Universidad de Buenos Aires
Facultad de Ciencias Económicas

CMA IADCOM - UBA
CENTRO DE INVESTIGACIÓN EN METODOS CUANTITATIVOS APLICADOS A LA ECONOMÍA Y LA GESTIÓN

DIRECTORA: PROFESORA EMÉRITA DRA. MARÍA TERESA CASPARRI

Introducción a la Probabilidad y a la Estadística

CÁTEDRA DE ESTADÍSTICA

MARÍA JOSÉ BIANCO

PRÓLOGO DE MARÍA TERESA CASPARRI

PRIMERA EDICIÓN

Roberto Darío Bacchini[1]
Lara Viviana Vázquez[2]
María José Bianco[3]
Javier I. García Fronti[4]

[1] Universidad de Buenos Aires. Facultad de Ciencias Económicas. Cátedra de Análisis Numérico Javier García Fronti. Ciudad Autónoma de Buenos Aires, Argentina.
[2] Universidad de Buenos Aires. Facultad de Ciencias Económicas. Càtedra de Bases Actuariales de las Inversiones y Financiaciones Pérez Raffo. Ciudad Autónoma de Buenos Aires, Argentina.
[3] Universidad de Buenos Aires. Facultad de Ciencias Económicas. Cátedra de Estadística I María José Bianco. Ciudad Autónoma de Buenos Aires, Argentina.
[4] Universidad de Buenos Aires. Facultad de Ciencias Económicas. Instituto de Investigaciones en Administración, Contabilidad y Métodos Cuantitativos para la Gestión (IADCOM). Centro de Investigación en Metodologías Básicas y Aplicadas a la Gestión (CIMBAGE). Ciudad Autónoma de Buenos Aires, Argentina.

Introducción a la probabilidad y la estadística / Roberto Darío Bacchini ... [et al.].
- 1a ed. - Ciudad Autónoma de Buenos Aires : Universidad de Buenos Aires. Facultad de Ciencias Económicas, 2018.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-950-29-1734-4

1. Teoría de las Probabilidades. 2. Estadísticas. I. Bacchini, Roberto Darío
CDD 519.5

AUTORES:

Roberto Darío Bacchini
Lara Viviana Vázquez
María José Bianco
Javier I. García Fronti

COLABORADORES:

Valeria Gogni
Matías Larrá
Andrea Lepera
Juana Llamas

Editor Responsable
Facultad de Ciencias Económicas, Universidad de Buenos Aires
Av. Córdoba 2122, 2do. Piso
Ciudad Autónoma de Buenos Aires, Argentina.
Contacto: cma@fce.uba.ar
Tel: 0054 011 5285-6539

ISBN 978-950-29-1734-4

Autoridades

Universidad de Buenos Aires

Rector: Dr. Alberto E. Barbieri

Facultad de Ciencias Económicas

Decano: Dr. Ricardo Pahlen Acuña

Instituto de Investigaciones en Administración, Contabilidad y Matemática

Centro de Investigación en Métodos Cuantitativos Aplicados a la Economía y la Gestión

Directora: Dra. María Teresa Casparri

Proyecto de Formación Docente en técnicas cuantitativas aplicadas (TCA)

Directora: Dra. María José Bianco

Subdirector: Roberto Armando García

Centro de Investigación en Métodos Cuantitativos Aplicados a la Economía y la Gestión (CMA - IADCOM)

Inaugurado en el año 2001, el **Centro de Investigación en Métodos Cuantitativos Aplicados a la Economía y la Gestión (CMA)** es actualmente parte del Instituto de Investigaciones en Administración, Contabilidad y Métodos Cuantitativos para la Gestión (IADCOM) de la Universidad de Buenos Aires, con sede en la Facultad de Ciencias Económicas.

El Centro se ha especializado en el estudio del riesgo de diversas actividades económicas y financieras en el contexto de países emergentes, haciendo especial énfasis en el bloque latinoamericano y particularmente en el caso de Argentina.

A lo largo del tiempo el Centro ha explotado diversos marcos conceptuales para la estimación del riesgo de activos financieros, proyectos de inversión real y de sectores económicos en su conjunto, en el marco de los principios de la gobernanza macroprudencial responsable.

CMA
IADCOM - UBA

Facultad de Ciencias Económicas
Universidad de Buenos Aires
Av. Córdoba 2122 2º Piso
Página web: www.economicas.uba.ar/cma
Teléfono: 5285-6539 – Correo Electrónico: cma@fce.uba.ar

Contenidos

Capítulo 1: Teoría de la Probabilidad
Darío Bacchini – Lara Vázquez – Valeria Gogni[5]

Capítulo 2: Variables Aleatorias y distribuciones
Darío Bacchini – Lara Vázquez – Valeria Gogni

Capítulo 3: Descripción de datos
Darío Bacchini – Lara Vázquez – Matías Larrá[6] *– Juana*[7] *Llamas*

Capítulo 4: Distribuciones de muestreo y Estimaciones
Darío Bacchini – Lara Vázquez – Andrea Lepera[8]

Capítulo 5: Pruebas de Hipótesis
Darío Bacchini – Lara Vázquez – Andrea Lepera

Capítulo 6: Regresión Lineal
Darío Bacchini – Lara Vázquez – Andrea Lepera

Capítulo 7: Números Índice
Darío Bacchini – Lara Vázquez – Valeria Gogni

Colaborador en el proceso de edición: Leonardo A. Dufour[9]

[5] Universidad de Buenos Aires. Facultad de Ciencias Económicas. Cátedra de Estadística I María José Bianco. Ciudad Autónoma de Buenos Aires, Argentina.
[6] Universidad de Buenos Aires. Facultad de Ciencias Económicas. Cátedra de Estadística I María José Bianco. Ciudad Autónoma de Buenos Aires, Argentina.
[7] Universidad de Buenos Aires. Facultad de Ciencias Económicas. Cátedra de Estadística I María José Bianco. Ciudad Autónoma de Buenos Aires, Argentina.
[8] Universidad de Buenos Aires. Facultad de Ciencias Económicas. Cátedra de Estadística I María José Bianco. Ciudad Autónoma de Buenos Aires, Argentina.
[9] Becario CIN por el proyecto UBACYT "Impacto económico de la nanotecnologia en la agroindustria Argentina: Valuación de inversiones e instrumentos de financiamiento" dirigido por Javier I. García Fronti.

Índice

Prólogo 8

1 Teoría de la Probabilidad 9
1.1 Teoría de Conjuntos: un repaso 10
1.2 Definición de Probabilidad 14
1.3 Axiomática 16
1.4 Probabilidad Conjunta y Marginal 17
1.5 Probabilidad Condicional e Independencia 18
1.6 Reglas de Conteo 24
1.7 Apéndice: Demostraciones 29

2 Variables Aleatorias y distribuciones de probabilidad 33
2.1 Definición 34
2.2 Distribución De Probabilidades 37
2.3 Cuantiles, Momentos y otras medidas 48
2.4 Distribuciones Discretas 56
2.5 Distribuciones Continuas 74
2.6 Anexo: Demostraciones 82

3 Descripción de Datos 85
3.1 Distribuciones de Frecuencia 87
3.2 Medidas de Posición 96
3.3 Apéndice: Demostraciones 112

4 Distribuciones de muestreo y Estimación 113
4.1 Muestreo Aleatorio: Técnicas 114
4.2 Distribuciones de Estadísticos 117
4.3 Distribución de $\bar{X}$ (media muestral): varianza poblacional conocida 120
4.4 Distribución de $\bar{p}$ (proporción muestral) 129
4.5 Distribución de s^2 (varianza muestral) en poblaciones Normales 131
4.6 Distribución de $\bar{X}$: varianza poblacional desconocida 135
4.7 Estimación: puntual y por intervalo 138
4.8 Propiedades deseables de un Estimador 140
4.9 Estimación Puntual: métodos 144
4.10 Intervalos de Confianza (IC) 146
4.11 IC para comparar poblaciones 152
4.12 Tamaño Muestral y poblaciones Finitas 157
4.13 Apéndice: Demostraciones 162

5 Pruebas de Hipótesis 166
5.1 Conceptos Generales del Testeo de Hipótesis 167
5.2 Testeo para Medias 175
5.3 Testeo para proporciones 182
5.4 Testeo para varianzas 185
5.5 Relación de las pruebas de hipótesis con los Intervalos de confianza 188

6 Regresión Lineal 190
6.1 Regresión Lineal 191

6.2 Estimación de la Recta de Regresión 206
6.3 Bondad del Ajuste y Coeficiente de determinación 214
6.4 Anexo: Demostraciones 217

7 Números Índice 220
7.1 Concepto 221
7.2 Índices Simples y Ponderados 225
7.3 Índices de Laspeyres y Paasche 230
7.4 Cambios en la Base 234

Bibliografía 236

Prólogo

Introducción a la Probabilidad y la Estadística es el resultado la labor de docencia y de investigación desarrollada por profesores y auxiliares docentes de nuestro Departamento Pedagógico de Matemática, en articulación con el programa de formación docente en métodos cuantitativos del Centro de Investigación en Métodos Cuantitativos aplicados a la Economía y la Gestión (IADCOM), de la Facultad de Ciencias Económicas de la Universidad de Buenos Aires.

Para mí es muy grato tener el honor de prologar este trabajo, el cual ha sido coordinado por la titular de cátedra de Estadística de nuestro departamento, María José Bianco, con la cual también comparto tareas de dictado de seminarios en el doctorado.

María José ha coordinado esta publicación, en conjunto con Darío Bacchini, Lara Vazquez y Javier García Fronti. Con el fin de entregar a los alumnos de la materia un libro de texto, se han procesado didácticamente los materiales y se fijaron siete unidades temáticas: 1. Teoría de la probabilidad, 2. Variables aleatorias y distribuciones de probabilidad, 3. Descripción de datos, 4. Distribuciones de muestreo y Estimación, 5. Prueba de hipótesis, 6. Regresión lineal y 7. Números índice. Dichos capítulos han sido elaborados por Darío Bacchini y Lara Vazquez, en colaboración con Matías Larrá, Juana Llamas, Andrea Lepera y Valeria Gogni.

Quiero terminar este prólogo remarcando el entusiasmo y dedicación del grupo docente involucrado en este trabajo y quiero desearles el mayor de los éxitos en la tarea de los próximos cuatrimestres. Asimismo, considero muy importante que este grupo de docentes publique este texto en nuestra facultad y permita el acceso libre a los estudiantes.

Profesora Emérita Dra. María Teresa Casparri

1 Teoría de la Probabilidad

Dario Bacchini
Lara Vazquez
Valeria Gogni

La vida está llena de incertidumbres. De hecho, casi todos los eventos que nos suceden llevan consigo algo de aleatoriedad. Por ejemplo, podemos decir que el ómnibus que nos lleva a la Facultad pasa regularmente a las 8.45 a.m.; pero, ¿podemos afirmar con toda *certeza* que mañana pasará *exactamente* a esa hora?

El lector puede imaginar, sólo con un pequeño esfuerzo, ejemplos como el del párrafo anterior. Sobre la base de este (y de los que se le hayan ocurrido a usted), podemos realizar las primeras definiciones referidas a diversos fenómenos.

- Un fenómeno se dice *determinístico*, si se sabe con toda certeza cuál será su comportamiento.
- Un fenómeno es *aleatorio*, cuando no podemos afirmar con certeza cuál será su comportamiento.

Ejemplo 1

Si lanzamos una piedra al aire, podemos afirmar con certeza que volverá a caer a la superficie de la tierra, pero no podemos saber con precisión el punto en el que caerá. Así, la caída es un fenómeno determinístico, mientras que el lugar en que se producirá dicha caída es aleatorio, ya que existe incertidumbre respecto del punto preciso en el que caerá.

Ejemplo 2

Un seguro de vida paga un monto determinado en caso de muerte del asegurado. El pago del monto es un fenómeno determinístico, ya que sabemos que la muerte indefectiblemente sucederá. Sin embargo, el momento de pago es aleatorio, ya que no podemos precisar con exactitud la edad a la cual fallecerá cada asegurado.

Como verá el lector, ejercitando un poco su imaginación, estamos rodeados de fenómenos aleatorios, y lidiamos a diario con los mismos casi sin notarlo.

Pensemos, además, en la cantidad de afirmaciones que oímos a diario, casi sin darnos cuenta, relacionadas con la "probabilidad" de ocurrencia de determinados fenómenos. Por ejemplo, frecuentemente escuchamos en el noticiero que hay alta probabilidad de lluvias, o a un locutor decir que la probabilidad de que un equipo de fútbol revierta un resultado adverso es casi nula, o bien, que la probabilidad de ganar en un determinado juego es una en cien. Sin embargo, seguramente, pocas veces hemos reparado en pensar qué quiere decir exactamente un valor determinado de "probabilidad".

En este capítulo, lo que pretendemos es justamente precisar algunas definiciones de probabilidad. La Teoría de la Probabilidad es la encargada de estudiar los fenómenos aleatorios y, mediante ciertos axiomas que veremos más adelante, se define lo que llamaremos *medida de probabilidad*. A su vez, a partir de dichos axiomas se desprenden una serie de propiedades de la probabilidad muy útiles para su aplicación al análisis de fenómenos concretos.

Así, mediante ciertos estudios probabilísticos se podrán realizar afirmaciones respecto de la probabilidad de que determinado artículo de una línea de producción sea defectuoso, la probabilidad de ganar cierto juego de azar o la probabilidad de que al extraer un individuo al azar del curso de estadística, el mismo sea un hombre y, además, sea fumador.

En el presente capítulo se presentarán los conceptos básicos relacionados con la Teoría de la Probabilidad, la cual constituye una piedra angular de la Estadística. Pero antes de entrar de lleno en el tema que nos compete, expondremos un breve repaso de la Teoría de Conjuntos, la cual será una herramienta fundamental para los desarrollos posteriores.

1.1 Teoría de Conjuntos: un repaso

La Teoría de Conjuntos, o al menos los conceptos básicos de ésta, es desarrollada en los estudios de nivel medio. Sin embargo, aquí se realiza una breve introducción a modo de repaso y con el fin de establecer la notación a usar a lo largo del capítulo.

De acuerdo con lo visto anteriormente, lo que nos interesa estudiar es el comportamiento de los fenómenos aleatorios. Dicho comportamiento puede relacionarse con el resultado de un determinado experimento. Por ejemplo, el experimento puede consistir en medir la hora en que pasa el ómnibus, u observar el punto de caída de una piedra o bien anotar el resultado de un

partido de fútbol. Teniendo en mente esta relación, pasemos a desarrollar la teoría desde esta óptica, considerando al comportamiento aleatorio de ciertos fenómenos como resultados de un experimento determinado.

Definimos, a continuación, ciertos elementos comunes de cualquier experimento:

- **Espacio muestral (Ω)**: conjunto de todos los posibles resultados que se pueden dar al realizar un experimento.
- **Evento Simple:** cada uno de los posibles resultados, considerados individualmente. Es decir, cada uno de los elementos del espacio muestral.
- **Evento compuesto:** conjunto de eventos simples.

En general, salvo aclaración en contrario, la letra griega omega (Ω) representará el espacio muestral, mientras que las letras mayúsculas del alfabeto latino (A, B,...) denotarán eventos, tanto simples como compuestos. Unos ejemplos clarificarán las definiciones enunciadas.

Ejemplo 3

Considere el lanzamiento de un dado. El espacio muestral está dado por $\Omega = \{1,2,3,4,5,6\}$, un evento simple es *A ="el resultado es 2*" y un evento compuesto es *B =" el resultado es un número par"*. Los eventos pueden escribirse también como $A = \{2\}$ y $B = \{2,4,6\}$.

Ejemplo 4

Si se considera un experimento dado por el lanzamiento de una moneda, el espacio muestral está dado por $\Omega = \{cara, ceca\}$, y en este caso sólo es posible considerar los eventos simples $A = \{cara\}$ y $B = \{ceca\}$.

Ejemplo 5

Considere el lanzamiento de dos monedas, una por vez. El espacio muestral está dado por $\Omega = \{CaCe, CaCe, CeCe, CeCa\}$ [10]. Un evento simple es $A = \{CaCe\}$, o de manera extensiva *A = "el primer lanzamiento es cara y el segundo ceca"*. Un evento compuesto $B = \{CaCa, CeCe\}$, o de manera extensiva *B = "los dos lanzamientos arrojan el mismo resultado"*[11].

A continuación, definimos algunas **operaciones básicas** relacionadas con conjuntos:

- **Unión** de dos conjuntos ($A \cup B$): está dada por el conjunto de todos los resultados que pertenecen al evento A *o* al evento B *o* a ambos.

Ejemplo 6.a

Si se considera el lanzamiento de un dado y se definen los eventos $A = \{1,2,3\}$ y $B = \{2,4,6\}$, entonces $A \cup B = \{1,2,3,4,6\}$.

- **Intersección** de dos conjuntos ($A \cap B$): está dada por el conjunto de los resultados que pertenecen tanto a A como a B, es decir a A *y* a B simultáneamente.

Ejemplo 6.b

Continuando con los conjuntos definidos en el Ejemplo 6.a, tenemos que $A \cap B = \{2\}$.

- **Complemento** de un conjunto (A^C): es el conjunto de todos los elementos del espacio muestral que no pertenecen al evento A.[12]

[10] Ca = cara; Ce = ceca.

[11] Nótese que este caso se considera relevante el orden en que se dan los resultados, ya que si no interesara el orden, los eventos $A = \{CaCe\}$ y $C = \{CeCa\}$ serían iguales, y el espacio muestral se reduciría a $\Omega = \{CaCa, CaCe, CeCe\}$.

[12] El complemento suele denotarse también como $\bar{A}$.

–

Ejemplo 6.c

Continuando los ejemplos anteriores: $A^C = \{4,5,6\}$ y $B^C = \{1,3,5\}$.

Por otro lado, podemos realizar algunas definiciones que están relacionadas con las características de los conjuntos:

– Cuando todos los elementos de un conjunto A pertenecen también a otro conjunto B, se dice que A **está incluido** en B, y se denota $A \subset B$.

– Un conjunto que no posee ningún elemento se denomina **conjunto vacío** y se denota por $\varnothing = \{\ \}$.

– Dos eventos A y B son **mutuamente excluyentes** si la ocurrencia de uno implica la no-ocurrencia del otro, es decir, la intersección de los conjuntos que representan a dos eventos mutuamente excluyentes es el conjunto vacío: $A \bigcap B = \varnothing$.

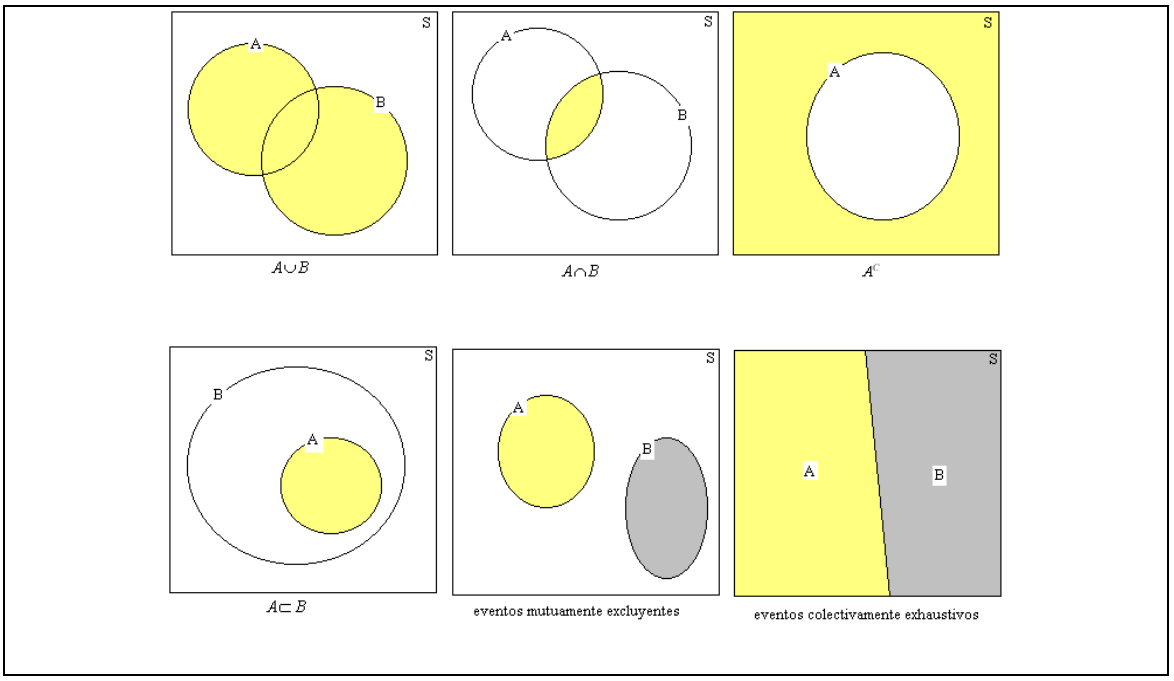
– Dos eventos A y B son **colectivamente exhaustivos** si la unión de los conjuntos que los representan conforman el espacio muestral: $A \bigcup B = \Omega$. Es decir, que con certeza ocurrirá al menos uno de ellos.

De acuerdo con las definiciones enunciadas hasta aquí, se pueden extraer las siguientes conclusiones:

– $A \subset \Omega$: Todo evento está incluido en el espacio muestral.

– $A \bigcap A^C = \varnothing$ y $A \bigcup A^C = \Omega$: Un evento y su complemento son mutuamente excluyentes y colectivamente exhaustivos.

En la Figura, se puede observar el diagrama de Venn de cada una de las operaciones y definiciones expuestas previamente.



Podemos notar a su vez, que cada operación define nuevos eventos, con los cuales se podrán efectuar nuevamente las operaciones definidas, realizando de esta manera, operaciones compuestas.

Ejemplo 7

Consideremos el lanzamiento de un dado. El espacio muestral, como ya hemos visto, es $\Omega = \{1,2,3,4,5,6\}$. Además, definimos los eventos *A = "el resultado es mayor a 3"*, *B = "el resultado es impar"*, y *C = "el resultado es 1"*. O bien, $A = \{4,5,6\}$, $B = \{1,3,5\}$, y $C = \{1\}$.

De acuerdo con las definiciones arriba enunciadas, podemos obtener los siguientes resultados: $A \cap C = \varnothing$ (A y C son mutuamente excluyentes) y $C \subset B$ (C está incluido en B).

Ejemplo 8

Supongamos que, con los datos del ejemplo anterior, deseamos hallar las operaciones compuestas $A^C \cap B$, y $\left(A \cap C^C\right) \cup B$. Siempre es recomendable operar paso a paso.

Para hallar $A^C \cap B$, primero obtenemos $A^C = \{1,2,3\}$ y luego, realizamos la intersección de este último con B. Finalmente, $A^C \cap B = \{1,3\}$.

Para la segunda operación deseada, calculamos primero $C^C = \{2,3,4,5,6\}$, luego realizamos la intersección con A, dando por resultado $A \cap C^C = \{4,5,6\}$, y finalmente, al realizar la unión con B, el resultado es $\left(A \cap C^C\right) \cup B = \{1,3,4,5,6\}$.

1.1.1 Propiedades de operaciones

Las operaciones entre conjuntos definidas en la sección anterior presentan algunas propiedades que vale la pena tener presentes. A modo de ejercicio, el lector puede comprobar las propiedades que siguen realizando, en cada una de ellas, el diagrama de Venn del miembro izquierdo y del miembro derecho por separado, y luego, compararlos para verificar la igualdad.

- **Asociatividad de la unión:** la unión de un conjunto A con la unión de otros dos conjuntos B y C, es igual a la unión de la unión de los dos primeros con el tercero. Es decir:

$$A \cup (B \cup C) = (A \cup B) \cup C$$

- **Asociatividad de la intersección:** la intersección de un conjunto *A* con la intersección de otros dos conjuntos *B* y *C*, es igual a la intersección de la intersección los dos primeros con el tercero. Es decir:

$$A \cap (B \cap C) = (A \cap B) \cap C$$

- **Distributividad de la intersección respecto de la unión:** La intersección de un evento A con la unión de otros dos eventos B y C, es la unión de las intersecciones de A con cada uno de ellos. Es decir:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

- **Distributividad de la unión respecto de la intersección:** La unión de un evento *A* con la intersección de otros dos eventos *B* y *C*, es la intersección de las uniones de *A* con cada uno de ellos. Es decir:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

- **Complemento de la unión:** el complemento de la unión de los conjuntos *A* y *B* es la intersección de los complementos de cada uno de ellos. Es decir:

$$\left(A \cup B\right)^C = A^C \cap B^C$$

- **Complemento de la intersección:** el complemento de la intersección de los conjuntos *A* y *B* es la unión de los complementos de cada uno de ellos. Es decir:

$$\left(A \cap B\right)^C = A^C \cup B^C$$

1.2 Definición de Probabilidad

En esta sección veremos que existen varias maneras de definir a la probabilidad, las cuales surgirán de acuerdo con el tipo de fenómeno que estemos analizando. A su vez, se observará que estas definiciones están estrechamente ligadas a las nociones intuitivas que se pueden llegar a tener respecto de la probabilidad.

1.2.1 Definición Clásica

Si preguntamos a cualquier persona que nos diga cuál es la probabilidad de obtener *ceca* al lanzar una moneda al aire, casi con seguridad nos contestará "un 50%". Asimismo, si consultamos cuál es la probabilidad de obtener *el número 6* al lanzar un dado, es muy posible que la respuesta sea "un sexto"; mientras que si preguntamos cuál es la probabilidad de obtener *un número par*, la respuesta será "un 50%". Estas respuestas intuitivas están ligadas a la definición clásica de probabilidad:

Sea Ω un espacio muestral finito que contiene N eventos simples, y sea A un evento que puede darse de n maneras distintas; es decir, que al realizar un experimento hay N resultados posibles de los cuales n son favorables al evento A. La probabilidad de que ocurra el evento A está dada por:

$$P(A) = \frac{resultados\ favorables}{resultados\ posibles} = \frac{n}{N}$$

Si relacionamos la definición precedente con el repaso de la Teoría de Conjuntos, podemos afirmar que la probabilidad de que se dé el evento A está dada por el cociente entre la cantidad de elementos del conjunto favorables al evento A y el número de elementos del conjunto Ω, siendo estos últimos igualmente probables.

Cabe aclarar que el evento A puede ser simple o compuesto, y en este segundo caso, puede resultar complicado determinar la cantidad de maneras en que puede darse el evento. A su vez, hay ocasiones en que resulta complicado determinar la cantidad de elementos que posee el espacio muestral Ω. Para ambos casos, resultan útiles **las *reglas de conteo*** (combinatoria, variaciones, etc.) que serán vistas en la sección 6 de este capítulo.

Ejemplo 9

Un individuo está por jugar a un juego en el que se lanzan dos dados equilibrados; gana $ 1 si el resultado de la suma de los números obtenidos en ambos dados es siete.

La cantidad de resultados posibles cuando se lanzan dos dados es 36 (estos resultados son igualmente probables): si el resultado del primer dado es 1, el segundo puede arrojar cualquiera de los números del 1 al 6, con lo cual ya tenemos seis resultados posibles; si el primer dado es 2, el segundo nuevamente podrá arrojar cualquier valor del 1 al 6, con lo cual ya sumamos doce resultados; y así sucesivamente hasta completar $6^2 = 36$ resultados posibles.

Luego, deberíamos determinar la cantidad de resultados favorables al evento "la suma de los dados es 7": éste puede darse de seis maneras distintas (1 y 6, 2 y 5, 3 y 4, 4 y 3, 5 y 2, 6 y 1).

En la siguiente tabla, se resumen todos los resultados posibles, y aparecen sombreados los resultados favorables al evento:

		Dado 2					
		1	**2**	**3**	**4**	**5**	**6**
Dado 1	**1**	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Así, la probabilidad de que el apostador gane, está dada por el cociente entre el número de resultados favorables al suceso y el número de resultados posibles:

$$P(A) = \frac{6}{36} = \frac{1}{6} = 0,1667$$

1.2.2 Definición Frecuentista

La Definición Frecuentista de probabilidad surge debido a la existencia de fenómenos aleatorios en los cuales no se puede determinar con precisión la probabilidad clásica de cada evento simple, es decir, que no podemos precisar cuántos resultados favorables a un evento existen y/o cuántos resultados posibles hay.

Consideremos algunos ejemplos en los cuales no se puede determinar con precisión los casos favorables y los casos posibles: un jefe de control de calidad desea determinar la probabilidad de que un artículo sea defectuoso, un fanático está interesado en la probabilidad de que su equipo de fútbol gane o un profesor que quiere saber la probabilidad de que sus alumnos aprueben.

Para estimar la probabilidad de cada uno de esos eventos, se recurre a la segunda manera de definir a la probabilidad, utilizando la ***frecuencia relativa*** de ocurrencia de los mismos.

Sea K el número de veces que se observa un fenómeno determinado, y sea k el número de veces en que ocurre un resultado favorable al evento A. La probabilidad de ocurrencia del evento A es la frecuencia relativa observada cuando el número total de observaciones crece indefinidamente:

$$P(A) = \lim_{K \to \infty} \frac{k}{K}$$

La gran mayoría de los fenómenos aleatorios con que nos enfrentaremos en la práctica son de este tipo, por lo cual esta definición de probabilidad será muy utilizada a lo largo de la presente obra.

Ejemplo 10

Consideremos un control de calidad de una empresa, en el cual se desea saber la probabilidad de que un determinado artefacto tenga una vida útil superior a las 1200 hs. Para ello, el departamento de control de calidad separa 500 unidades de la producción y mide la vida útil de cada unidad. Los resultados se observan en la siguiente tabla:

Duración (en hs.)	frec. abs.	frec. rel.
menos de 800	10	2%
800 a 899	40	8%
900 a 999	55	11%
1000 a 1099	70	14%
1100 a 1199	85	17%
1200 a 1299	115	23%
1300 a 1399	84	17%
1400 o más	41	8%
	500	100%

Así, de acuerdo a la Definición Frecuentista (y considerando que 500 es un número suficientemente grande), la probabilidad de que la vida útil sea mayor o igual a 1200 hs. es:

$$P(A) = \frac{115 + 84 + 41}{500} = 0,23 + 0,17 + 0,08 = 0,38$$

Esta definición de probabilidad da lugar a las pruebas de hipótesis, que serán tratadas en el Capítulo 7. Consideremos el lanzamiento de un dado y supongamos que queremos detectar si el mismo está cargado. Para ello, podríamos lanzar el dado un gran número de veces y observar la frecuencia relativa de ocurrencia de cada resultado; por ejemplo, si lanzamos el dado 600 veces, deberíamos esperar que 100 veces se dé cada uno de los resultados posibles. Sin embargo, difícilmente esto ocurra, y supongamos que el resultado 2 se dio 140 veces. Lo que se pretende al realizar un test de hipótesis, es probar si la evidencia empírica es suficiente como

para afirmar que el dado está efectivamente cargado a favor del número 2, o si la observación de una cantidad elevada de dicho resultado se debió simplemente al azar propio del experimento. Continuaremos con este tema en el capítulo correspondiente.

1.2.3 Definición Subjetiva

La Definición Subjetiva de probabilidad está relacionada con el grado de creencia que tiene quien lleva a cabo un experimento respecto de la probabilidad de ocurrencia del mismo.

Así, por ejemplo, al lanzar un nuevo producto al mercado, un gerente de ventas puede creer que el mismo tendrá un 70% de aceptación en el público, es decir, que la probabilidad (subjetiva) de que un individuo acepte el producto es de 0,7. Esta probabilidad suele llamarse también *probabilidad a priori*, ya que refleja el grado de creencia antes de que se realice cualquier prueba empírica. Las probabilidades a priori suelen modificarse luego mediante algún tipo de experimento como, por ejemplo, una encuesta para ver la aceptación que podría tener el producto. Una vez que el experimento se realiza, se modifican las probabilidades a priori para obtener las probabilidades a posteriori, las cuales serán utilizadas para tomar decisiones.

Este tipo de análisis de problemas es lo que se conoce como Análisis Bayesiano, mediante el cual se modifican las probabilidades subjetivas (*a priori*) utilizando el Teorema de Bayes, el cual será expuesto más adelante. La tarea consiste en analizar la información suministrada por los resultados de algún tipo de experimento (por ejemplo, como dijimos anteriormente, una encuesta), para obtener probabilidades condicionadas a dicha información. Este tipo de análisis está íntimamente relacionado con la dependencia estadística de ciertos fenómenos, el cálculo de probabilidades condicionales y el Teorema de Bayes, temas desarrollados más adelante en el presente Capítulo. Cabe destacar que el Análisis Bayesiano tiene una amplitud mucho mayor que la mencionada. Sin embargo en esta obra no se tratarán con profundidad problemas de este tipo.

Antes de iniciar el estudio de probabilidades condicionales y de fenómenos estadísticamente independientes, desarrollaremos los axiomas principales que debe cumplir cualquier medida de probabilidad.

1.3 Axiomática

Todas las definiciones anteriores están íntimamente ligadas a la parte experimental de la Estadística. Sin embargo, en los últimos años, la Teoría de la Probabilidad ha evolucionado de manera sorprendente y las definiciones se han hecho más rigurosas desde un punto de vista matemático.

En este contexto, el ruso Andrei Kolmogorov (1933) definió la medida o función de probabilidad mediante una serie de axiomas. Éstos, si bien son válidos para cualquiera de las definiciones de probabilidad expuestas anteriormente, amplían la definición incluyendo a cualquier medida que los verifique.

Dado un espacio muestral Ω, llamamos **medida de probabilidad** a una función P que va del espacio muestral al conjunto de los números reales si satisface los siguientes axiomas:

a) *Si A es un evento cualquiera, entonces* $P(A) \geq 0$

b) $P(\Omega) = 1$

c) Si A_i $(i=1,2...)$ *son eventos mutuamente excluyentes, entonces:*

$$P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$$

Es decir, que la probabilidad "P" asigna a cada elemento del espacio muestral un número que verifica los axiomas expuestos.

A partir de estos tres axiomas, se desprenden las siguientes **conclusiones**[13]:

– Conocida la probabilidad de un evento A, se puede conocer la de su complemento A^C mediante la siguiente relación:

[13] Las demostraciones de estas conclusiones se encuentran en el Apéndice del final del presente capítulo.

$$P\left(A^C\right)=1-P(A)$$

- La función de probabilidad está incluida en el intervalo real $[0,1]$, es decir:

$$0 \le P(A) \le 1$$

- La probabilidad del evento vacío es nula, es decir:

$$P(\varnothing)=0$$

- Si A y B son dos eventos cualesquiera[14], entonces la probabilidad de su unión es:

$$P(A \cup B)=P(A)+P(B)-P(A \cap B)$$

- Si A, B y C son tres eventos cualesquiera, entonces la probabilidad de su unión es:

$$P\left(A \cup B \cup C\right)=P(A)+P(B)+P(C)-P\left(A \cap B\right)-P\left(A \cap C\right)-P\left(B \cap C\right)+P\left(A \cap B \cap C\right)$$

- Si A está incluido en B, entonces la probabilidad de A es menor o igual a la probabilidad de B:

$$A \subseteq B \quad \Rightarrow \quad P(A) \le P(B)$$

- Si A está incluido en B, entonces la probabilidad de la intersección de los dos conjuntos coincide con la probabilidad de A:

$$A \subseteq B \quad \Rightarrow \quad P(A \cap B)=P(A)$$

1.4 Probabilidad Conjunta y Marginal

En la presente sección, expondremos conceptos relacionados con la probabilidad de eventos que ocurren simultáneamente y la probabilidad de eventos simples. Ambos conceptos ya han sido estudiados y ejemplificados en apartados anteriores, pero no han sido definidos de manera precisa.

1.4.1 Probabilidad Conjunta

Si bien hasta aquí no hemos definido el concepto de Probabilidad Conjunta, hemos estado trabajando con él de manera implícita. La probabilidad conjunta de dos eventos *A* y *B* es simplemente la probabilidad de que ambos sucedan al mismo tiempo.

Ejemplo 11

Consideremos el lanzamiento de un dado. La probabilidad del evento *A = "el resultado está entre 2 y 4, ambos inclusive"* está dada por:

$$P(A)=\frac{3}{6}=\frac{1}{2}$$

Definamos los siguientes eventos simples: A_1 = *"el resultado es mayor o igual a 2"*, A_2 = *" el resultado es menor o igual a 4"*. Claramente podemos ver qué $A=A_1 \cap A_2$. De este modo, la probabilidad conjunta de los eventos A_1 y A_2 está dada por:

$$P\left(A_1 \cap A_2\right)=P(A)=\frac{1}{2}$$

Basándonos en el ejemplo anterior, podemos formalizar la definición:

[14] No necesariamente mutuamente excluyentes.

Sea A un evento que surge como resultado de la intersección de los eventos $A_1, A_2, ..., A_n$, es decir: $A_1 \cap A_2 \cap ... \cap A_n = A$. La Probabilidad Conjunta de los eventos $A_1, A_2, ..., A_n$ es la probabilidad del evento que surge como intersección de todos ellos:

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P\left(\bigcap_{j=1}^{n} A_j\right) = P(A)$$

1.4.2 Probabilidad Marginal

La Probabilidad Marginal es simplemente la probabilidad de ocurrencia de un evento *A*, sin pensar en la existencia de otro evento *B* que suceda de modo simultáneo con *A*.

Ejemplo 12

Consideremos el Ejemplo 11. La probabilidad conjunta de los eventos A_1 y A_2 es:

$$P(A_1 \cap A_2) = \frac{1}{2}$$

La Probabilidad Marginal de cada uno de los eventos es:

$$P(A_1) = \frac{5}{6} \qquad P(A_2) = \frac{4}{6}$$

Nótese que la probabilidad marginal es simplemente la probabilidad de un evento determinado. Lo mismo ocurre con la probabilidad conjunta. Sin embargo, utilizamos el término *marginal* o *conjunta* para hacer referencia a que la probabilidad es calculada en un contexto en el cual se estudian los fenómenos de manera simultánea.

1.5 Probabilidad Condicional e Independencia

En la presente sección analizaremos la influencia que tiene sobre un evento determinado la información que se posee sobre otro evento relacionado con el mismo, si es que existe tal influencia.

1.5.1 Probabilidad Condicional

Cuando se trabaja con fenómenos aleatorios, muchas veces podemos contar con cierta información que modificaría nuestra estimación de la probabilidad del mismo. En estos casos, se dice que la probabilidad del evento en cuestión está **condicionada** a la ocurrencia de otro evento.

Ejemplo 13

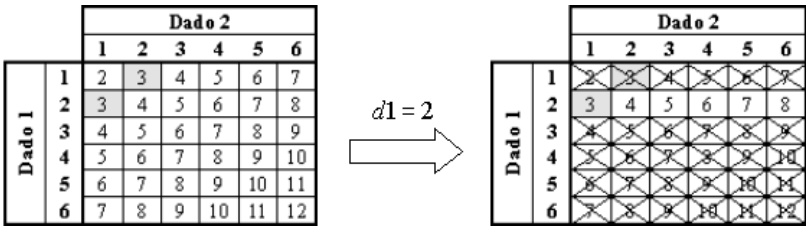
Consideremos el lanzamiento de dos dados. El resultado del primero de ellos se denotará por d_1 y el resultado del segundo por d_2. La probabilidad de que la suma sea 3 está dada por:

$$P(d_1 + d_2 = 3) = \frac{2}{36} = \frac{1}{18}$$

Sin embargo, si sabemos que el resultado del primer dado es 2, la única manera de que la suma sea 3 es que el resultado del segundo sea 1, por lo tanto, la probabilidad será:

$$P(d_1 + d_2 = 3 \text{ } \textit{sabiendo que} \text{ } d_1 = 2) = \frac{1}{6}$$

En la siguiente Tabla se ilustra el razonamiento seguido en el ejemplo:



A continuación, definimos formalmente el cálculo de probabilidades condicionadas.

Sean A y B dos eventos de un espacio muestral Ω. La probabilidad de que se produzca el evento A condicionada a (sabiendo) que ocurrió el evento B, $P(A|B)$, es el cociente entre la probabilidad conjunta y la probabilidad del evento conocido:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \textit{siendo} \qquad P(B) > 0 \qquad (1)$$

Si bien el Ejemplo 13 se resolvió de manera directa utilizando la definición clásica de probabilidad, podría resolverse utilizando la fórmula anterior, de la siguiente manera:

Ejemplo 14

Consideremos el ejemplo anterior. La probabilidad de que la suma de los dos dados sea 3, sabiendo que el resultado del primer dado fue 2 es:

$$P(d1+d2=3|d1=2) = \frac{P(d1+d2=3 \cap d1=2)}{P(d1=2)} = \frac{1/36}{1/6} = \frac{1}{6}$$

Puede observarse que el condicionamiento es equivalente a "recortar" el espacio muestral: se eliminan del espacio muestral aquellos eventos que resultan imposibles de acuerdo a la información con la que contamos. Esta afirmación puede verse claramente en la figura del Ejemplo 13.

1.5.2 Eventos Estadísticamente Independientes

Lógicamente, puede suceder que tengamos información sobre la ocurrencia de un evento determinado B, y sin embargo la probabilidad marginal de ocurrencia del evento A no se vea alterada. Esto quiere decir, que la ocurrencia de B no tiene ninguna influencia sobre el evento A, es decir, que los eventos son estadísticamente independientes.

Dos eventos A y B son **estadísticamente independientes**, si la ocurrencia de uno no afecta la probabilidad de ocurrencia del otro, es decir que:

$$P(A|B) = P(A) \qquad (2)$$

De las definiciones de probabilidad condicional y eventos independientes, se desprende la regla del producto de probabilidades de eventos independientes.

Si A y B son dos eventos **estadísticamente independientes**, entonces la **probabilidad conjunta es igual el producto** de las probabilidades marginales:

$$P(A \cap B) = P(A).P(B)$$

Se destaca que **la independencia es una relación simétrica** entre eventos, esto quiere decir que si *A* es independiente de *B*, entonces *B* es independiente de *A*. Como ejercicio, el lector puede demostrar esto a partir de las definiciones expuestas.

Ejemplo 15

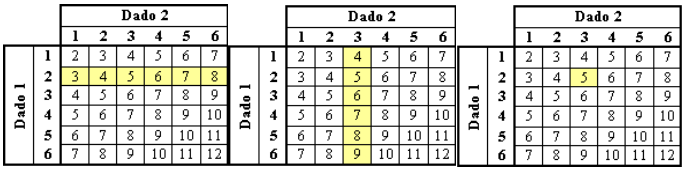
Consideremos el lanzamiento de dos dados y los siguientes eventos: A_1 = *"el resultado del primer dado es dos"* y A_2 = *"el resultado del segundo es tres"*. La probabilidad marginal de cada uno de ellos es:

$$P(A_1) = P(d_1 = 2) = \frac{1}{6}; \qquad P(A_2) = P(d_2 = 3) = \frac{1}{6}$$

La probabilidad conjunta es[15]:

$$P(A_1 \cap A_2) = \frac{1}{36}$$

Como puede observarse, la probabilidad conjunta es el producto de las probabilidades marginales.



Ejemplo 16

Consideremos el lanzamiento de dos dados y los siguientes eventos: B_1 = *"el resultado del primer dado es dos"* y B_2 = *"la suma de los resultados de los dos dados es cinco"*. La probabilidad marginal de cada uno de ellos es:

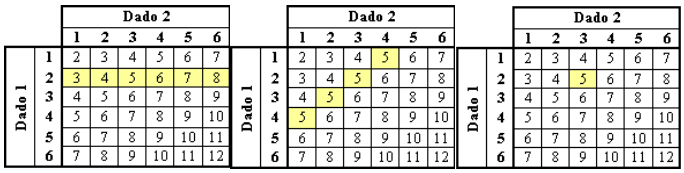
$$P(B_1) = P(d_1 = 2) = \frac{1}{6}; \qquad P(B_2) = P(d_1 + d_2 = 5) = \frac{4}{36} = \frac{1}{9}$$

La probabilidad conjunta es:

$$P(B_1 \cap B_2) = \frac{1}{36}$$

ya que existe una única manera de que simultáneamente, el resultado del primer dado sea 2 y la suma sea 5 (el primero resultado debe ser 2 y el segundo 3).

En este caso, los eventos son dependientes, ya que el producto de las probabilidades marginales no iguala a la probabilidad conjunta.



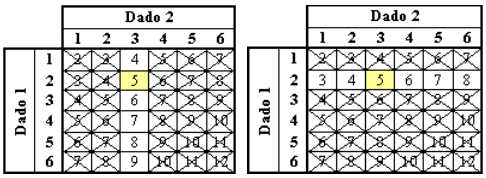
Ejemplo 17

Calculando las probabilidades condicionales del ejemplo 15, podemos verificar los siguientes resultados:

$$P(A_1 | A_2) = P(d_1 = 2 | d_2 = 3) = \frac{1}{6}; \qquad P(A_2 | A_1) = P(d_2 = 3 | d_1 = 2) = \frac{1}{6}$$

[15] Hay una posibilidad sobre las 36 combinaciones posibles al lanzar dos dados.

Las cuales son idénticas a las probabilidades marginales de A_1 y A_2, implicando de esta manera la independencia de los eventos. La gráfica auxiliar al cálculo se ilustra a continuación:

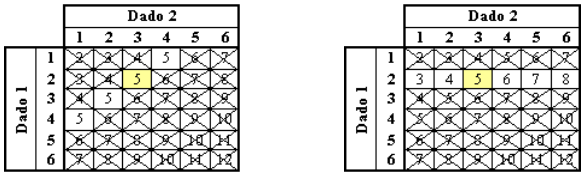


Al calcular las probabilidades condicionales del Ejemplo 16, tenemos que:

$$P(B_1|B_2) = P(d_1 = 2|d_1 + d_2 = 5) = \frac{1}{4};$$

$$P(B_2|B_1) = P(d_1 + d_2 = 5|d_1 = 2) = \frac{1}{6}$$

Éstas son diferentes de las probabilidades marginales, implicando dependencia entre B_1 y B_2. Las siguientes figuras ilustran los cálculos.



1.5.3 Ley de Probabilidad Total

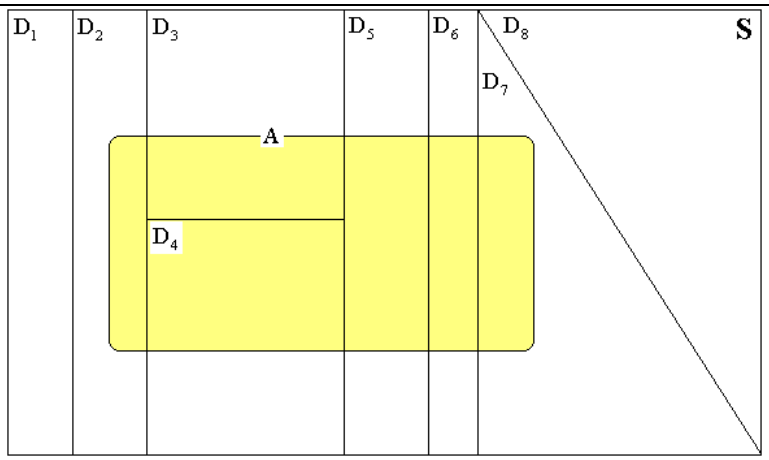
Utilizando el esquema de probabilidades condicionales, si no se conoce directamente la probabilidad de ocurrencia de un evento *A*, la misma puede obtenerse utilizando la ley de la probabilidad total, la cual determina la probabilidad de un evento por medio de las probabilidades conjuntas del mismo con otros eventos mutuamente excluyentes y colectivamente exhaustivos.

Sea *A* un evento de un espacio muestral Ω y sea $D_j\ (j = 1,2,...,n.)$ una partición del espacio muestral (es decir, que los D_j son mutuamente excluyentes y colectivamente exhaustivos), entonces la **probabilidad total** del evento *A* está dada por[16]:

$$P(A) = \sum_{j=1}^{n} P(A \cap D_j) = \sum_{j=1}^{n} P(A|D_j).P(D_j)$$

Esta fórmula puede comprobarse simplemente inspeccionando la siguiente figura, donde el espacio muestral se divide en ocho partes. Una demostración más rigurosa se expone en el Apéndice del final del capítulo.

[16] Nótese que pudimos haber dicho simplemente la probabilidad de *A*, o bien, la probabilidad marginal de *A*. La forma de expresar la probabilidad dependerá del contexto.



Evento *A* a partir de su intersección con una partición del espacio muestral.

Ejemplo 18

Consideremos una bolsa con cubitos y bolitas de madera de dos colores (rojo y verde). Se sabe que el 20% de las piezas rojas son bolitas, es decir, $P(b/r)=0,2$ y el 40% de las verdes son bolitas, es decir, $P(b/v)=0,4$. Además, se conoce que el 70% de las piezas son rojas ($P(r)=0,7$).

La probabilidad de extraer una bolita puede calcularse mediante el empleo de la fórmula de cálculo de probabilidad total, teniendo en cuenta que el porcentaje de piezas verdes será el complemento del porcentaje de piezas rojas: $P(v)=0,3$. Finalmente, la probabilidad deseada es:

$$P(b)=P(r\cap b)+P(v\cap b)$$
$$P(b)=P(r).P(b/r)+P(v).P(b/v)$$
$$P(b)=0,7.\ 0,2+0,3\,.\,0,4$$
$$P(b)=0,26$$

Ejemplo 19

Consideremos el ejemplo anterior. Si en total hay 250 piezas en la bolsa, tendremos que 175 (70% de 250) son rojas y 75 (30% de 250) son verdes. De las piezas rojas, 35 son bolitas (20% de 175); mientras que de las verdes, 30 son bolitas (40% de 75). Esto nos da un total de 65 bolitas sobre las 250 piezas, es decir que:

$$P(b)=\frac{65}{250}=0,26$$

En la siguiente Tabla, se resumen todas las cantidades de piezas y colores de acuerdo con los datos del ejemplo:

	bolitas	cubitos	**Total colores**
verdes	30	45	75
rojas	35	140	175
Total forma	65	185	**250**

Ejemplo 20

En el año 2005 el gobierno cree que la inflación estará entre el 8,5% y el 10%. Sin embargo, de acuerdo con sus estimaciones, hay una probabilidad de 0,02 de que esté por debajo del mínimo esperado y una probabilidad del 0,20 de que supere el máximo previsto.

Además, dada la relación inversa existente entre el nivel de desempleo y la tasa de inflación, el gobierno prevé que, si la inflación está en los niveles esperados, hay una probabilidad de 0,65 de que el desempleo sea inferior o igual al 13%. Esta probabilidad aumenta a 0,80, si la inflación supera el máximo previsto, y cae a sólo el 0,05, si la inflación está por debajo del mínimo esperado.

De este modo, si denotamos π a la tasa de inflación y u a la tasa de desempleo, entonces la probabilidad de que esta última esté por debajo de 13% es:

$$\begin{aligned} P(u \le 13\%) &= P(u \le 13\% | \pi < 8,5\%) \times P(\pi < 8,5\%) \\ &\quad + P(u \le 13\% | 8,5\% \le \pi \le 10\%) \times P(8,5\% \le \pi \le 10\%) \\ &\quad + P(u \le 13\% | \pi > 10\%) \times P(\pi > 10\%) \\ &= 0,05 \times 0,02 + 0,65 \times (1 - 0,02 - 0,20) + 0,80 \times 0,20 \\ &= 0,668 \end{aligned}$$

1.5.4 Teorema de Bayes

Basado en las probabilidades condicionales y la ley de la probabilidad total, el reverendo Thomas Bayes expuso el siguiente Teorema[17]:

Dado un evento A y n eventos mutuamente excluyentes y colectivamente exhaustivos $B_1, B_2, ..., B_n$, entonces la probabilidad de cualquiera de los eventos B_j condicionado al evento A puede calcularse como:

$$P(B_j | A) = \frac{P(A | B_j) P(B_j)}{\sum_{i=1}^{n} P(A | B_i) P(B_i)}$$

En el Apéndice se encuentra la demostración del Teorema.

Ejemplo 21

En un centro médico especializado en problemas respiratorios, el 80% de los fumadores que se fueron a atender resultó tener cáncer, mientras que de los no fumadores atendidos sólo el 10% tenía cáncer. Se sabe, además, que el 60% de los pacientes no son fumadores. ¿Cuál es la probabilidad de que un paciente con cáncer sea fumador?

Definimos los eventos:

B_1 = *"el paciente es no fumador"*, B_2 = *"el paciente es fumador"*, y A = *"el paciente tiene cáncer"*.

De acuerdo con la información que contamos, conocemos las siguientes probabilidades:

$P(B_1) = 0,60$; $P(B_2) = 0,40$; $P(A | B_1) = 0,10$; $P(A | B_2) = 0,80$

Sobre la base de éstas, podemos hallar la probabilidad deseada, es decir $P(B_2 | A)$. Utilizando el Teorema de Bayes tenemos que:

$$P(B_2 | A) = \frac{P(A | B_2) P(B_2)}{P(A | B_1) P(B_1) + P(A | B_2) P(B_2)}$$

Reemplazando con los datos de la clínica:

$$\begin{aligned} P(B_2 | A) &= \frac{0,80 \times 0,40}{0,10 \times 0,60 + 0,80 \times 0,40} \\ &= 0,8421 \end{aligned}$$

[17] Bayes, T. (1963).

Ejemplo 22

Consideremos el Ejemplo 20. Si el gobierno conoce primero el dato de la tasa de desempleo, y sabe que la misma superó el 13%, entonces las probabilidades de cada uno de los niveles de inflación se modificarían. Si escribimos π_1, π_2 y π_3 para cada uno de los niveles previstos de inflación (menor a 8,5%, entre 8,5% y 10%, y mayor a 10%, respectivamente), entonces las probabilidades buscadas son $P(\pi_1|u>13\%)$, $P(\pi_2|u>13\%)$ y $P(\pi_2|u>13\%)$, las cuales pueden hallarse aplicando la fórmula de Bayes. El denominador de la fórmula puede expresarse como:

$$\sum_{i=1}^{3} P(u>13\%|\pi_i)\times P(\pi_i) = 0{,}95\times 0{,}02 + 0{,}35\times 0{,}78 + 0{,}20\times 0{,}20 = 0{,}332$$

Podemos observar que es simplemente la probabilidad de que el desempleo supere el 13%, es decir, uno menos la probabilidad de que sea menor o igual al 13%, la cual fue hallada en el Ejemplo 20.

Luego, las probabilidades buscadas son:

$$\begin{aligned} P(\pi_1|u>13\%) &= \frac{P(u>13\%|\pi_1)\times P(\pi_1)}{0{,}332} \\ &= \frac{0{,}95\times 0{,}02}{0{,}332} \\ &= 0{,}06 \end{aligned}$$

$$\begin{aligned} P(\pi_2|u>13\%) &= \frac{P(u>13\%|\pi_2)\times P(\pi_2)}{0{,}332} \\ &= \frac{0{,}35\times 0{,}78}{0{,}332} \\ &= 0{,}82 \end{aligned}$$

$$\begin{aligned} P(\pi_3|u>13\%) &= \frac{P(u>13\%|\pi_3)\times P(\pi_3)}{0{,}332} \\ &= \frac{0{,}20\times 0{,}20}{0{,}332} \\ &= 0{,}12 \end{aligned}$$

1.6 Reglas de Conteo

En muchos fenómenos se puede identificar claramente cuántos resultados son posibles al realizar un experimento y cuántos son favorables a cierto evento A, y con dichos valores calcular la probabilidad del evento utilizando la definición clásica. Sin embargo, la tarea al realizar el conteo de casos favorables y casos posibles puede resultar sumamente ardua.

Por ejemplo, consideremos la probabilidad de que al sacar tres cartas de una baraja francesa, dos de ellas sean negras. Para ello, deberíamos contar cuántas combinaciones posibles hay al sacar tres cartas de una baraja francesa, y luego contar cuántas de ellas son favorables al evento "dos son negras". Esta tarea sería muy engorrosa si no se utilizan las reglas de conteo que se exponen en esta sección.

Al momento de trabajar con reglas de conteo, un factor importante a considerar es la relevancia del orden en el cual suceden las observaciones. De esta manera, dependiendo de si el orden altera o no el resultado del experimento se estará trabajando con reglas distintas. Básicamente, las reglas de conteo son las variaciones, permutaciones y combinaciones. Antes de abordar el detalle de cada una de ellas, debe tenerse en cuenta las diferencias principales entre las mismas: en las combinaciones el orden es irrelevante y el resultado depende de los elementos que conformen la observación; en las **variaciones**, por el contrario, dos observaciones representan resultados distintos a pesar de tener los mismos elementos si el orden en el cual los mismos se presentan varía. Finalmente, al trabajar con **permutaciones** se evalúan las distintas alternativas para ordenar un grupo de elementos.

1.6.1 Variaciones y Permutaciones

Consideremos dos lanzamientos consecutivos de una moneda. Los resultados posibles, considerando el orden en que ocurren, son cuatro:

$$\Omega = \{CaCa; CaCe; CeCa; CeCe\}$$

Consideremos ahora tres lanzamientos consecutivos, entonces hay ocho resultados posibles:

$$\Omega = \{CaCaCa; CaCaCe; CaCeCa; CeCaCa; CaCeCe; CeCaCe; CeCeCa; CeCeCe\}$$

Consideremos 5 lanzamientos, o 10 lanzamientos, o, más aún, 20 lanzamientos. La tarea de contar uno por uno todos los posibles resultados sería muy complicada ¿no? Para contar la cantidad de resultados posibles en estos casos se utilizan las variaciones.

Cuando un fenómeno puede ocurrir de n maneras distintas (hay n resultados posibles), y el mismo se repite r veces, la cantidad total de resultados distintos que se pueden obtener (considerando el orden en que ocurre el resultado de cada ensayo) es una variación de n elementos tomados de r en r :

$$V_{(n,r)} = n^r$$

Ejemplo 23

Si lanzamos una moneda al aire hay dos resultados posibles ($n = 2$), cara o ceca Si lanzamos 2 veces consecutivas una moneda ($r = 2$), los resultados posibles son $V_{(2;2)} = 2^2 = 4$.

Si se lanzan 3 veces, entonces hay $V_{(2;3)} = 2^3 = 8$ resultados posibles.

Si se realizan 20 lanzamientos, habrá $V_{(2;20)} = 2^{20} = 1.048.576$ posibles resultados (teniendo en cuenta el orden en que ocurren las caras y las cecas obtenidas).

Ejemplo 24

En los ejemplos anteriores hemos visto que si lanzamos 2 veces un dado, hay 36 resultados posibles si se tienen en cuenta el orden en que ocurren los números (es decir, un 6 y un 1 no es lo mismo que un 1 y un 6). Esta cantidad no es ni más ni menos que las variaciones de 6 tomados de 2 en 2:

$$V_{(6;2)} = 6^2 = 36$$

Por otra parte, hay ocasiones en que se combinan distintos fenómenos. Por ejemplo, lanzamos un dado y una moneda y queremos analizar cuántos posibles resultados se obtienen. Estas circunstancias generan la segunda regla de conteo.

Si hay r fenómenos donde el primero posee n_1 resultados posibles, el segundo n_2 resultados posibles,…, y el r-ésimo n_r resultados posibles, entonces el número total de resultados distintos que se pueden obtener al combinar los r fenómenos es:

$$n_1 \times n_2 \times ... \times n_r$$

Ejemplo 25

Si se lanza una moneda ($n_1 = 2$) y un dado ($n_2 = 6$), la cantidad de resultados posibles es:

$$2 \times 6 = 12$$

Este resultado es bastante intuitivo, considerando que puede ocurrir “cara” con cada uno de los seis resultados del dado y “ceca” con cada uno de los mismos.

Ejemplo 26

Si se lanza una moneda ($n_1 = 2$), dos dados ($n_2 = n_3 = 6$) y se extrae una carta de una baraja española ($n_4 = 40$), la cantidad de resultados posibles es:

$$2 \times 6 \times 6 \times 40 = 2.880$$

Un caso particular de la segunda regla de conteo mencionada es lo que se denomina *Variación sin repetición*. En ese caso, lo que se considera es que el fenómeno sujeto a experimentación es siempre el mismo pero los eventos, una vez que suceden, no vuelven a ocurrir. Es decir que, con cada repetición del experimento, el número de eventos posibles disminuye en uno respecto de los posibles casos del experimento anterior. Por lo tanto:

$$n_1 = n \;\; ; \;\; n_2 = n-1 \; ; \; n_3 = n-2 \; ..... \;\; n_r = n-(r-1)$$

Cuando un fenómeno puede ocurrir de n maneras distintas (hay n resultados posibles), el mismo se repite r veces y, además, una vez obtenido un resultado determinado el mismo no puede volver a darse, la cantidad total de resultados distintos (considerando el orden en que ocurre el resultado de cada ensayo) genera las **variaciones sin repetición** de n elementos tomados de a r:

$$\begin{aligned} VR_{(n,r)} &= n \cdot (n-1) \cdot (n-2) \cdot ..... \cdot (n-(r-1)) \\ &= \frac{n!}{(n-r)!} \end{aligned}$$

Ejemplo 27

Podemos considerar que una persona posee cuatro CDs de música distintos y decide llevar en sus vacaciones sólo dos de ellos. Además, decide escucharlos en su automóvil en el mismo orden en el cual los selecciona. A su vez, en función de la duración del viaje, difícilmente terminará de escuchar el segundo disco. Como no tiene preferencias entre esos discos, la elección de los mismos la realizará al azar. ¿Cuántas combinaciones posibles de discos a escuchar tiene la persona?

En este caso podríamos fácilmente calcularlo por extensión, obteniendo las siguientes doce combinaciones:

$$\{AB; AC; AD; BA; BC; BD; CA; CB; CD; DA; DB; DC\}$$

Llegaríamos al mismo resultado con la fórmula expuesta, remplazando $n = 4$ ya que hay 4 CDs y $r = 2$ porque se seleccionarán dos de ellos (sin repetición):

$$VR_{(4,2)} = \frac{4!}{(4-2)!} = \frac{4!}{2!} = 12$$

Ejemplo 28

Si se toma un mazo de barajas españolas (40 cartas) y quiere saberse la cantidad de maneras posibles que existe de tomar dos cartas diferentes (considerando importante el orden en el cual sean seleccionamos las mismas). Debe tenerse en cuenta que, una vez tomada una carta del mazo original, la misma ya no formará parte de él. El cálculo intuitivo de la cantidad de posibilidades en este caso no es sencillo. Sin embargo, podemos realizar el cálculo deseado utilizando la fórmula de *variaciones sin repetición* con $n = 40$ y $r = 2$, obteniendo un total de ¡1560 combinaciones posibles!:

$$VR_{(40,2)} = \frac{40!}{(40-2)!} = \frac{40!}{38!} = 1560$$

Es muy importante tener en cuenta que en las reglas analizadas hasta aquí es importante el orden en que ocurren los eventos. Es decir que, por ejemplo, al lanzar dos veces una moneda, no es lo mismo $CaCe$ que $CeCa$, o al lanzar dos dados, no es lo mismo un dos y un tres que un tres y un dos.

La segunda regla de conteo que analizaremos, y que también considera el orden de los resultados es la **permutación**. Ésta considera las distintas maneras de ordenar un grupo de

elementos. Uno de los casos más sencillos es el caso de la **permutación simple**. Lo que se refleja en este caso es la cantidad de maneras en las que puede ordenarse un grupo de n elementos:

Si se poseen n elementos, la cantidad de maneras de ordenarlos es:

$$n! = n \times (n-1) \times (n-2) \times ... \times 2 \times 1$$

Puede verse que esta alternativa de cálculo es equivalente al de una *variación sin repetición* en donde el número de experimentos, r, es igual al número de resultados posibles para el primero de ellos. Es decir:

$$VR_{(n;n)} = \frac{n!}{(n-n)!} = n!$$

Esta equivalencia es lógica dado que las distintas formas de ordenar el grupo constituyen los distintos eventos para la variación.

Ejemplo 29

Puede considerarse, a modo de ejemplo, el caso en el cual cinco deportistas deban realizar una prueba. El orden en el cual cada uno de ellos la efectúe depende de un sorteo el cual consiste en retirar de una urna el nombre de cada uno de ellos. La cantidad de maneras de ordenar a estos deportistas es entonces 120 y está dado por: $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

Una alternativa a la permutación simple es la de considerar la cantidad de muestras ordenadas distintas que pueden obtenerse de un grupo. Este concepto es similar al de la *variación sin repetición*, y la fórmula de cálculo es la misma:

Si se extraen r elementos de un conjunto de n, la cantidad de muestras **ordenadas** distintas que pueden obtenerse es la **permutación** de n tomados de a r:

$$P_{(n;r)} = VR_{(n;r)} = \frac{n!}{(n-r)!}$$

Ejemplo 30

Continuando con el ejemplo anterior, podría darse el caso en que el primer día realicen la prueba sólo tres de los cinco deportistas. ¿Cuántas alternativas distintas de deportistas seleccionados y orden en el que se realizarán las pruebas existen? Este cálculo equivale a determinar la permutación de 5 elementos (los deportistas) tomados de a 3 (tres): $P_{(5;3)} = \frac{5!}{(5-3)!} = 60$

En ocasiones se presentan casos en los cuales el orden pierde importancia, por ejemplo si queremos saber solamente la suma de los dados, o la cantidad de cecas que salen. En estos casos las reglas de conteo cambian, de acuerdo a lo que se verá en el siguiente apartado.

1.6.2 Combinatorias

Según hemos hecho referencia en el párrafo anterior, hay casos en los cuales no resulta relevante el orden en el cual se dan los resultados, sino cuáles son esos resultados en sí. Por ejemplo, en el caso en que lancemos un dado dos veces de manera tal que avancemos en un juego tantos casilleros como indica la suma de ellos, el orden de los resultados no resultará relevante: si obtenemos un 5 y luego un 2 significará lo mismo que obtener un 2 y luego un 5; en ambos casos avanzaremos 7 casilleros. Cuando trabajamos con variaciones o permutaciones, el orden resulta relevante: por ejemplo, en el caso en que en el juego en cuestión deban cumplirse las *"prendas"* relativas al casillero al cual nos lleve el primer dado.

Cuando se trabaja con **combinatorias** lo que se busca calcular es la cantidad de grupos distintos de *r* elementos que pueden formarse con los *n* elementos que conforman un conjunto.

Si se extraen r elementos de un conjunto de n, la cantidad de muestras distintas que pueden obtenerse (**sin importar el orden**) es la **combinatoria** de n elementos tomados de a r:

$$C_{(n;r)} = \frac{n!}{(n-r)!r!}$$

La combinatoria de n tomados de a r suele escribirse como:

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Ejemplo 31

Si se considera el Ejemplo 30, con la combinatoria puede calcularse cuántos grupos de deportistas distintos realizarían la prueba el primer día de la competición. En esta situación no resulta relevante el orden en el que participarán los tres deportistas seleccionados sino cuáles son los mismos. La cantidad de grupos distintos que deberán realizar la prueba el primer día es la combinatoria de cinco elementos tomados de a 3:

$$C_{(5;3)} = \frac{5!}{(5-3)!3!} = 10$$

Es decir, que hay diez grupos distintos de tres deportistas que debieran realizar la prueba el primer día.

Ejemplo 32

Dado un grupo de cien lamparitas, quince de ellas resultan ser defectuosas. ¿Cuál es la probabilidad de que, tomando dos lamparitas al azar, las dos resulten ser defectuosas?

Lo primero que debemos calcular, de acuerdo a la definición clásica, es la cantidad de eventos posibles: es decir, cuántos conjuntos de dos lamparitas pueden formarse. En este caso, tomamos dos lamparitas ($r = 2$) de entre cien ($n = 100$):

$$Casos\ Posibles = C_{(100;2)} = \frac{100!}{98!2!} = 4950$$

Los casos favorables son la cantidad de grupos de dos lamparitas que pueden formarse sólo considerando aquellas defectuosas:

$$Casos\ Favorables = C_{(15;2)} = \frac{15!}{13!2!} = 105$$

La probabilidad entonces de tomar dos lamparitas defectuosas es

$$\frac{105}{4950} = 0,021$$

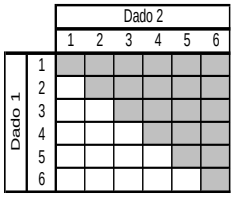
Puede también considerarse la **combinación** de distintos elementos existiendo la posibilidad de reposición. Por ejemplo, para el caso de la suma que se obtiene al lanzar dos veces un mismo dado, el hecho de que en el primer lanzamiento haya salido un dos no invalida que el segundo resultado sea también un dos.

Si se consideran r elementos de un conjunto de n, la cantidad de muestras distintas que pueden obtenerse (**sin importar el orden**) en caso de que la obtención de un resultado no invalide nuevamente su ocurrencia es:

$$CR_{(n;r)} = \frac{(n+r-1)!}{(n-1)!r!}$$

Ejemplo 33

Si se considera la cantidad de combinaciones posibles que surgen de dos lanzamientos de un dado, independientemente del orden, los resultados que serán distintos son los sombreados con gris en el cuadro siguiente, es decir, 21 combinaciones diferentes.



Prescindiendo de la representación gráfica, podría haberse utilizado la fórmula anterior con $n=6$ (resultados posibles en un lanzamiento) y $r=2$ (cantidad de lanzamientos):

$$CR_{(6;2)} = \frac{(6+2-1)!}{(6-1)!2!} = 21$$

1.7 Apéndice: Demostraciones

1.7.1 Conclusiones de la Axiomática

1. $P\left(A^C\right) = 1 - P(A)$

De acuerdo con la definición del complemento, tenemos que $A \cup A^C = \Omega$, por lo cual:

$$P\left(A \cup A^C\right) = P(\Omega)$$

Según el Axioma b) expuesto en la sección 1.3, $P(\Omega) = 1$, y dado que *A* y A^C son mutuamente excluyentes podemos usar el Axioma c) para escribir $P\left(A \cup A^C\right) = P(A) + P\left(A^C\right)$.

Igualando ambas expresiones resulta:

$$P(A) + P(A^c) = P(\Omega)$$
$$P(A) + P(A^c) = 1$$
$$P(A) = 1 - P(A^c)$$

2. $0 \le P(A) \le 1$

De acuerdo al Axioma a) expuesto en la sección 1.3, tenemos que para cualquier evento *A* se verifica que $P(A) \ge 0$. A su vez, por la demostración anterior $P(A) + P\left(A^C\right) = 1$, por lo cual, indefectiblemente $P(A) \le 1$.

Combinando las dos desigualdades del párrafo anterior, obtenemos lo que queríamos demostrar.

3. $P(\varnothing)=0$

De acuerdo con la definición de evento vacío, tenemos que $\Omega\cup\varnothing=\Omega$ y $\Omega\cap\varnothing=\varnothing$. Como la intersección es vacía y la unión es el espacio muestral, entonces $\varnothing$ y Ω son eventos complementarios, es decir $\varnothing=\Omega^C$.

A su vez, sabemos que $P(\Omega)+P(\Omega^C)=1$. Pero por el Axioma b) $P(\Omega)=1$, por lo cual $P(\Omega^C)=P(\varnothing)=0$.

4. $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

Primero, escribimos la unión de *A* y *B* como:

$$A\cup B=(A\cap B)\cup(A\cap B^C)\cup(B\cap A^C) \qquad (1)$$

A su vez, podemos escribir los eventos *A* y *B* como:

$$\begin{aligned} A&=(A\cap B)\cup(A\cap B^C)\\ B&=(B\cap A)\cup(B\cap A^C)\end{aligned} \qquad (2)$$

Observamos que los eventos entre paréntesis de (1) y (2) son mutuamente excluyentes, por lo cual podemos aplicar el Axioma c) para escribir:

$$\begin{aligned} P(A\cup B)&=P(A\cap B)+P(A\cap B^C)+P(B\cap A^C)\\ P(A)&=P(A\cap B)+P(A\cap B^C)\\ P(B)&=P(B\cap A)+P(B\cap A^C)\end{aligned} \qquad (3)$$

Si tomamos la primer expresión de (3) y restamos de ésta, miembro a miembro, la segunda y la tercera ecuación, obtenemos:

$$P(A\cup B)-P(A)-P(B)=-P(A\cap B)$$

Despejando, $P(A\cup B)$ obtenemos lo que queríamos demostrar.

5.

$$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$$

La demostración es similar a la anterior y la dejamos como ejercicio para el lector

6. $A\subseteq B \quad\Rightarrow\quad P(A)\le P(B)$

Pude observarse que:

$$A\subseteq B\Rightarrow(A\cup B)=B;\ (A\cap B)=A;(A\cap B^C)=\varnothing$$

Entonces, utilizando la expresión de (3), puede escribirse:

$$P(A\cup B)=P(B)=P(A\cap B)+P\left(A\cap B^{C}\right)+P\left(B\cap A^{C}\right)$$
$$P(B)=P(A)+P(\varnothing)+P\left(B\cap A^{C}\right)$$
$$P(B)=P(A)+P\left(B\cap A^{C}\right)$$

Como, por Axioma, $P\left(B\cap A^{C}\right)\geq 0 \Rightarrow P(B)\geq P(A)$

7. $A\subseteq B \quad \Rightarrow \quad P(A\cap B)=P(A)$

Si $A\subseteq B \Rightarrow (A\cap B)=A$, de donde se deduce que $P(A\cap B)=P(A)$

1.7.2 Probabilidad Total

Si D_j $(j=1,2,...,n)$ son eventos mutuamente excluyentes y colectivamente exhaustivos, entonces tenemos que:

$$\bigcup_{j=1}^{n} D_j = D_1\cup D_2\cup ...\cup D_n=\Omega \qquad (4)$$

A su vez, de acuerdo con las propiedades de eventos, se tiene que:

$$A=A\cap\Omega$$

Con lo cual, remplazando Ω por la expresión (4), y usando las propiedades de intersección e unión de conjuntos, tenemos que:

$$\begin{aligned} A &= A\cap(D_1\cup D_2\cup ...\cup D_n) \\ &= (A\cap D_1)\cup(A\cap D_2)\cup ...\cup(A\cap D_n) \\ &= \bigcup_{j=1}^{n}(A\cap D_j) \end{aligned}$$

Luego,

$$P(A)=P\left[\bigcup_{j=1}^{n}(A\cap D_j)\right] \qquad (5)$$

Asimismo, como los eventos entre corchetes en el miembro derecho de (5) son mutuamente excluyentes, podemos usar el Axioma c):

$$P\left[\bigcup_{j=1}^{n}(A\cap D_j)\right]=\sum_{j=1}^{n}P(A\cap D_j) \qquad (6)$$

Finalmente, igualando (5) con (6), obtenemos la ley de la probabilidad total:

$$P(A)=\sum_{j=1}^{n}P(A\cap D_j)$$

1.7.3 Teorema de Bayes

De acuerdo con la definición de la probabilidad condicional, tenemos que:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad siendo \qquad P(B) > 0$$

o bien, considerando un evento cualquiera B_j condicionado a otro evento *A*, la fórmula precedente se puede escribir como:

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} \qquad siendo \qquad P(A) > 0 \qquad (7)$$

Por otro lado, si consideramos el evento *A* condicionado a B_j, tenemos que:

$$P(A|B_j) = \frac{P(A \cap B_j)}{P(B_j)} \qquad \Rightarrow \qquad P(A|B_j)P(B_j) = P(A \cap B_j) \qquad (8)$$

siendo $P(B_j) > 0$. A su vez, como $P(B_j \cap A) = P(A \cap B_j)$, podemos remplazar la expresión (8) en el numerador de (7):

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} \qquad siendo \qquad P(A) > 0 \qquad (9)$$

Luego, si B_j pertenece a un grupo de eventos mutuamente excluyentes y colectivamente exhaustivos, y de acuerdo con la ley de probabilidad total, tenemos que la probabilidad del evento *A* se puede escribir como:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \qquad (10)$$

Finalmente, reemplazando (10) en el denominador de (9), se obtiene el Teorema de Bayes:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

2 Variables Aleatorias y distribuciones de probabilidad

Dario Bacchini
Lara Vazquez
Valeria Gogni

En el capítulo anterior se ha desarrollado la Teoría de la Probabilidad, una de las piedras angulares de la estadística. Esta Teoría está estrechamente relacionada con las variables aleatorias, que serán presentadas en este capítulo, y que permitirán de alguna manera cuantificar los experimentos aleatorios con los que trabajemos.

Al realizar un experimento aleatorio, sus resultados posibles, es decir, los elementos del espacio muestral, pueden ser cualitativos o cuantitativos. Por ejemplo, si arrojamos un dado, los posibles resultados son cuantitativos, ya que podemos asignar fácilmente una cantidad numérica a cada uno de ellos; en cambio, si arrojamos una moneda al aire, los resultados "cara" o " cecas" no tienen una relación obvia con un conjunto de números, por lo cual este experimento es cualitativo.

Las variables aleatorias asignan valores numéricos a cada posible resultado de un espacio muestral, ya sea éste cualitativo o cuantitativo, y su uso es de gran utilidad cuando se desean realizar estudios cuantitativos relacionados con ciertos fenómenos.

2.1 Definición

Una variable aleatoria asigna un valor numérico a cada resultado posible del espacio muestral de un fenómeno o experimento aleatorio. En base a ello, podemos formalizar la definición.

Una variable aleatoria $X(\omega)$ es una función que asigna un número real a cada posible resultado ω de un espacio muestral Ω. Es decir que:

$$X: \Omega \rightarrow A \quad ; \quad A \subseteq \mathbb{R} \qquad (1)$$

En la definición puede observarse que los valores que toma la variable aleatoria son un subconjunto de los números reales. Sin embargo, si se trata de un fenómeno cuyos resultados posibles constituyen un conjunto numerable (ya sea finito o infinito), los valores de la variable aleatoria serán un subconjunto de los números enteros, $\mathbb{Z}$, o de los números naturales, $\mathbb{N}$.

Ejemplo 1

Consideremos el lanzamiento de una moneda. Podemos asignar al resultado ceca el valor cero, y al resultado cara, el valor uno. El espacio muestral es $\Omega = \{\omega_1; \omega_2\} = \{cara; ceca\}$, y la variable aleatoria definida sobre el mismo es $X(ceca) = 0; X(cara) = 1$.

Desde ya, que, en este caso, la especificación es totalmente arbitraria, teniendo presente que el lanzamiento de una moneda es un experimento intrínsecamente cualitativo. Por ejemplo, se podría haber optado por definir a la variable aleatoria como $X(ceca) = 1; X(cara) = -1$.

Ejemplo 2

Consideremos el lanzamiento de un dado. El espacio muestral es $\Omega = \{1,2,3,4,5,6\}$. Al ser el espacio muestral intrínsecamente cuantitativo, la variable aleatoria más lógica para definir, llamémosla X, sería asignar a la misma el resultado del espacio muestral. Es decir que, en este caso, tenemos que $\Omega = A$, en términos de la definición expuesta más arriba.

Sin embargo, se podría definir otra variable, digamos Y, que asigne 1 a los resultados pares y 0 a los impares. En la siguiente tabla, se exponen ambas variables.

ω	$X(\omega)$	$Y(\omega)$
1	1	0
2	2	1
3	3	0
4	4	1
5	5	0
6	6	1

Ejemplo 3

Si deseamos estudiar las estaturas de un grupo de personas, los resultados serán aleatorios en la medida que seleccionemos al azar las personas a medir. En este caso, al tratarse de una variable cuantitativa, el espacio muestral coincidirá con los valores que pueda tomar la variable ($\Omega = A$).

Los valores que puede tomar la variable son de 0 a 3 metros de altura (exagerando los límites superior e inferior), es decir, que $A = (0;3)$, un subconjunto de los números reales.

En la definición expuesta precedentemente, hemos mencionado que los valores que puede tomar una variable aleatoria están dados por los elementos de un conjunto A. Además, destacamos que A puede ser un subconjunto de los números reales, o un subconjunto de los números enteros (o quizás de los naturales). Esto nos lleva a la distinción entre variables continuas y discretas.

Variables aleatorias discretas y continuas

En los Ejemplos 1 y 2, se definieron variables cuyos valores posibles formaban un conjunto discreto, mientras que en el Ejemplo 3 se definió una variable aleatoria que podía tomar valores dentro de un intervalo de la recta real. Esta distinción que realizamos en cuanto a los posibles valores que puede tomar una variable aleatoria nos lleva a diferenciar entre variables aleatorias continuas y discretas.

Un espacio muestral Ω es **discreto** si es numerable, ya sea finito o infinito. Por otro lado, Ω es **continuo** si sus elementos forman un conjunto infinito no numerable.

Ejemplo 4

Como se ha ilustrado en los ejemplos precedentes, el lanzamiento de un dado, o el lanzamiento de una moneda, claramente constituyen experimentos aleatorios cuyos espacios muestrales son discretos, ya que la cantidad de resultados posibles es finita y numerable en ambos casos (6 en el dado y 2 en la moneda).

Por otro lado, la cantidad de mililitros de lluvia que cae en un año en una zona subtropical constituye un fenómeno aleatorio, y en principio su resultado puede ser cualquier número real positivo. El espacio muestral, en este caso, es continuo, ya que los resultados incluidos en el mismo forman un conjunto infinito no numerable.

Asimismo, podemos clasificar a las variables aleatorias en discretas o continuas, de acuerdo a la imagen de la función que las define, es decir, de acuerdo con los elementos del conjunto A.

Una **variable aleatoria** es **discreta** si la imagen de la misma está constituida por un conjunto numerable:

$$X : \Omega \to A \qquad (A \text{ es numerable})$$

Ejemplo 5

Las variables aleatorias definidas en los Ejemplo 1 y 2 son discretas, ya que el conjunto imagen es un subconjunto de los números enteros y, por lo tanto, es numerable.

Ejemplo 6

Consideremos un juego en el cual se lanza una moneda al aire tres veces consecutivas, y se paga $ 0,50 al jugador por cada ceca que sale. En la siguiente tabla, se exponen los elementos del espacio muestral y los valores correspondientes de la variable aleatoria:

ω	$X(\omega)$
CaCaCa	$ 0,00
CaCaCe	$ 0,50
CaCeCe	$ 1,00
CeCeCe	$ 1,50

Siendo los posibles valores un conjunto numerable finito, la variable aleatoria es discreta.

En el ejemplo anterior, no consideramos el orden en que se consiguió cada resultado, ya que, a los efectos de la apuesta, lo único que importa es cuántas cecas se obtuvieron. Por ejemplo, la obtención de $CaCeCa$ y $CeCaCa$, arroja el mismo resultado de la variable, $0,50.-, que $CaCaCe$. Sin embargo, cuando analicemos las probabilidades asociadas a cada valor de la variable aleatoria, tendremos que analizar las distintas maneras en que se pueden presentar cada uno de ellos.

Antes de proseguir, remarcamos algo sumamente importante. **No necesariamente los valores de una variable discreta son números enteros (o naturales), sino que lo que importa es que todos los resultados constituyan un conjunto numerable (finito o infinito).**

En el ejemplo anterior, claramente, los valores no son enteros. Sin embargo, se puede enumerar cada uno de ellos, es decir, que podemos contabilizar cada resultado posible. Lo que importa es que haya una correspondencia uno a uno con los números naturales, esto es, que el conjunto sea numerable.

Una **variable aleatoria** es **continua** si la imagen de la misma está constituida por un intervalo de los números reales:

$$X: \Omega \to (a,b) \ ; \ (a,b) \subseteq \mathbb{R}$$

Ejemplo 7

La variable definida en el Ejemplo 3, "estatura", es continua, ya que los posibles valores que puede tomar la misma constituyen un intervalo de la recta real.

Ejemplo 8

Consideremos la cantidad de mililitros de lluvia por metro cuadrado que cae en un año en una zona subtropical (Ejemplo 4). Podemos definir una variable aleatoria, asignando a la misma el valor observado en el espacio muestral (es decir, que, si cayeron 500 mililitros, la variable toma el valor 500). Claramente, la variable es continua, ya que sus valores posibles forman un conjunto infinito no numerable.

Finalmente, hacemos una última aclaración en cuanto a los espacios muestrales (dominio) y los valores que puede tomar una variable aleatoria (imagen). En principio, en el caso de espacios muestrales cuantitativos, uno está tentado a asociar los valores del mismo a los valores de la

variable. Sin embargo, como vimos en el Ejemplo 2, esto no necesariamente es así. Además, es importante destacar que **un espacio muestral continuo puede tener asociada una variable aleatoria discreta**. ¿Le parece raro? Considere el siguiente ejemplo.

Ejemplo 9

En los Ejemplos 4 y 8 se ha considerado la cantidad de mililitros de lluvia por metro cuadrado que cae en un año en una zona subtropical. En el caso de Bolivia, por ejemplo, las precipitaciones anuales son de 555mm.[18] Puede considerarse, entonces, el caso en el que el gobierno entregue un determinado subsidio a los productores si el nivel de lluvias es inferior a los 480 mm. (z dólares) o superior a los 630 mm. (w dólares). En este caso, la **variable aleatoria discreta** monto del subsidio (X) puede tomar únicamente los valores $\{0\ ;\ z\ ;\ w\}$. Esta variable se encuentra asociada a un **espacio muestral continuo** dado por el nivel de lluvias anual que se presente. De esta manera:

ω	$X(\omega)$
< 480 mm	u$s z
480 mm - 630 mm	u$s 0
> 630 mm	u$s w

Notación

Antes de continuar, hacemos un paréntesis para remarcar ciertas cuestiones referidas a la notación que utilizaremos de aquí en adelante.

En general, y siguiendo la convención de la mayoría de los autores, las variables aleatorias serán expresadas por letras mayúsculas (X, Y, W, etc), mientras que los valores particulares que puedan asumir las variables, serán denotados con letras minúsculas (x, y, w, etc).

En consecuencia, y recordando la notación de Teoría de Probabilidades vista en el capítulo anterior, la probabilidad de que una variable aleatoria tome un valor determinado se expresará como $P(X = x)$, siendo x un valor particular del dominio. En el caso del Ejemplo 9, podríamos calcular la probabilidad de que no exista pago de subsidio, lo cual se expresaría con la siguiente notación: $P(X = 0)$, siendo 0 el valor particular considerado.

2.2 Distribución De Probabilidades

En la sección anterior hemos definido a las variables aleatorias, las cuales asignan un valor numérico a los distintos eventos de un espacio muestral. De esta manera, podemos definir, entonces una función real que asigna una probabilidad a cada valor que puede tomar la variable aleatoria.

Expresemos lo mismo en otras palabras. En el capítulo anterior, hemos mencionado que a un experimento aleatorio se le puede asignar una medida de probabilidad, la cual debía cumplir con ciertos axiomas. Así, y de acuerdo a la relación biunívoca existente entre los fenómenos aleatorios y las variables aleatorias definidas sobre los mismos, lógicamente a cada valor que tome la variable se le podrá asignar la probabilidad relacionada con el evento subyacente en el valor de la variable.

En base a lo expresado anteriormente, podemos enunciar la siguiente definición:

[18] "El Libro del Mundo", Arte Gráfico Editorial Argentino (1997)

La **distribución de probabilidades** de una variable aleatoria es el conjunto de todos los valores que puede tomar la misma y sus respectivas probabilidades.

Para ejemplificar y considerar con mayor profundidad esta definición, analizaremos por separado el caso de variables aleatorias discretas y continuas.

2.2.1 Función de Probabilidad de variables discretas

Empecemos con un ejemplo de una variable aleatoria discreta definida sobre un espacio muestral discreto.

Ejemplo 10

Definimos la siguiente variable aleatoria relacionada con el lanzamiento de una moneda: $X(ceca)=0; X(cara)=1$, claramente podremos establecer la probabilidad asociada a cada valor que toma la misma como:

$$P(X=0)=P(ceca)=0{,}5 \text{ ; y } P(X=1)=P(cara)=0{,}5$$

Teniendo presente lo mencionado al inicio de esta sección, y con el ejemplo anterior como herramienta para clarificar la manera de asignar probabilidades a cada valor de una variable aleatoria, podemos realizar la siguiente definición:

Sea *X* una variable aleatoria discreta definida sobre un espacio muestral Ω. Su función de probabilidad es una función que cumple las siguientes condiciones:

$$a)\ \forall x \in A: \quad p(x) \geq 0$$

$$b)\ \sum_{x \in A} p(x) = 1 \qquad (2)$$

donde A, como siempre, es el conjunto numérico que representa todos los posibles valores que puede asumir la variable aleatoria.

De acuerdo a la expresión (1) expuesta al inicio de la Sección 1, la expresión (2), y a las propiedades vistas en el anterior, respecto de las medidas de probabilidad, podemos expresar lo siguiente:

$$X : \Omega \rightarrow A$$
$$p : A \rightarrow [0;1]$$

Es decir que, dado un espacio muestral Ω, podemos definir una función X que asigna a cada resultado (simple o compuesto) un número perteneciente a un conjunto numérico *A*. A su vez, podemos definir una función *p* que asigna a cada valor de *X* (perteneciente al conjunto *A*) un número contenido en el intervalo de la unidad (teniendo presente que la suma de *p* sobre todos los posibles valor de *X* debe ser uno).

De acuerdo con ello, se puede observar que estamos frente a una composición de funciones en la cual *p* depende en última instancia del resultado ω del espacio muestral Ω:

$$p\left[X(\omega)\right]$$

De este modo, como la aleatoriedad está presente en el evento ω, y *X* es simplemente una transformación de dicho evento, el punto b) de (2) podría escribirse como:

$$P(\Omega) = \sum_{\omega \in \Omega} p\left[X(\omega)\right] = \sum_{x \in A} p(x) = 1$$

Esta expresión relaciona el punto b) de la expresión (2) con el Axioma b) de la sección 1.3 del capítulo anterior.

Clarifiquemos la definición y los conceptos expuestos mediante un simple ejemplo.

Ejemplo 11

Consideremos un juego en el cual se lanza 2 veces una moneda, y el apostador gana $1 por cada cara que salga.

El espacio muestral está dado por los posibles resultados que se obtengan al lanzar dos veces una moneda: CeCe, CaCe, CeCa, CaCa y la probabilidad de cada uno de ellos es de $(1/2)^2 = 1/4$. A su vez, los valores de la variable aleatoria "ganancia del apostador" correspondientes a cada resultado del espacio muestral son $X(CeCe) = 0$, $X(CaCe) = 1$, $X(CeCa) = 1$ y $X(CaCa) = 2$.

De esta manera, las probabilidades asociadas a cada posible ganancia del apostador dependen en última instancia del resultado del espacio muestral:

$$P(X = 0) = P[X(\omega) = 0] = P(\omega = CeCe) = 1/4$$

$$P(X = 1) = P[X(\omega) = 1] = P(\omega = CaCe \text{ ó } \omega = CeCa) = 1/4 + 1/4 = 1/2$$

$$P(X = 2) = P[X(\omega) = 2] = P(\omega = CaCa) = 1/4$$

A su vez, podemos comprobar que la suma de las probabilidades sobre todos los posibles valores de la variable aleatoria y sobre todos los posibles resultados del espacio muestral es 1:

$$\sum_{k=0}^{2} P(X = k) = P(X = 0) + P(X = 1) + P(X = 2) = 1/4 + 1/2 + 1/4 = 1$$

$$\sum_{\omega \in \Omega} P(\omega) = P(CeCe) + P(CaCe) + P(CeCa) + P(CaCa) = 1/4 + 1/4 + 1/4 + 1/4 = 1$$

Si queremos saber la probabilidad de que la variable aleatoria tome ciertos valores en un intervalo determinado, no tenemos más que sumar las probabilidades de cada uno de los valores de ese intervalo.

La probabilidad de que *X* pertenezca a un intervalo $[a;b]$ incluido entre sus posibles valores, está dada por[19]:

$$P(a \le X \le b) = \sum_{x=a}^{b} p(x)$$

Ejemplo 12

Consideremos la siguiente variable aleatoria X y su correspondiente función de probabilidad[20]:

x	1	2	3	4	5	6	7	8	9	10
$P(X = x)$	0,05	0,06	0,07	0,10	0,20	0,18	0,17	0,10	0,04	0,03

Nótese antes que nada que:

$$\sum_{x=1}^{10} P(X = x) = 1$$

En base a la tabla, podemos calcular la probabilidad de que la variable se encuentre entre los valores 2 y 6, ambos inclusive:

$$P(2 \le X \le 6) = \sum_{x=2}^{6} p(x) = 0,06 + 0,07 + 0,10 + 0,20 + 0,18 = 0,61$$

Asimismo, si no incluimos los extremos del intervalo, tenemos la siguiente probabilidad:

[19] Notamos que el intervalo se escribió utilizando corchetes para indicar que se incluyen ambos extremos del mismo.
[20] Omitimos los resultados del espacio muestral que generan la variable aleatoria.

$$P(2 < X < 6) = \sum_{x=3}^{5} p(x) = 0{,}07 + 0{,}10 + 0{,}20 = 0{,}37$$

En el ejemplo anterior, se destaca la necesidad de definir muy precisamente qué valores de la variable son relevantes al calcular la probabilidad, ya que al excluir los extremos del intervalo la probabilidad se modifica notablemente.

Hasta aquí los conceptos básicos relacionados con las probabilidades vinculadas a variables aleatorias discretas. A continuación, analizaremos cómo asignar probabilidades a variables aleatorias continuas, y luego, estudiaremos algunas propiedades comunes a ambos tipos de variables.

2.2.2 Función de Densidad de variables continuas

Cuando el espacio muestral es continuo, no tendremos asociado a cada posible resultado una probabilidad. Pensemos en la definición clásica: tendremos un resultado favorable e infinitos resultados posibles. Así, la probabilidad de un resultado particular será cero.

En el caso de una variable aleatoria continua, tenemos una *función de densidad* que nos permitirá calcular la probabilidad asociada a un intervalo de valores de la variable aleatoria.

Sea X una variable aleatoria definida sobre un espacio muestral continuo Ω, cuyos posibles valores pertenecen al conjunto A incluido en el conjunto de los números reales ($A \subseteq \mathbb{R}$). Su función de densidad es una función que cumple las siguientes condiciones:

$$1. f(x) \geq 0$$
$$2. \int_{x \in A} f(x)\,dx = 1$$

Los comentarios realizados para las variables aleatorias discretas, referidos a la composición de funciones desde el espacio muestral al intervalo de la unidad, son válidos también en este caso, y remitimos al lector a lo expuesto más arriba.

Hemos dicho que la función de densidad nos permitirá calcular la probabilidad dentro de un intervalo de valores (incluido en el conjunto A).

La probabilidad de que una variable aleatoria continua, X, pertenezca a un intervalo $[a;b]$, está dada por la integral de su función de densidad entre los extremos de dicho intervalo:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$$

De acuerdo con esta definición, podemos ver que la probabilidad puntual de una variable aleatoria continua es efectivamente nula:

$$P(X = a) = P(a \leq X \leq a) = \int_{a}^{a} f(x)\,dx = 0 \qquad (3)$$

Ejemplo 13

Consideremos una variable aleatoria X, definida sobre el intervalo $[0;3]$, cuya función de densidad es:

$$f(x) = \frac{x^2}{9} \qquad 0 \leq x \leq 3$$

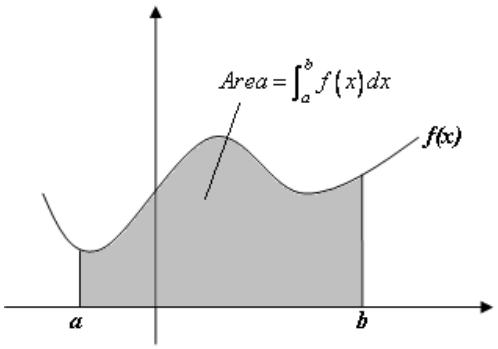
Para comprobar que es una función de densidad, en primer lugar, observamos que es no negativa en el intervalo en el cual está definida. Luego, comprobamos que la integral de la función en el intervalo es igual a uno:

$$\int_{0}^{3} \frac{x^2}{9}\,dx = \left.\frac{x^3}{27}\right|_{0}^{3} = 1$$

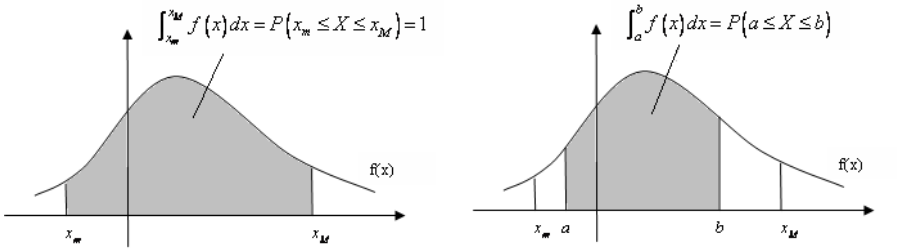
Si, por ejemplo, deseamos saber cuál es la probabilidad de que la variable aleatoria se encuentra en el intervalo de la unidad, simplemente calculamos la integral de la función de densidad en el mismo:

$$P(0 \le X \le 1) = \int_0^1 \frac{x^2}{9} dx = \frac{1}{27}$$

Del estudio de funciones, recordamos que la integral de una función no negativa sobre un intervalo determinado nos proporcionaba el área que se encuentra entre la curva de la función y el eje de las abscisas, como ilustra la siguiente figura:

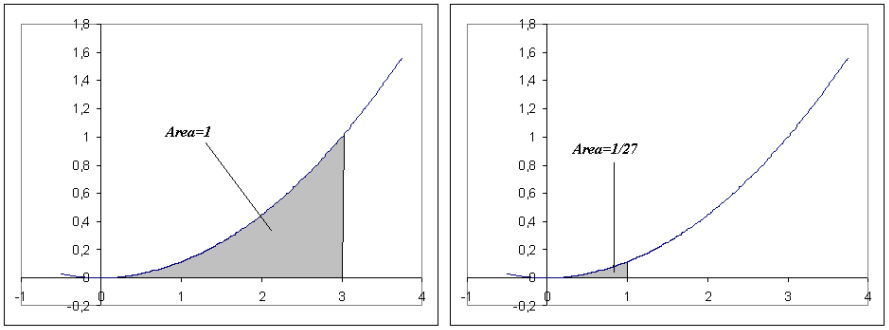


De esta manera, el área que se encuentra entre la curva de una función de densidad y el eje de abscisas, en el dominio correspondiente a la variable aleatoria, $\left[x_m; x_M\right]$, debe ser uno. Asimismo, la probabilidad de que la variable aleatoria se encuentre entre dos valores *a* y *b*, representa el área que se encuentra entre el eje de abscisas, las rectas $x = a$, $x = b$ y la curva de la función. En las siguientes figuras, se observa lo mencionado.



Ejemplo 14

En el ejemplo anterior hemos comprobado que la integral de la función de densidad $f(x) = x^2/9$ en el dominio de la variable aleatoria, $X \in [0;3]$, es uno. A su vez, hemos visto que la probabilidad de que la variable aleatoria se encuentre entre 0 y 1 era de 1/27, que es justamente la integral de la función de densidad entre dichos valores. Gráficamente, podemos observar que los valores mencionados se corresponden con las áreas debajo de la función de densidad.



Ejemplo 15

Consideremos la siguiente función de densidad que depende del parámetro α.

$$f(x)=\frac{x^3}{\alpha}$$

¿Cuál es el valor de este parámetro si el dominio de la variable aleatoria es el intervalo $[0;10]$?

Para el valor de α hallado ¿cuál es la probabilidad de que la variable aleatoria sea menor que 5 y la probabilidad de que la misma se encuentre entre 6 y 9?

En primer lugar, para calcular el valor de α, debemos tener presente la condición que debe cumplir la función de densidad, es decir, que la integral entre el máximo y el mínimo valor del dominio debe ser igual a 1:

$$\int_0^{10} f(x)dx=\int_0^{10}\frac{x^3}{\alpha}dx=\frac{1}{\alpha}\frac{x^4}{4}\Big|_{x=0}^{x=10}=\frac{1}{\alpha}\left(\frac{10^4}{4}-\frac{0^4}{4}\right)=\frac{1}{\alpha}2500$$

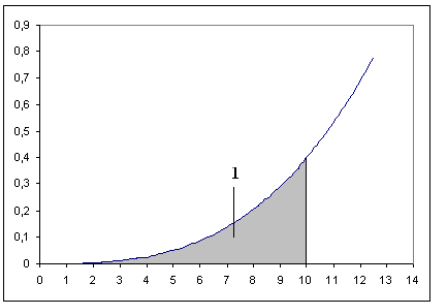
Para que la integral sea igual a 1, necesariamente $\alpha=2500$.

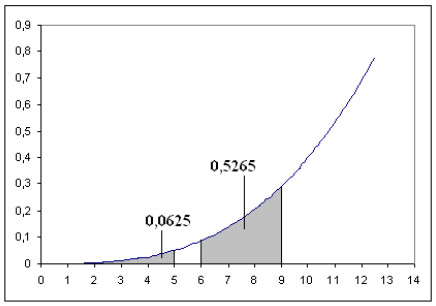
Luego, con este valor de α, calculamos las probabilidades deseadas integrando en el intervalo correspondiente:

$$P(X<5)=\int_0^5\frac{x^3}{2500}dx=\frac{1}{2500}\left(\frac{5^4-0^4}{4}\right)=0,0625$$

$$P(6<X<9)=\int_6^9\frac{x^3}{2500}dx=\frac{1}{2500}\left(\frac{9^4-6^4}{4}\right)=0,5265$$

Gráficamente:





En el siguiente apartado estudiaremos las funciones de distribución, las cuales constituyen una herramienta fundamental a la hora de analizar variables aleatorias y calcular probabilidades asociadas a las mismas. En la exposición se trabajará de manera simultánea con variables aleatorias discretas y continuas.

2.2.3 Función de Distribución

Hasta aquí hemos visto cómo calcular la probabilidad de que una variable aleatoria tome cierto valor o pertenezca a cierto intervalo. Ahora pasaremos a presentar la función de distribución, la cual es ampliamente utilizada para el cálculo de probabilidades.

La Función de Distribución $F_X(x)$ asociada a una variable aleatoria *X* es la función que permite calcular la probabilidad de que la variable aleatoria asuma un valor menor o igual al argumento de la función. Es decir,

$$F_X(x) = P(X \leq x)$$

Así, en el caso discreto tenemos que:

$$F_X(x) = \sum_{y \leq x} p(y) \qquad (4)$$

mientras que en el caso continuo:

$$F_X(x) = \int_{y \leq x} f(y)\,dy \qquad (5)$$

En las definiciones anteriores, ***y*** es una variable auxiliar que se utiliza como índice para realizar la suma en el caso discreto, o como variable de integración en el caso continuo.

Ejemplo 16

Consideremos el Ejemplo 15, donde la función de densidad era $f(x) = x^3/2500$ y el dominio era $[0;10]$. La función de distribución es:

$$F(x) = \int_0^x \frac{y^3}{2500}\,dy = \frac{1}{2500}\frac{y^4}{4}\Big|_{y=0}^{y=x} = \frac{x^4}{10000}$$

A partir de la función de distribución, podemos calcular la probabilidad de que la variable aleatoria sea menor o igual a un valor especificado. Por ejemplo, la probabilidad de que la variable sea menor o igual a 5 es de:

$$P(X \leq 5) = F(5) = \frac{5^4}{10000} = 0,0625$$

Podemos observar que la probabilidad de que la variable de los Ejemplos 15 y 16 sea *menor* a 5 coincide con la probabilidad de que sea *menor o igual* a 5. Esto se debe a que la probabilidad puntual de cada valor de una variable aleatoria continua es cero, tal cual se expuso en la expresión (3).

Veamos un ejemplo de variable aleatoria discreta.

Ejemplo 17

Consideremos el Ejemplo 12, donde la función de probabilidad estaba dada de manera tabular:

x	1	2	3	4	5	6	7	8	9	10
p(x)	0,05	0,06	0,07	0,10	0,20	0,18	0,17	0,10	0,04	0,03

La función de distribución indica la probabilidad acumulada hasta cada valor, y se obtiene mediante la suma de las probabilidades asociadas a los valores de la variable menores o iguales al valor correspondiente. Por ejemplo,

$$F(1) = P(X \le 1) = P(X = 1) = 0,05 \qquad F(2) = P(X \le 2) = P(X = 1) + P(X = 2) = 0,11$$

Con esta lógica, podemos construir de manera tabular la función de distribución para todos los valores de la variable aleatoria[21]:

x	1	2	3	4	5	6	7	8	9	10
F(x)	0,05	0,11	0,18	0,28	0,48	0,66	0,83	0,93	0,97	1,00

En rigor, la Función de Distribución debe definirse sobre toda la recta real. De esta manera, cuando una variable aleatoria tiene un dominio acotado, digamos $X \in [x_m; x_M]$, entonces la Función de Distribución definida sobre toda la recta real es:

$$F_X(x) = \begin{cases} 0 & x \le x_m \\ \int_{x_m}^{x} f(y)\,dy & x_m < x \le x_M \\ 1 & x \ge x_M \end{cases}$$

En palabras, la función vale cero para valores menores al mínimo valor del dominio y vale 1 para valores mayores al máximo. Para valores intermedios, se utilizan las expresiones (4) o (5) del inicio de esta sección, según la variable aleatoria sea discreta o continua.

Ejemplo 18

Consideremos el Ejemplo 16. La función de distribución, definida sobre toda la recta real es:

$$F(x) = \begin{cases} 0 & x \le 0 \\ \dfrac{x^4}{10000} & 0 < x \le 10 \\ 1 & x > 10 \end{cases}$$

En el caso de variables discretas, al definir la función de distribución sobre toda la recta real debemos tener presente que la misma es constante entre los valores que puede tomar la variable y "salta" en cada uno de los valores del dominio.

Ejemplo 19

Consideremos el Ejemplo 17. La función de distribución, definida sobre toda la recta real es:

$$F(x) = \begin{cases} 0 & x \le 1 \\ 0,05 & 1 \le x < 2 \\ 0,11 & 2 \le x < 3 \\ 0,18 & 3 \le x < 4 \\ 0,28 & 4 \le x < 5 \\ 0,48 & 5 \le x < 6 \\ 0,66 & 6 \le x < 7 \\ 0,83 & 7 \le x < 8 \\ 0,93 & 8 \le x < 9 \\ 0,97 & 9 \le x < 10 \\ 1 & x \ge 10 \end{cases}$$

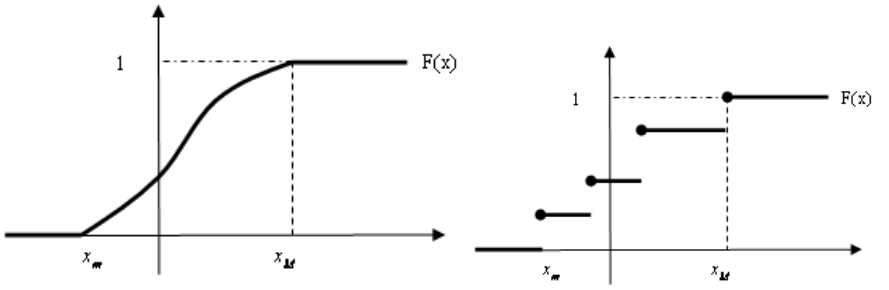
$x < 1$	0
$1 \le x < 2$	0,05
$2 \le x < 3$	0,11
$3 \le x < 4$	0,18
$4 \le x < 5$	0,28
$5 \le x < 6$	0,48
$6 \le x < 7$	0,66
$7 \le x < 8$	0,83
$8 \le x < 9$	0,93
$9 \le x < 10$	0,97
$x \ge 10$	1

[21] Note la semejanza con las frecuencias relativas acumuladas descritas en la primera sección del Capítulo 3

De acuerdo con la definición de **Función de Distribución**, y con los ejemplos anteriores, puede notarse que $F(x)$ es una función **no negativa monótonamente no decreciente**. Es decir, que:

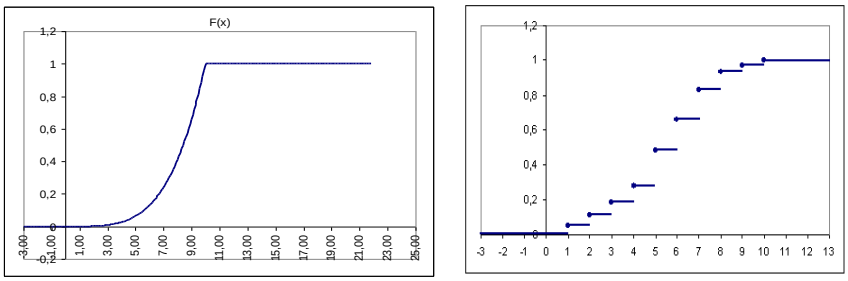
$$si \qquad b \geq a \quad \Rightarrow \qquad F(b) \geq F(a)$$

Si la variable es continua, la función será una curva suave no decreciente que va desde cero hasta uno, cuando la variable independiente crece desde menos infinito hasta más infinito (o desde el mínimo valor de su dominio hasta el máximo). Por otra parte, si la variable es discreta, la función de distribución crecerá de a saltos desde cero hasta uno, cuando la variable crezca desde menos infinito hasta más infinito (o desde el mínimo hasta el máximo). Los valores en los cuales salta la función de distribución son aquéllos en los que la función de probabilidad toma un valor positivo.



Ejemplo 20

Las curvas correspondientes a las funciones de distribución de los Ejemplos 18 y 19 son las siguientes:



Podemos observar que la función de distribución cumple con las siguientes condiciones:

$$\lim_{x \to -\infty} F(x) = 0$$
$$\lim_{x \to +\infty} F(x) = 1$$

Por otra parte, la principal función de la función de distribución es que si conocemos los valores que toma para cada argumento, fácilmente podemos calcular las probabilidades relacionados con intervalos de la variable aleatoria o la probabilidad puntual en el caso de variables aleatorias discretas.

La probabilidad de que una variable aleatoria pertenezca a un determinado intervalo $(a;b]$ está dada por la diferencia entre la función de distribución evaluada en el extremo superior y la función de distribución evaluada en el extremo inferior. Es decir:

$$P(a < X \leq b) = F(b) - F(a) \qquad (6)$$

Es importante notar que, para el caso de variables discretas, en la fórmula precedente no se incluye el valor del extremo inferior del intervalo, ya que, de acuerdo a la definición de la función de distribución, estamos calculando la probabilidad de que la variable sea menor o igual a *b*, y estamos restando la probabilidad de que sea menor o igual al valor **a**. Así, el valor ***a*** de la variable no está incluido en el cálculo de la probabilidad.

Ejemplo 21

Consideremos la siguiente variable aleatoria discreta, con sus correspondientes funciones de probabilidad y de distribución:

x	p(x)
0	0,125
1	0,375
2	0,375
3	0,125

x	F(x)
$x < 0$	0
$0 \le x < 1$	0,125
$1 \le x < 2$	0,5
$2 \le x < 3$	0,875
$x \ge 3$	1

$$F(x) = \begin{cases} 0 & x < 0 \\ 0{,}125 & 0 \le x < 1 \\ 0{,}5 & 1 \le x < 2 \\ 0{,}875 & 2 \le x < 3 \\ 1 & x \ge 3 \end{cases}$$

Usando la expresión (6) anterior podemos calcular la probabilidad de que la variable sea mayor a 1 y menor o igual a 3:

$$P(1 < X \le 3) = F(3) - F(1) = 1 - 0{,}5 = 0{,}5$$

Análogamente, pudimos haber calculado la probabilidad anterior como:

$$P(1 < X \le 3) = P(X = 2) + P(X = 3) = 0{,}375 + 0{,}125 = 0{,}5$$

En la fórmula anterior, tuvimos presente lo mencionado en el párrafo precedente, respecto de no incluir el valor inferior del intervalo (1 en este caso) al sumar las probabilidades.

En el caso de variables aleatorias continuas, no es necesario hacer la salvedad mencionada precedentemente, ya que la probabilidad puntual de que la variable aleatoria asuma el valor ***a*** es nula.

Ejemplo 22

Consideremos una variable aleatoria x , definida sobre los reales positivos, cuya función de densidad es:

$$f(x) = 0{,}5e^{-0{,}5 \cdot x} \qquad x > 0$$

La función de distribución para $x > 0$ es la integral de la función de densidad:

$$\begin{aligned} F(x) &= \int_0^x 0{,}5e^{-0{,}5 \cdot y} dy \\ &= 1 - e^{-0{,}5 \cdot x} \end{aligned}$$

Por lo tanto:

$$F(x) = \begin{cases} 0 & x \le 0 \\ 1 - e^{-0{,}5 \cdot x} & x > 0 \end{cases}$$

La probabilidad de que la variable esté entre 5 y 10 es:

$$\begin{aligned} P(5 < X \le 10) &= F(10) - F(5) \\ &= \left(1 - e^{-0{,}5 \times 10}\right) - \left(1 - e^{-0{,}5 \times 5}\right) \\ &= e^{-2{,}5} - e^{-5} \\ &\cong 0{,}075347 \end{aligned}$$

Teniendo en cuenta que la probabilidad puntual de una variable aleatoria continua es cero, podemos establecer las siguientes relaciones:

Si X es una variable aleatoria continua, se cumplen las siguientes relaciones:

$$\begin{aligned} P(a < X \le b) &= P(a \le X \le b) \\ &= P(a < X < b) \\ &= P(a \le X < b) \end{aligned}$$

Con esta definición, podemos re-expresar (3) correspondiente al cálculo de la probabilidad puntual de la siguiente manera:

$$P(X = a) = P(a \le X \le a) = F(a) - F(a) = 0$$

Además, de acuerdo con su definición, la función de distribución se obtiene mediante el cálculo de la integral de la función de densidad. De esta manera, la función de densidad es la derivada de la función de distribución:

$$\frac{dF(x)}{dx} = \frac{d}{dx}\int_{-\infty}^{x} f(y)\,dy = f(x)$$

Ejemplo 23

Consideremos el ejemplo anterior, donde la función de distribución era:

$$F(x) = 1 - e^{-0,5 \cdot x}$$

La función de densidad puede obtenerse como:

$$f(x) = \frac{d}{dx}\left(1 - e^{-0,5 \cdot x}\right) = -\left(-0,5e^{-0,5 \cdot x}\right) = 0,5e^{-0,5 \cdot x} \quad x > 0$$

Podemos observar que coincide con aquélla definida en el Ejemplo 22.

Por otra parte, en el caso de variables aleatorias discretas, cada valor de la función de probabilidad puede calcularse como la diferencia entre dos valores consecutivos de la función de distribución:

$$p(x_{k+1}) = F(x_{k+1}) - F(x_k)$$

En esta expresión, x_k indica el k-ésimo valor que puede asumir la variable, y x_{k+1} el siguiente. Esta salvedad la hacemos ya que no necesariamente los valores coincidirán con dos enteros consecutivos. Por ejemplo, si la variable puede tomar los valores 0, 2 y 4, tenemos que $p(2) = F(2) - F(0)$, siendo 0 y 2 números no consecutivos, pero sí adyacentes en el dominio de la variable.

Ejemplo 24

Consideremos el Ejemplo 21. Podemos ver las siguientes relaciones[22]:

$$\begin{aligned} p(0) &= F(0) - F(-1) \\ &= 0,125 - 0 \\ &= 0,125 \end{aligned} \qquad \begin{aligned} p(1) &= F(1) - F(0) \\ &= 0,5 - 0,125 \\ &= 0,375 \end{aligned}$$

[22] Recordemos que para cualquier $x < 0$, $F(x) = 0$.

$$p(2) = F(2) - F(1) \qquad p(3) = F(3) - F(2)$$
$$= 0,875 - 0,5 \qquad = 1 - 0,875$$
$$= 0,375 \qquad = 0,125$$

2.2.4 Distribución de Probabilidad

Para caracterizar completamente a una variable aleatoria debemos conocer su distribución de probabilidad, la cual es una combinación de los conceptos expresados hasta aquí.

La **Distribución de Probabilidades** (o simplemente la "distribución") de una variable aleatoria está dada por el dominio de la misma (los posibles valores que puede tomar) y por su función de probabilidad o densidad, o bien, por su función de distribución, ya que mediante una puede obtenerse la otra.

Ejemplo 25

De los Ejemplos 22 y 23 podemos expresar la distribución de la variable x como:

$$F(x) = \begin{cases} 0 & x \le 0 \\ 1 - e^{-0,5 \cdot x} & x > 0 \end{cases}$$

O bien, como:

$$f(x) = 0,5e^{-0,5 \cdot x} \qquad x > 0$$

Y $f(x) = 0$ en otro caso.

En el Ejemplo 21, cualquiera de las dos tablas expuestas en el mismo representa la distribución de la variable aleatoria de dicho ejemplo.

En la sección siguiente analizaremos algunas medidas referidas a una variable aleatoria que permiten conocer ciertas características respecto de la distribución de sus valores.

2.3 Cuantiles, Momentos y otras medidas

2.3.1 Cuantiles

Los cuantiles de una distribución son valores de la variable aleatoria que acumulan una determinada probabilidad, es decir, que la probabilidad de que la variable aleatoria sea menor al cuantil, es un valor predeterminado.

Cuando se desean obtener los cuantiles, el trabajo sería a la inversa: se conoce la probabilidad, y se desea saber cuál es el valor de la variable que acumula dicha probabilidad.

El **cuantil *q*** de una distribución, x_q, es el valor de la variable aleatoria tal que la probabilidad de que la misma sea menor o igual a ese valor es *q*. Es decir:

$$P(X \le x_q) = q$$

o expresando en términos de la Función de Distribución:

$$F(x_q) = q$$

Ejemplo 26

Consideremos la función de distribución:

$$F(x) = \begin{cases} 0 & x \le 0 \\ 1 - e^{-0,5 \cdot x} & x > 0 \end{cases}$$

El cuantil q de esta variable aleatoria es:

$$q = 1 - e^{-0,5 \times x_q} \qquad \Rightarrow \qquad x_q = -2 \times \ln(1-q)$$

Por ejemplo, el valor de la variable que acumula 0,90 es:

$$\begin{aligned} x_{0,9} &= -2 \times \ln(1-0,9) \\ &\cong 4,6051702 \end{aligned}$$

Podemos comprobar que: $F(4,6051702) = 1 - e^{-0,5 \times 4,6051702} = 0,9$.

Algunos autores prefieren separar los cuantiles en cuartiles, deciles y percentiles, de acuerdo a la división que se haga del dominio de la variable aleatoria. Así, por ejemplo, los cuartiles separan en cuatro a la función de probabilidad (o densidad), el primer cuartil acumula 25%, el segundo 50%, el tercero 75% y el cuarto 100%. Los deciles dividen en diez el dominio de la variable, el primero acumula un 10%, el segundo un 20%, y así sucesivamente. Finalmente, los percentiles dividen en cien el dominio, y el primer cuantil acumula un 1%, el segundo un 2%, y así sucesivamente.

Puede observarse que la definición de cuantil es mucho más general e incluye a las otras definiciones (por ejemplo, el segundo cuartil, quinto decil y percentil cincuenta es el cuantil 0,5), y es por ello que en la presente obra se trabajará exclusivamente con cuantiles, para evitar la confusión del lector.

2.3.2 Momentos Absolutos y Centrados

Los momentos son valores numéricos que están relacionados con la distribución de probabilidades de una variable aleatoria y permiten conocer ciertas características de la misma.

Los momentos de una variable aleatoria *X* son un promedio ponderado de una función de dicha variable, $g(X)$, donde los ponderadores son los valores de su función de probabilidad o densidad, según sea discreta o continua la variable aleatoria. El resultado obtenido se denomina valor esperado de la función $g(X)$:

$$\begin{aligned} E\left[g(X)\right] &= \sum_{\forall x} g(x)\, p(x) \\ &\quad o \qquad\qquad (7) \\ E\left[g(X)\right] &= \int_{\forall x} g(x)\, f(x)\, dx \end{aligned}$$

De acuerdo con la forma que tome la función $g(x)$ en la definición anterior, surgen distintos momentos de la variable.

El **momento absoluto de orden *r*** de una variable aleatoria surge de definir $g(x) \equiv x^r$ en la expresión (7):

$$\begin{aligned} E\left[X^r\right] &= \sum_{\forall x} x^r\, p(x) \\ &\quad o \qquad\qquad (8) \\ E\left[X^r\right] &= \int_{\forall x} x^r\, f(x)\, dx \end{aligned}$$

El momento absoluto de orden uno se denomina **esperanza matemática** o valor esperado de la variable, y suele denotarse por μ:

$$E(X) = \sum_{\forall x} x.p(x)$$

o

$$E(X) = \int_{\forall x} x.f(x)dx$$

Puede observarse que este valor es un promedio ponderado de los valores de la variable aleatoria. Los ponderadores dan más peso a los valores más probables, y de ahí su nombre de valor esperado.

Ejemplo 27

Consideremos la siguiente función de densidad:

$$f(x) = -\frac{1}{36}x^2 + \frac{1}{6}x \qquad 0 \le x \le 6$$

En primer lugar, comprobamos que se trata de una función de densidad, ya que es no negativa en el intervalo $[0;6]$ y la integral en dicho soporte es 1:

$$\int_0^6 \left(-\frac{x^2}{36} + \frac{x}{6}\right)dx = \left(-\frac{x^3}{36\times 3} + \frac{x^2}{6\times 2}\right)\Bigg|_{x=0}^{x=6} = -\frac{6^3}{108} + \frac{6^2}{12} = 1$$

La esperanza matemática es:

$$E(X) = \int_0^6 x\left(-\frac{x^2}{36} + \frac{x}{6}\right)dx = \left(-\frac{x^4}{36\times 4} + \frac{x^3}{6\times 3}\right)\Bigg|_{x=0}^{x=6} = -\frac{6^4}{144} + \frac{6^3}{18} = -9 + 12 = 3$$

Ejemplo 28

Consideremos una variable aleatoria cuya función de densidad es:

$$f(x) = 0,01e^{-0,01x} \qquad x > 0$$

La esperanza matemática es[23]:

$$\begin{aligned} E(X) &= \int_0^\infty x.0,01e^{-0,01x}dx \\ &= \cancel{0,01}\left[x\frac{-e^{-0,01x}}{\cancel{0,01}}\Bigg|_{x=0}^{x\to\infty} - \int_0^\infty \frac{-e^{-0,01x}}{\cancel{0,01}}dx\right] \\ &= \underbrace{\left[\lim_{x\to\infty}\left(-xe^{-0,01x}\right) - \left(0.e^{-0,01\times 0}\right)\right]}_{0} + \frac{-e^{-0,01x}}{0,01}\Bigg|_{x=0}^{x\to\infty} \\ &= -\frac{1}{0,01}\left[\lim_{x\to\infty}\left(e^{-0,01\cdot x}\right) - \left(e^{-0,01\cdot 0}\right)\right] \\ &= 100 \end{aligned}$$

Ejemplo 29

Consideremos una variable aleatoria discreta cuya función de probabilidad está dada por:

[23] Para el cálculo se utilizó el método de integración por partes.

x	p(x)
0	0,125
1	0,375
2	0,375
3	0,125

El valor esperado es:

$$\begin{aligned} E(X) &= \sum_{x=0}^{3} x.p(x) \\ &= 0\times 0{,}125 + 1\times 0{,}375 + 2\times 0{,}125 + 3\times 0{,}125 \\ &= 1{,}5 \end{aligned}$$

En el ejemplo anterior, podemos observar que la esperanza matemática no pertenece al dominio de la variable aleatoria, lo cual sucede en muchas ocasiones. La interpretación debe realizarse pensando en que la variable surge de un experimento aleatorio y, si repetimos muchas, pero muchas veces el mismo, el promedio de las observaciones será la esperanza matemática. De hecho, la variable aleatoria del ejemplo anterior puede pensarse como la cantidad de caras que se obtienen al lanzar tres veces consecutivas una moneda. Si realizamos los tres lanzamientos una gran cantidad de veces, la cantidad promedio de caras que se obtienen será de aproximadamente 1,5.

El **momento centrado de orden *r*** de una variable aleatoria surge de definir $g(x) \equiv \left[x - E(X)\right]^r$ en la expresión (7):

$$E\left\{\left[X - E(X)\right]^r\right\} = \sum_{\forall x} \left[x - E(X)\right]^r \, p(x)$$

o

$$E\left\{\left[X - E(X)\right]^r\right\} = \int_{\forall x} \left[x - E(X)\right]^r \, f(x)dx$$

El momento centrado de orden 2 se denomina **varianza**:

$$Var(X) = \sum_{\forall x} \left[x - E(X)\right]^2 \, p(x)$$

o

$$Var(X) = \int_{\forall x} \left[x - E(X)\right]^2 \, f(x)dx$$

El concepto de varianza indica el grado de dispersión de los valores de la variable aleatoria en torno de su valor esperado.

Así, al realizar un experimento aleatorio, “esperamos” que el valor de la variable aleatoria asociada a dicho experimento sea $E(X)$. Lógicamente esto difícilmente ocurra, y está relacionado con la varianza de la variable. Un alto valor de la varianza indica que los valores están muy dispersos respecto de su esperanza, por lo cual es muy probable que el valor obtenido al realizar una vez el experimento difiera considerablemente de $E(X)$. Por otro lado, una varianza pequeña indica un alto grado de concentración de la variable en torno a su valor esperado y, de este modo, al realizar una vez el experimento es más probable que obtengamos un valor cercano a $E(X)$.

Es difícil utilizar directamente a la varianza como un indicador del grado de dispersión (sobre todo al momento de interpretar el resultado), ya que la unidad de medida de la variable aleatoria queda elevada al cuadrado, de manera que, es más útil para estos fines el desvío estándar.

El **desvío estándar** de una variable aleatoria es la raíz cuadrada de la varianza:

$$d.e.(X) = \sqrt{Var(X)}$$

Ejemplo 30

En el Ejemplo 27 calculamos la esperanza de una variable aleatoria cuya función de densidad era:

$$f(x) = -\frac{1}{36}x^2 + \frac{1}{6}x \qquad 0 \le x \le 6$$

La esperanza resultó ser $E(X) = 3$. De este modo, la varianza será:

$$\begin{aligned} Var(X) &= \int_0^6 (x-3)^2 \left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx \\ &= \int_0^6 \left(x^2 + 3^2 - 6x\right) \left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx \\ &= \int_0^6 x^2 \left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx + 3^2 \int_0^6 \left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx - 6\int_0^6 x\left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx \end{aligned}$$

El segundo término de la suma en la expresión anterior contiene la integral de la función de densidad en todo el dominio, por lo cual dicha integral es igual a 1. El último término contiene la integral de x multiplicada por la función de densidad, es decir, que el resultado será igual a la esperanza matemática, cuyo valor es 3. Entonces,

$$\begin{aligned} Var(X) &= \int_0^6 x^2 \left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx + 3^2 \times 1 - 6 \times 3 \\ &= \left(-\frac{1}{36}\frac{x^5}{5} + \frac{1}{6}\frac{x^4}{4}\right)\Bigg|_{x=0}^{x=6} + 9 - 18 \\ &= -\frac{216}{5} + 54 - 9 \\ &= 1{,}8 \end{aligned}$$

Luego, el desvío estándar es:

$$\begin{aligned} d.e.(X) &= \sqrt{1{,}8} \\ &\cong 1{,}341641 \end{aligned}$$

Ejemplo 31

Calculemos la varianza de la variable del Ejemplo 29.

$$\begin{aligned} Var(X) &= \sum_{x=0}^{3} (x - 1{,}5)^2 \, p(x) \\ &= (0-1{,}5)^2 \times 0{,}125 + (1-1{,}5)^2 \times 0{,}375 + (2-1{,}5)^2 \times 0{,}375 + (3-1{,}5)^2 \times 0{,}125 \\ &= 0{,}75 \end{aligned}$$

Luego, el desvío estándar es:

$$\begin{aligned} d.e.(X) &= \sqrt{0{,}75} \\ &\cong 0{,}866025 \end{aligned}$$

La varianza puede calcularse de manera alternativa como:

$$Var(X) = E(X^2) - E(X)^2$$

Ejemplo 32

Consideremos la función de densidad del Ejemplo 28. El momento absoluto de orden dos es[24]:

$$\begin{aligned}
E(X^2) &= \int_0^{\infty} x^2 0,01 e^{-0,01x} dx \\
&= \cancel{0,01}\left[\underbrace{x^2 \frac{-e^{-0,01x}}{\cancel{0,01}}\Big|_{x=0}^{x\to\infty}}_{0} - \int_0^{\infty} \frac{-2xe^{-0,01x}}{\cancel{0,01}} dx \right] \\
&= 2\left[\underbrace{x\frac{-e^{-0,01x}}{0,01}\Big|_{x=0}^{x\to\infty}}_{0} - \int_0^{\infty} \frac{-e^{-0,01x}}{0,01} dx \right] \\
&= \frac{2}{0,01^2} \underbrace{\left(-e^{-0,01x}\Big|_{x=0}^{x\to\infty} \right)}_{1} \\
&= \frac{2}{0,01^2} \\
&= 20000
\end{aligned}$$

Luego, la varianza será:

$$\begin{aligned}
Var(X) &= E(X^2) - E(X)^2 \\
&= 20000 - 100^2 \\
&= 10000
\end{aligned}$$

El desvío estándar es:

$$\begin{aligned}
d.e.(X) &= \sqrt{10000} \\
&= 100
\end{aligned}$$

De acuerdo a lo visto en los ejemplos precedentes, el valor del desvío estándar, no es muy útil si no lo comparamos con la esperanza. Por ejemplo, consideremos dos variables positivas, una con valor esperado de 1.000.000 y la otra con esperanza igual a 5, siendo el desvío estándar de ambas igual a 10. En la primera, la desviación es muy pequeña en relación a la media, mientras que en la segunda es mucho mayor. Por ello, resulta importante el estudio del desvío en relación a la media de la variable.

Sobre la base de lo mencionado en el párrafo anterior, surge el concepto de Coeficiente de Variación.

El **Coeficiente de Variación** de una variable aleatoria es el cociente entre el desvío estándar y la esperanza matemática:

$$c.v.(X) = \frac{d.e.(X)}{E(X)}$$

Ejemplo 33

El coeficiente de variación de los Ejemplos 29 y 31 es:

[24] Nuevamente, utilizamos la integración por partes en el cálculo.

$$c.v.(X) \cong \frac{0,866025}{1,5}$$
$$\cong 0,577350$$

Es decir, que el desvío estándar es aproximadamente el 57,73% de la media.

Por otra parte, el desvío estándar de la variable de los Ejemplos 27 y 30 es aproximadamente de 44,72% de la media, ya que

$$c.v.(X) \cong \frac{1,341641}{3}$$
$$\cong 0,447214$$

2.3.3 Medidas de Forma

Existen otras medidas calculadas a partir de los momentos que están relacionadas con la forma que tiene la función de probabilidad (o densidad).

La Simetría nos indica si la distribución de los valores de la variable alrededor de su esperanza es simétrica; es decir, indica si los valores por encima y por debajo de la media tienen la misma relevancia en la distribución.

El **Coeficiente de Asimetría** de una variable aleatoria está dada por el cociente entre el momento centrado de orden 3 y el desvío estándar elevado al cubo:

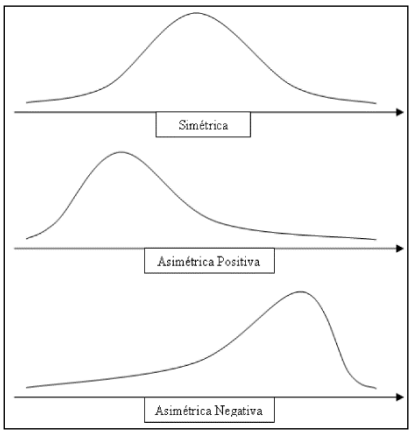
$$As(X) = \frac{E\left\{\left[X - E(X)\right]^3\right\}}{d.e.(X)^3} = \frac{E\left\{\left[X - E(X)\right]^3\right\}}{Var(X)^{3/2}}$$

Si el coeficiente de asimetría es cero, entonces la función de probabilidad (densidad) es simétrica, y se dice que la variable aleatoria lo es. Además, si la variable es simétrica, todos sus momentos centrados de orden impar son cero.

Por otra parte, si el coeficiente de asimetría es positivo, se dice que la variable es asimétrica positiva o hacia la derecha. Esto indica que la "cola" derecha (positiva o de los valores mayores de la variable) es más larga que la "cola" izquierda. Finalmente, si $As(X) < 0$, la variable es asimétrica negativa o hacia la izquierda, indicando que la "cola" izquierda es más larga.

En la figura se ilustra la forma de la distribución de acuerdo al valor del coeficiente de asimetría.

Otro indicador de la forma de la función de probabilidad es el *coeficiente de kurtosis*[25] (o simplemente kurtosis). Este valor nos indica el grado de apuntamiento de la distribución.



[25] O bien curtosis, según la terminología usada por algunos autores.

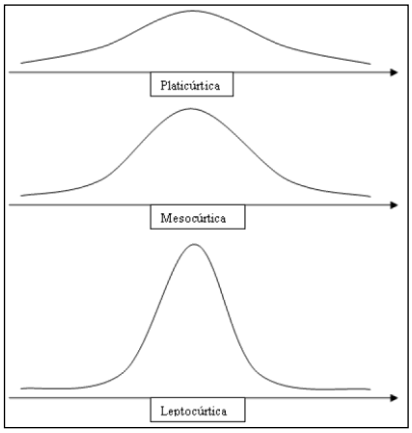
La **Kurtosis** de una variable aleatoria está dada por el cociente entre el momento centrado de orden 4 y la varianza al cuadrado:

$$Ku(X) = \frac{E\left\{\left[X - E(X)\right]^4\right\}}{d.e.(X)^4} = \frac{E\left\{\left[X - E(X)\right]^4\right\}}{Var(X)^2}$$

Si el coeficiente de kurtosis es mayor a 3, decimos que la distribución es muy apuntada (*leptocúrtica*). Si es menor a 3, diremos que la distribución es muy chata (*platicúrtica*). Finalmente, si es igual a 3 diremos que no es ni muy apuntada ni muy chata (*mesocúrtica*)[26].

La kurtosis, al igual que la varianza, y como se desprende de su forma de cálculo, puede interpretarse como un indicador de la dispersión. Una kurtosis elevada indica mayor dispersión.

En la figura se ilustra la forma de la distribución de acuerdo al valor del coeficiente de kurtosis.



Ejemplo 34

Calculemos el Coeficiente de Asimetría y de Kurtosis de la variable aleatoria de los Ejemplos 27 y 30.

El momento centrado de orden 3 es:

$$\begin{aligned}
E\left\{\left[X - E(X)\right]^3\right\} &= \int_0^6 (x-3)^3 \left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx \\
&= \int_0^6 \left(x^3 - 9x^2 + 27x - 27\right)\left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx \\
&= \int_0^6 x^3\left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx - 9\int_0^6 x^2\left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx \\
&\quad + 27\int_0^6 x\left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx - 27\int_0^6 \left(-\frac{1}{36}x^2 + \frac{1}{6}x\right) dx \\
&= \left(-\frac{1}{36}\frac{x^6}{6} + \frac{1}{6}\frac{x^5}{5}\right)\Bigg|_{x=0}^{x=6} - 9\left(-\frac{1}{36}\frac{x^5}{5} + \frac{1}{6}\frac{x^4}{4}\right)\Bigg|_{x=0}^{x=6} \\
&\quad + 27\times 3 - 27\times 1 \\
&= \frac{6^3}{5} - 9\times\frac{6^3}{20} + 81 - 27 \\
&= 0
\end{aligned}$$

Por lo cual, $As(X) = 0$, y la distribución es simétrica.

[26] Muy puntiaguda o poco puntiaguda en relación a la distribución Normal, la cual será tratada en el Capítulo 3.

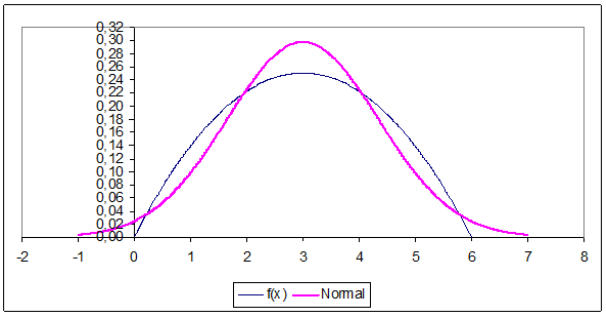
El momento centrado de orden 4 es:

$$
\begin{aligned}
E\left\{\left[X-E(X)\right]^4\right\} &= \int_0^6 (x-3)^4\left(-\frac{1}{36}x^2+\frac{1}{6}x\right)dx \\
&= \int_0^6 \left(x^4-12x^3+54x^2-108x+81\right)\left(-\frac{1}{36}x^2+\frac{1}{6}x\right)dx \\
&= \int_0^6 x^4\left(-\frac{1}{36}x^2+\frac{1}{6}x\right)dx-12\int_0^6 x^3\left(-\frac{1}{36}x^2+\frac{1}{6}x\right)dx+ \\
&+54\int_0^6 x^2\left(-\frac{1}{36}x^2+\frac{1}{6}x\right)dx-108\int_0^6 x\left(-\frac{1}{36}x^2+\frac{1}{6}x\right)dx+ \\
&+81\int_0^6\left(-\frac{1}{36}x^2+\frac{1}{6}x\right)dx \\
&= \left(-\frac{1}{36}\frac{x^7}{7}+\frac{1}{6}\frac{x^6}{6}\right)\Bigg|_{x=0}^{x=6}-12\times\frac{6^3}{5}+54\times\frac{6^3}{20}-108\times 3+81 \\
&= \frac{6^4}{7}-12\times\frac{6^3}{5}+54\times\frac{6^3}{20}-324+81 \\
&\cong 6,942857
\end{aligned}
$$

Luego, el coeficiente de kurtosis es:

$$Ku(X)=\frac{6,942857}{1,8^2}\cong 2,142857$$

Al ser el coeficiente de kurtosis menor a 3, podemos decir que la distribución es menos apuntada que distribución normal. El siguiente gráfico compara, la distribución Normal, que veremos en el próximo capítulo, con la distribución que estamos analizando:



2.4 Distribuciones Discretas

En esta sección se analizan las principales familias de distribuciones para el caso de variables aleatorias discretas. Éstas son: Distribución de Bernoulli, Distribución Binomial, Distribución de Poisson y Distribución Hipergeométrica.

2.4.1 Distribución de Bernoulli

La distribución de Bernoulli está asociada a eventos dicotómicos. Es decir, fenómenos en los cuales el resultado puede tomar únicamente dos valores, que llamaremos de manera genérica "éxito" y "fracaso". Por ejemplo, si se lanza una moneda al aire, el resultado será cara (éxito) o ceca (fracaso). Asimismo, si se pregunta a una persona si simpatiza por un equipo de fútbol, la respuesta será sí (éxito) o no (fracaso), si preguntamos a un individuo si fuma nuevamente la respuesta será sí (éxito) o no (fracaso).

La variable de Bernoulli es una variable aleatoria relacionada con el tipo de fenómenos ejemplificados en el párrafo anterior, y asigna el valor cero al evento "fracaso" y el valor uno al

evento “éxito”. La probabilidad de obtener éxito es p y por lo tanto la probabilidad de fracaso es $1-p$.

Sea X una variable que indica la ocurrencia de “éxito” en un evento determinado, entonces X tiene una **distribución de Bernoulli**, cuya función de probabilidad es:

$$P(X=0|p)=1-p \qquad P(X=1|p)=p$$

Ejemplo 1

Consideremos el lanzamiento de una moneda al aire, y supongamos que estamos interesados en que el resultado sea “cara”. La probabilidad de éxito es la probabilidad de obtener una cara al lanzar la moneda, es decir $p=1/2$. La variable Bernoulli asociada al lanzamiento de la moneda tiene la siguiente distribución.

$$P(X=0|p=1/2)=1-1/2=1/2 \qquad P(X=1|p=1/2)=1/2$$

Ejemplo 2

Supongamos que en una institución compuesta por 150 personas se está realizando un estudio sobre la cantidad de fumadores que integran la misma. En base al estudio se encontró que 50 son fumadores y 100 no.

Sea X una variable Bernoulli que indica “éxito” en caso de que una persona seleccionada al azar fume. El parámetro p indica la proporción de fumadores, es decir que $p=50/150=1/3$. Entonces, la distribución de X es:

$$P(X=0|p=1/3)=1-1/3=2/3 \qquad P(X=1|p=1/3)=1/3$$

2.4.2 Distribución Binomial

La Distribución Binomial es en alguna medida una generalización de la Distribución de Bernoulli. Más precisamente, una variable aleatoria binomial es la suma de variables Bernoulli. Este tipo de variables tiene importantes aplicaciones: estudios de mercado, encuestas electorales, estudio de la mortalidad, seguros de vida, juegos de azar, etc.

Consideremos por ejemplo 10 lanzamientos consecutivos de una moneda, y supongamos que el resultado “cara” representa el éxito. Cada lanzamiento constituye una variable Bernoulli con parámetro $p=1/2$. La cantidad total de éxitos en los 10 lanzamientos es una variable aleatoria binomial.

Supongamos que en una población el 30% de las personas son fumadores. Si se seleccionan 15 personas y se les pregunta si fuma o no, el resultado de cada pregunta es una variable Bernoulli con $p=0,30$, mientas que el total de éxitos (personas que responden que sí fuman) es una variable binomial.

En general, se realizan n ensayos, y la probabilidad de “éxito”[27] en cada uno de ellos es p. Se desea calcular la probabilidad de que ocurran exactamente k éxitos en las n pruebas. Si X es la variable aleatoria que indica el número de éxitos, se desea conocer: $P(X=k|n,p)$. Nótese que la probabilidad está condicionada a la cantidad de ensayos y la probabilidad de éxito en cada uno de ellos.

A su vez, el dominio de la variable binomial está dado por los números enteros comprendidos entre 0 y n, ambos inclusive. Esto es bastante obvio, ya que si repetimos n veces un experimento, a lo sumo obtendremos n éxitos, y como mínimo no obtendremos ninguno. Por ejemplo, si lanzamos 10 veces una moneda, como máximo obtendremos 10 caras, y como mínimo ninguna.

[27] Es importante notar que “éxito”, en términos estadísticos, no necesariamente implica algo deseable. Por ejemplo, en los seguros de vida, la variable de interés es la cantidad de decesos que se producen en un grupo determinado.

En base a lo expuesto hasta aquí, ya hemos identificado el dominio y los parámetros relevantes para el cálculo de la función de probabilidad:

Dominio	**Parámetros**	
$D=\{0,1,2,...,n-1,n\}$	n (cantidad de ensayos)	p (probabilidad de éxito en cada ensayo)

Además, existen dos supuestos fundamentales en los se basa la distribución binomial, a saber:

(a) La probabilidad de éxito es constante en cada ensayo.

(b) Los ensayos son independientes.

Para desarrollar la función de probabilidad analizaremos primero algunos ejemplos.

Ejemplo 3

Consideremos el lanzamiento de una moneda tres veces consecutivas. Sea X la variable que indica la cantidad de caras que se obtienen. Deseamos conocer la probabilidad de que $X=2$, es decir, que se obtengan exactamente dos caras en los tres ensayos.

En base a estos datos, podremos determinar que la cantidad de ensayos independientes es $n=3$ y la probabilidad de éxito en cada ensayo es $p=1/2$.

La probabilidad de obtener la secuencia $CaCaCe$ es:

$$p\times p\times(1-p)=0,5\times 0,5\times 0,5=0,125$$

A su vez, las secuencias $CaCeCa$ y $CeCaCa$ también se corresponden con el evento "salen exactamente dos caras". Estas dos tienen igual probabilidad que la calculada anteriormente, y al ser eventos mutuamente excluyentes, debemos sumar las probabilidades, o simplemente multiplicar la anterior por 3. Es decir que:

$$P\left(X=2\middle|n=3;p=1/2\right)=3\times 0,125=0,375$$

Ejemplo 4

Consideremos el lanzamiento de cinco dados. Sea X la variable que indica la cantidad de dados que caen con el 1 o el 2 hacia arriba. Deseamos saber la probabilidad de que tres dados caigan en 1 o 2, es decir, la probabilidad de que $X=3$.

En este caso, la cantidad de ensayos independientes es $n=5$ y la probabilidad de éxito es $p=1/3$.

El suceso de que los tres primeros sean éxitos (caigan con 1 o 2 hacia arriba) y los dos restantes sean fracasos (caigan con 3 o 4 o 5 o 6 hacia arriba) tiene una probabilidad de:

$$p\times p\times p\times(1-p)\times(1-p)=(1/3)^3\times(2/3)^2=4/243$$

Sin embargo, ésta no es la probabilidad de que $X=3$, ésta es solamente una de las maneras en que se da dicho evento. En este caso, determinar la cantidad de maneras en que se puede producir el éxito es un tanto más complicado que en el ejemplo anterior. Adelantamos que son 10 las formas en que se pueden dar tres éxitos y dos fracasos en los cinco ensayos. Por ello, la probabilidad deseada es:

$$P\left(X=3|n=5;p=1/3\right)=10\times 4/243=40/243\cong 0{,}1646$$

En los ejemplos anteriores, hemos visto que primero calculamos la probabilidad de una de las maneras en que se puede dar el evento de interés, y luego contamos de cuántas maneras distintas se podría dar el mismo.

Más precisamente, en primer lugar, se determinó que la probabilidad de que los primeros k ensayos sean éxitos y los restantes $n-k$ sean fracaso es:

$$p\times p\times p\times \ldots\times (1-p)\times (1-p)=p^k\times (1-p)^{n-k}$$

Luego, para determinar la cantidad de formas en que se puede dar este evento, simplemente debemos contar la cantidad de maneras en que pueden ocurrir los k éxitos y los $n-k$ fracasos. Es decir que se debe determinar cuántas formas existen de ordenar n elementos, donde k tienen una característica determinada. Este número es simplemente la Combinatoria de "n elementos tomado de a k", analizado en el Capítulo 1:

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

En base a lo mencionado, podemos realizar la siguiente definición:

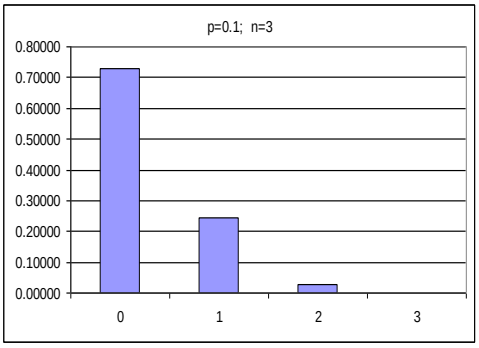
Consideremos n ensayos independientes, en los cuales hay una probabilidad constante, p, de obtener un resultado determinado llamado de manera genérica "éxito". La probabilidad de obtener un "fracaso" es, obviamente, $1-p$.

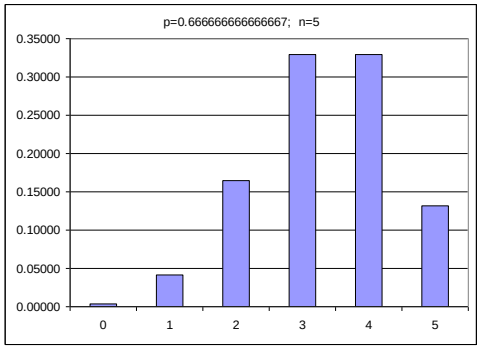
Sea X la variable aleatoria que indica la cantidad de éxitos que se obtienen en los n ensayos independientes. Entonces X tiene una **distribución de probabilidad binomial** dada por:

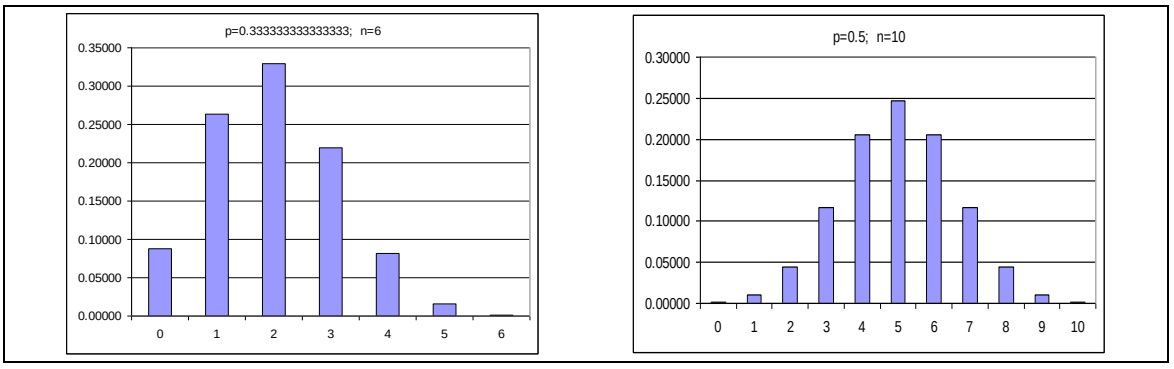
$$P\left(X=k|n,p\right)=\binom{n}{k}p^k\left(1-p\right)^{n-k}\qquad\qquad k=0,1,2,\ldots,n-1,n.$$

Cabe destacar que cuando $n=1$ la distribución binomial se reduce a la distribución de Bernoulli.

En la siguiente figura, se ilustra la distribución binomial para distintos valores de los parámetros n y p.







Al igual que en cualquier distribución discreta, la función de distribución es simplemente la suma de las probabilidades hasta el valor de interés, es decir:

$$F_B\left(k|n,p\right)=P\left(X\le k|n,p\right)=\sum_{j=0}^{k}P\left(X=j|n,p\right)=\sum_{j=0}^{k}\binom{n}{j}p^j\left(1-p\right)^{n-j}$$

Ejemplo 5

Supongamos que en una población determinada hay un tercio de fumadores y dos tercios no fumadores. Si se toma una muestra aleatoria con reposición de 10 personas, ¿cuál es la probabilidad de que exactamente 2 sean fumadores? ¿Cuál es la probabilidad de que haya como máximo 3 fumadores?

Claramente X, la cantidad de fumadores que se encuentren, es una variable binomial con parámetros $n=10$ y $p=1/3$. La probabilidad de que se encuentren exactamente 2 fumadores entonces es:

$$\begin{aligned}P\left(X=2|n=10,p=1/3\right)&=\binom{10}{2}\left(1/3\right)^2\left(2/3\right)^{10-2}\\&=45\left(1/3\right)^2\left(2/3\right)^8\\&\cong 0{,}19509\end{aligned}$$

Por otro lado, la probabilidad de que haya como máximo 3 fumadores es la función de distribución evaluada en 3:

$$\begin{aligned}F_B\left(3|n=10;p=1/3\right)&=\sum_{j=0}^{3}\binom{10}{j}\left(1/3\right)^j\left(2/3\right)^{10-j}\\&\cong 0{,}55926\end{aligned}$$

Ejemplo 6

En una urna hay 8 bolitas, de las cuales 2 son blancas y las restantes negras. Suponga que se extraen al azar, con reposición, 5 bolitas (se extrae una bolita y se la vuelve a colocar en la urna para la siguiente extracción). Se desean responder las siguientes preguntas: ¿Cuál es la probabilidad de que dos sean blancas? ¿Cuál es la probabilidad de que todas sean negras? ¿Cuál es la probabilidad de que al menos dos sean blancas?

En este caso, la cantidad de bolitas blancas es una variable aleatoria binomial X, con parámetros $n=5$ y $p=2/8=1/4$. Entonces, la primera pregunta es la probabilidad de que la variable sea igual a 2:

$$\begin{aligned}P\left(X=2|n=5,p=1/4\right)&=\binom{5}{2}\left(1/4\right)^2\left(3/4\right)^3\\&\cong 0{,}26367\end{aligned}$$

La segunda pregunta que se formula es equivalente a que la cantidad de bolitas blancas sea cero. Es decir que,

$$P(\text{todas negras}) = P(X = 0 | n = 5, p = 1/4)$$
$$= \binom{5}{0} (1/4)^0 (3/4)^5$$
$$\cong 0,23730$$

Finalmente, la probabilidad de que al menos dos sean blancas puede calcularse utilizando los conceptos de eventos complementarios. La probabilidad de que haya al menos dos éxitos, es uno menos la probabilidad de que se den "menos de dos", ya que se trata de eventos complementarios, y del Capítulo 1 sabemos que $P(A) = 1 - P(\bar{A})$. Además, "menos de dos éxitos", es lo mismo que decir que "como máximo un éxito". Con esto en mente, podemos calcular:

$$P(X \geq 2 | n = 5, p = 1/4) = 1 - P(X < 2 | n = 5, p = 1/4)$$
$$= 1 - P(X \leq 1 | n = 5, p = 1/4)$$
$$= 1 - F_B(1 | n = 5, p = 1/4)$$
$$\cong 0,89595$$

A continuación, expondremos las principales características de la distribución binomial. La demostración se expone al final del capítulo.

Sea X una distribución binomial con parámetros n y p. Entonces:

$$E(X | n, p) = n \times p$$

$$Var(X | n, p) = n \times p \times (1 - p)$$

Ejemplo 7

El valor esperado de la variable del Ejemplo 5 es:

$$E(X | n = 10, p = 1/3) = 10/3 \cong 3,33$$

La varianza y el desvío estándar están dados por:

$$V(X | n = 10, p = 1/3) = 10 \times (1/3) \times (2/3)$$
$$= 20/9$$
$$\cong 2,22$$

$$\Rightarrow$$

$$d.e.(X | n = 10, p = 1/3) = \sqrt{20/9}$$
$$= 2/3 \times \sqrt{5}$$
$$\cong 1,49$$

Ejemplo 8

Considerando los datos del Ejemplo 6, la media y el desvío estándar de la variable aleatoria: Cantidad de bolitas blancas es:

$$E(X | n = 5, p = 1/4) = 5 \times 1/4 = 5/4$$

$$d.e.(X | n = 5, p = 1/4) = \sqrt{5 \times 1/4 \times 3/4} = \sqrt{15/16} = \sqrt{15}/4 \cong 0,9682$$

Finalmente, las medidas de forma también pueden calcularse en función de los parámetros n y p.

Sea X una distribución binomial con parámetros n y p. Entonces los coeficientes de asimetría y de kurtosis son:

$$As(X|n,p)=\frac{1-2p}{\sqrt{n\times p\times(1-p)}}$$

$$Ku(X|n,p)=3+\frac{\left[1-6p(1-p)\right]}{np(1-p)}$$

En la fórmula anterior del coeficiente de asimetría, se puede observar que la distribución es simétrica solamente si $p=1/2$.

2.4.3 Distribución Geométrica

En la distribución binomial se trabaja con ensayos repetidos de variables Bernoulli, en donde la variable aleatoria que se representa es el número de éxitos que se obtienen dado un número determinado de ensayos. Tal como se ha visto, el número de ensayos es uno de los parámetros del modelo.

Sin embargo, al trabajar con ensayos de Bernoulli no necesariamente nos va a interesar la cantidad de éxitos sino, por ejemplo, la cantidad de ensayos que debieran repetirse hasta alcanzar el primer éxito. La distribución asociada a este tipo de eventos es la **distribución geométrica**. Nos interesa conocer: cuántas veces debe repetirse un determinado experimento hasta alcanzar el primer éxito. Lo que en esta distribución se considera fijo es, entonces, el número de éxitos, el cual es igual a 1.

Ejemplo 9

Suponga un juego de dados tal que en cada ronda el jugador arroje repetidas veces el mismo dado y se vayan sumando los valores obtenidos en cada tiro, siempre que ese valor sea distinto de uno. En el caso de que el resultado sea igual a uno, se anula todo el puntaje obtenido por el jugador en esa ronda. El jugador decide cuántas veces arrojará el dado en su turno: puede arriesgarse y continuar o conformarse con lo obtenido hasta determinado tiro ante el riesgo de que el próximo sea el número 1. Gana el juego quien sume primero 100 puntos.

Lo que nos interesará es, entonces, modelar en cuál de los lanzamientos saldrá el número 1. Este será el evento "éxito", aunque tendrá una connotación negativa. En caso de que el jugador decida arrojar el dado cuatro veces, está dejando de arriesgarse ante la posibilidad de que el uno salga en el quinto intento. ¿Cuál es la probabilidad de que esto ocurra?

Para resolver este problema debe tenerse en cuenta que estamos trabajando con eventos independientes, que la probabilidad de éxito (obtener un uno) es igual a $1/6$ y la de fracaso es $5/6$. El evento "que salga un uno en el quinto intento" es igual a que en los otros cuatro intentos el resultado haya sido distinto a uno. Con lo cual tendríamos:

$P(T_5=1\cap T_{1,2,3,4}\neq 1)=P(T_1\neq 1)\times P(T_2\neq 1)\times P(T_3\neq 1)\times P(T_4\neq 1)\times P(T_5=1)$, dado que los eventos son independientes.

Asimismo, como la probabilidad de éxito es constante en cada uno de los experimentos, tenemos:

$$P(T_5=1\cap T_{1,2,3,4}\neq 1)=P(T\neq 1)^4\times P(T=1)=\left(\frac{5}{6}\right)^4\times\left(\frac{1}{6}\right)\cong 0.08038$$

La variable *T* es una variable de Bernoulli, la variable que estamos analizando en este caso es la variable *Y*, número de ensayos hasta obtener el primer éxito. Entonces, estamos haciendo:

$$P(Y=5)=\left(\frac{5}{6}\right)^4\times\frac{1}{6}$$

que es el mismo resultado obtenido anteriormente.

En una serie de ensayos independientes de Bernoulli, donde la probabilidad de éxito es igual a p, se considera la variable aleatoria X, donde X es el número de ensayos hasta obtener el primer éxito. La variable X tiene **distribución geométrica**, con función de probabilidad:

$$P(Y = y|p) = (1-p)^{y-1} \times p$$

Como puede notarse en la definición, existe un único parámetro que caracteriza a la distribución: la probabilidad de éxito. El dominio es el conjunto de los números naturales (sin considerar el cero): en el caso del ejemplo anterior, se podría querer calcular la probabilidad de que arroje infinitas veces el dado hasta obtener un uno. La probabilidad de ese evento sería muy pequeña. Hay que tener en cuenta, de todas formas, que, al tratarse de ensayos de Bernoulli, la probabilidad de que arrojemos, por ejemplo, cien veces el dado hasta obtener un uno es igual a la probabilidad de que en el quinto tiro obtengamos nuestro primer uno y en las 95 jugadas siguientes salga cualquier otro número menos ese.

La siguiente tabla resume las características antedichas:

Dominio	Parámetro
$D = \{1, 2, ..., n, ..., \infty\}$	p (probabilidad de éxito en cada ensayo)

Es importante la interpretación de la función de distribución, $P(Y \leq n)$, ya que la misma implica la probabilidad de que el primer éxito ocurra antes del enésimo ensayo. Al momento de realizar el cálculo estamos frente a una progresión geométrica de la forma:

$$P(Y \leq n) = \sum_{i=1}^{n-1} (1-p)^{i-1} \times p = 1 - (1-p)^{n-1}$$

Veamos el siguiente ejemplo:

Ejemplo 10

Consideremos el caso de un matrimonio joven que desea tener un hijo. Asuma que las probabilidades de tener un hijo varón o una hija mujer son las mismas. ¿Cuál es la probabilidad de que su primer hijo sea varón? ¿Cuál es la probabilidad de que el primer hijo varón sea el tercero?

Primero debemos definir la probabilidad de éxito (probabilidad de tener un hijo varón), la cual es igual a $1/2$. La probabilidad de que el primer hijo sea varón es $P(Y = 1) = 0,5^0 \times 0,5 = 0,5$

El caso en donde se busca que $Y = 1$ la probabilidad es equivalente a la de un único ensayo de Bernoulli, es decir, es equivalente a considerar la probabilidad de éxito: que el resultado del ensayo sea éxito.

El segundo caso planteado es: $P(Y = 3) = 0,5^2 \times 0,5 = 0,125$. Esto implica que la probabilidad de que el primer hijo varón sea el tercero es de 0,125

Los momentos de esta variable aleatoria son los que se muestran en el recuadro siguiente:

Sea X una variable aleatoria con **distribución geométrica**, su esperanza y su varianza son:

$$E(X|p) = \frac{1}{p}$$

$$Var(X|p) = \frac{1-p}{p^2}$$

Ejemplo 11

Si retomamos el juego de dados del Ejemplo 9, podemos considerar la media de la distribución:

$$E\left(Y\middle|p=1/6\right)=\frac{1}{1/6}=6$$

Es importante tener en cuenta la interpretación y no, únicamente, el cálculo. Que la media sea igual a 6 implica que, en promedio, el primer número uno se obtendrá en el sexto lanzamiento del dado. Eso nos llevaría a no intentar más de cinco lanzamientos consecutivos. Si proseguimos el análisis con el cálculo de la varianza, se puede observar que el desvío estándar es:

$$Var\left(X\middle|\frac{1}{6}\right)=\frac{5/6}{1/36}=30 \quad \Rightarrow d.e\left(X\middle|\frac{1}{6}\right)=\sqrt{30}$$

Por lo tanto, el coeficiente de variación es mayor al 90%, lo cual indica que existe una gran dispersión en relación a la media:

$$Coef.Var=\frac{\sqrt{30}}{6}\cong 0.9129$$

2.4.4 Distribución Hipergeométrica

En la Sección 2.4.2 se analizó la distribución Binomial, donde se tomaba una muestra *con reposición* de una población determinada. De esta manera, la probabilidad de éxito se mantiene constante en cada ensayo.

Si la muestra se toma sin reposición de una población finita, la probabilidad de éxito en cada ensayo no es constante, y el cálculo de la probabilidad debe modificarse.

En este caso, deben considerarse tres parámetros: el tamaño de la población, la cantidad de éxitos que hay en la misma y el tamaño de la muestra. Para ilustrar el mecanismo, empezaremos por un ejemplo.

Ejemplo 12

Supongamos que se posee una baraja española de 40 naipes, y se desea calcular la probabilidad de obtener 2 cartas de oro si se extraen 4 cartas sin reposición. Es decir, si X es la variable que representa la cantidad de oros que se extraen, se desea la probabilidad de que $X=4$. El tamaño de la población es 40, el tamaño de la muestra es 4 y la cantidad de éxitos en la muestra es 10 (hay 10 cartas de cada palo: oros, espadas, bastos y copas)

Cuando se saca la primera carta, la probabilidad de extraer un oro es:

$$p_1=10/40$$

Sin embargo, al extraer la segunda carta, la probabilidad de éxito dependerá del resultado de la primera extracción. En primer lugar, quedan solamente 39 cartas en la baraja. Además, si la primera carta fue oro, quedan solamente 9 casos favorables y 30 desfavorables. Es decir, que la probabilidad de que las dos primeras sean oro será:

$$p_1\times p_2=(10/40)\times(9/39)$$

Siguiendo con esta lógica, la probabilidad de que las dos primeras cartas sean de oro y las dos últimas no lo sean es:

$$\begin{aligned}p_1\times p_2\times(1-p_3)\times(1-p_4)&=(10/40)\times(9/39)\times(1-8/38)\times(1-8/37)\\&=(10/40)\times(9/39)\times(30/38)\times(29/37)\end{aligned}$$

Aquí es importante notar que cualquier otro orden tiene la misma probabilidad. Consideremos, por ejemplo, el caso en que la primera carta y la última sean oro. En este caso se tiene una probabilidad de:

$$\begin{aligned}p_1\times(1-p_2)\times(1-p_3)\times p_4&=(10/40)\times(1-10/39)\times(1-9/38)\times(9/37)\\&=(10/40)\times(30/39)\times(29/38)\times(9/37)\end{aligned}$$

Podemos notar que los denominadores son los mismos, y lo que se modificó fue solamente el orden de los numeradores.

De este modo, en base a las reglas de conteo, sabemos que hay 6 maneras en que se pueden ordenar 4 elementos tomados de a dos:

$$\binom{4}{2} = \frac{4!}{2!2!} = 6$$

Finalmente, la probabilidad buscada es:

$$P(X=2) = 6 \times (10/40) \times (9/39) \times (30/38) \times (29/37) \cong 0,2142$$

Por suerte, no será necesario el trabajo tedioso utilizado en el ejemplo anterior, ya que existe una manera directa de calcular las probabilidades en estos casos, y es mediante la distribución *Hipergeométrica*.

Antes de exponer la función de probabilidad, haremos algunos comentarios en relación al dominio y los parámetros relevantes.

En primer lugar, como se ha mencionado, los parámetros relevantes son el tamaño de la población, h, el tamaño de la muestra extraída, m, y la cantidad de "éxitos" que hay en la población, a. En el ejemplo previo, se extrajo una muestra de tamaño $m=4$ cartas de una población con $h=40$ elementos, siendo $a=10$ la cantidad de elementos favorables al evento "es oro".

Si se extraen m elementos, obviamente se obtendrán como máximo m "éxitos". Además, si se extrae una muestra de tamaño superior al número de éxitos en la población ($m > a$), lógicamente la cantidad de éxitos en la muestra no podrá superar la cantidad de éxitos de la población. De esta manera, el dominio serán los enteros desde 0 hasta el mínimo entre m y a, ambos inclusive.

Por ejemplo, si se extraen cuatro cartas ($m=4$) y se analiza la cantidad de oros ($a=10$), el máximo de éxitos será cuatro, que es el mínimo entre 4 y 10. Sin embargo, si se extraen seis cartas ($m=6$), y se analiza la cantidad de ases que salen ($a=4$), el máximo de éxitos será cuatro, porque por más que se sigan sacando cartas nunca se podrán extraer más de 4 ases. En este caso, el máximo de éxitos es 4, que es el mínimo entre $m=6$ y $a=4$.

En la siguiente Tabla se ilustra lo comentado en los párrafos precedentes:

Dominio	**Parámetros**		
$D=\{0,1,2,...,\min(m;a)\}$	m (tamaño de la muestra)	a (cantidad de éxitos en la población)	h (tamaño de la población)

A continuación, exponemos la función de probabilidad de una variable aleatoria hipergeométrica. La deducción se encuentra en el Apéndice del final del capítulo.

Consideremos una población de tamaño h, donde a elementos son favorables a un evento (hay a "éxitos" en la población), y los restantes $h-a$ no lo son. Supongamos que se extrae una muestra de tamaño m.

Sea X la variable aleatoria que representa la cantidad de éxitos obtenidos en la muestra. Entonces X sigue una **distribución hipergeométrica**, cuya función de probabilidad está dada por:

$$P(X=k|m,a,h)=\frac{\binom{a}{k}\binom{h-a}{m-k}}{\binom{h}{m}}$$

$$=\frac{\frac{a!}{k!(a-k)!}\times\frac{(h-a)!}{(m-k)!\left[h-a-(m-k)\right]!}}{\frac{h!}{m!(h-m)!}}$$

con $k=0,1,...,m$; $k\le a$; $m-k\le h-a$ ($h;m;a$ enteros positivos).

Las condiciones impuestas a los parámetros al final de la definición anterior son bastante obvias. La cantidad de éxitos de la muestra no puede superar la cantidad de éxitos de la población ($k\le a$), y la cantidad de fracasos de la muestra no puede superar la cantidad de fracasos de la población ($m-k\le h-a$). Finalmente, como se mencionó anteriormente, la cantidad de éxitos muestrales está limitada al tamaño de la muestra: $k=0,1,...,m$.

Ejemplo 13

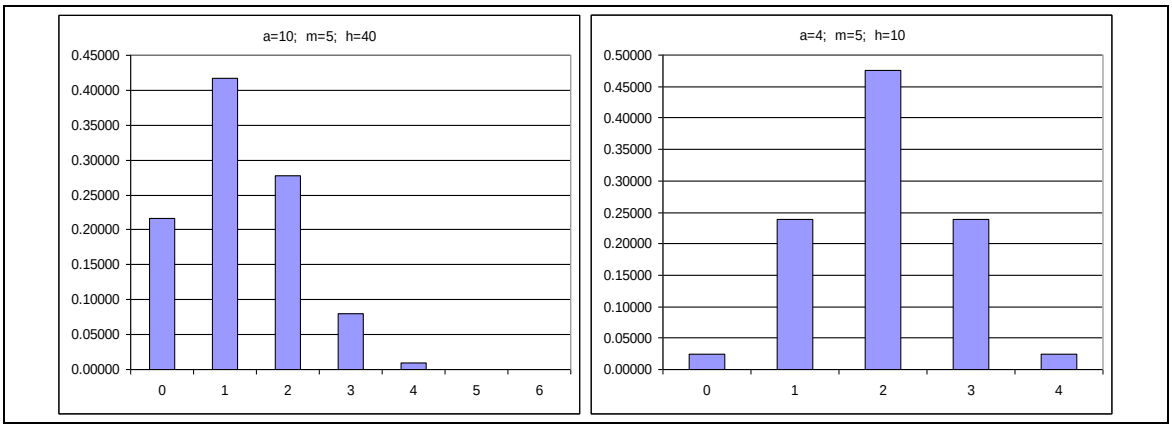
Supongamos que se extraen 5 cartas (sin reposición) de una baraja española y se desea calcular la probabilidad de que 3 sean espadas.

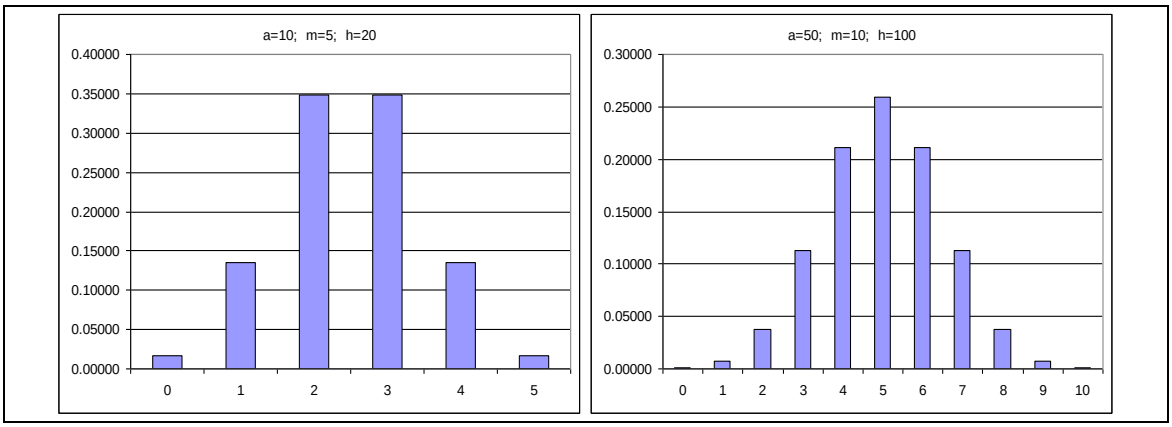
El tamaño de la población es $h=40$, la cantidad de éxitos es $a=10$ y el tamaño de la muestra es $m=5$ Se desea conocer la probabilidad de que la variable X (número de espadas) sea igual a 3.

Utilizando la definición anterior:

$$P(X=3|m=5,a=10;h=40)=\frac{\binom{10}{3}\binom{30}{2}}{\binom{40}{5}}\cong 0,07933$$

En la siguiente figura, se ilustra la distribución para distintos valores de los parámetros





La función de distribución es simplemente la suma de los valores de probabilidad para los cuales la variable es menor o igual al valor de interés:

$$\begin{aligned} F_H\left(k|m,a,h\right) &= P\left(X \le k|m,a,h\right) \\ &= \sum_{j=0}^{k} P\left(X = j|m,a,h\right) \\ &= \sum_{j=0}^{k} \frac{\binom{a}{j}\binom{h-a}{m-j}}{\binom{h}{m}} \end{aligned}$$

Ejemplo 14

Supongamos que de los 100 alumnos que están actualmente cursando estadística, 25 han aprobado Análisis Matemático I con una nota superior o igual a 7. Si se extraen 10 alumnos distintos al azar (sin reposición), ¿cuál es la probabilidad de que más de 5 hayan aprobado Análisis Matemático I con nota superior o igual a 7? En este caso $h=100$, $a=25$ y $m=10$. La probabilidad deseada es ¡menor al 1,5%!:

$$\begin{aligned} P\left(X > 5|m=10,a=25,h=100\right) &= 1 - P\left(X \le 5|m=10,a=25,h=100\right) \\ &= 1 - F_H\left(5|m=10,a=25,h=100\right) \\ &\cong 1 - 0,98551 \\ &\cong 0,01449 \end{aligned}$$

En términos de los parámetros, se puede determinar las principales características de la distribución. Para una demostración, véase Canavos (1997).

Sea X una variable aleatoria con distribución **hipergeométrica** con parámetros m, a y h. Entonces

$$E\left(X|m,a,h\right) = \frac{m \times a}{h}$$

$$Var\left(X|m,a,h\right) = \frac{m \times a \times \left(h-a\right)}{h^2} \times \frac{h-m}{h-1}$$

Ejemplo 15

Considerando el ejemplo anterior, la esperanza y la varianza de la variable "número de alumnos que aprobaron Análisis Matemático I nota superior o igual a 7" son:

$$E\left(X|m=10,a=25,h=100\right) = \frac{10 \times 25}{100} = 2,5$$

$$Var\left(X\middle|m=10,a=25,h=100\right)=\frac{10\times25\times75}{100^2}\times\frac{90}{99}\cong1,7045$$

Finalmente, las medidas de forma están dadas por las siguientes relaciones[28]:

Sea X una variable aleatoria con distribución **hipergeométrica** con parámetros m, a y h. Entonces el coeficiente de asimetría es

$$As\left(X\middle|m,a,h\right)=\frac{\left(h-2a\right)\left(h-2m\right)\sqrt{h-1}}{\left(h-2\right)\sqrt{m\times a\left(h-a\right)\left(h-m\right)}}$$

y el coeficiente de curtosis está dado por:

$$Ku\left(X\middle|a,m,h\right)=\frac{h^2\left(h-1\right)}{m\times a\left(h-2\right)\left(h-3\right)\left(h-a\right)\left(h-m\right)}\times$$
$$\times\left\{h\left(h+1\right)-6m\left(h-m\right)+3\frac{a}{h^2}\left(h-a\right)\times\left[h^2\left(m-2\right)-h\times m^2+6m\left(h-m\right)\right]\right\}$$

Aproximación por la distribución binomial

Como habrá visto, la distribución hipergeométrica tiene un grado de complicación superior a la binomial. Por suerte, para ciertos valores de los parámetros, la primera distribución puede aproximarse con esta última.

En primer lugar, notemos que si definimos a la probabilidad del primer éxito como:

$$q=a/h$$

Entonces la esperanza y la varianza pueden expresarse como:

$$E\left(X\middle|m,a,h\right)=m\times q$$

$$Var\left(X\middle|a,m,h\right)=m\times q\times\left(1-q\right)\times\frac{h-m}{h-1}$$

La esperanza entonces es idéntica a la de una distribución binomial con parámetros m y $p=a/h$. La varianza, en cambio, es más pequeña que la de una distribución binomial con los parámetros mencionados, debido al factor $\frac{h-m}{h-1}$. Sin embargo, cuando $h\to\infty$, las varianzas coinciden también. De manera más general, se demuestra que existe una convergencia entre las funciones funciones de probabilidad de ambas distribuciones.

La función de probabilidad de la distribución hipergeométrica converge a la distribución binomial cuando el tamaño de la población tiende a infinito:

$$\lim_{h\to\infty}P_H\left(X=k\middle|h,m,a\right)=P_B\left(X=k\middle|n=m,p=a/h\right)$$

La demostración de la proposición anterior no se expondrá en esta obra. La misma puede consultarse en Canavos (1997) o Novales (2000).

Ejemplo 16

Considere una empresa que produce 1500 artefactos por día. Al finalizar la producción diaria se realiza un control de calidad. Los ingenieros responsables del mismo afirman que diariamente hay 50 productos

[28] No ahondaremos aquí sobre las mismas. El lector interesado puede consultar Canavos (1997).

fallados. Si se extrae una muestra de tamaño 40 , los ingenieros están en lo correcto, la probabilidad de obtener 3 artículos defectuosos es:

$$P_H\left(X=3 \middle| m=40, a=50, h=1500\right) \cong 0{,}10442$$

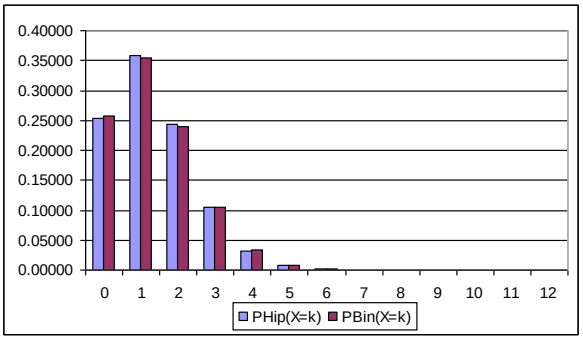
Si se utiliza la distribución binomial para aproxima esta probabilidad, los parámetros serían $n=40$ y $p=50/1500=1/30$, y la probabilidad resultante es:

$$P_B\left(X=3 \middle| n=40, p=1/30\right) \cong 0{,}10438$$

En la tabla siguiente, se comparan ambas distribuciones. Se puede observar que la probabilidad de obtener más de ocho artículos defectuosos es despreciable en ambos casos (la tabla debería continuar hasta $k=40$, que es el tamaño muestral, pero se han omitido los valores con probabilidad despreciable). En la figura se comparan los gráficos de ambas distribuciones, notándose la similitud de las mismas.

Parámetros y Características				
Hipergeométrica			Binomial	
m	*a*	*h*	*n*	*p*
40	50	1500	40	1/30
E (X)	*V (X)*		*E (X)*	*V (X)*
1 1/3	1.2554		1 1/3	1.2889

Probabilidades		
k	PHip(X=k)	PBin(X=k)
0	0.25301	0.25767
1	0.35863	0.35541
2	0.24268	0.23898
3	0.10442	0.10438
4	0.03211	0.03329
5	0.00751	0.00827
6	0.00139	0.00166
7	0.00021	0.00028
8	0.00003	0.00004
9	0.00000	0.00000
10	0.00000	0.00000
11	0.00000	0.00000



2.4.5 Distribución de Poisson.

Una de las distribuciones discretas más importantes por la cantidad de aplicaciones que posee es la Distribución de Poisson. Mencionamos, a continuación, algunas de sus aplicaciones:

- Líneas de espera.
- Compañías de Seguros.
- Evolución de Nacimientos y Muertes.
- Mercado de Capitales.
- Control de Calidad.

En términos generales, la distribución de Poisson está asociada a eventos “raros” que ocurren de manera independiente a una tasa constante en el tiempo o en el espacio. La variable cuenta la cantidad de veces que ocurre el evento en un intervalo (de tiempo o espacio) determinado. Al ser una variable que indica cantidad, su dominio son los números naturales incluyendo el cero (enteros no negativos). La función de probabilidad cuenta con un solo parámetro, λ, el cual representa la media y la varianza.

Dominio	Parámetros
$D=\{0,1,2,\ldots\}$	λ (intensidad de ocurrencia)

A continuación, se define formalmente la distribución de Poisson.

Consideremos cierto evento aleatorio que ocurre de manera independiente a una tasa constante en el tiempo (o el espacio). Sea X el número de eventos que ocurren. Entonces, X sigue una **distribución de Poisson** cuya función de probabilidad es:

$$P(X=k|\lambda)=\frac{\exp(-\lambda)\times\lambda^k}{k!} \qquad k=0,1,2.... \qquad \lambda>0$$

Otra aplicación importantísima de esta variable es que permite aproximar la distribución binomial cuando el tamaño de la muestra es grande y la probabilidad de éxito en cada ensayo es pequeña. Esta utilidad será analizada más adelante.

Antes de continuar, exponemos lo mencionado más arriba en cuanto a las características de la distribución.

Sea X una variable aleatoria con distribución **Poisson** con parámetro λ. Entonces,

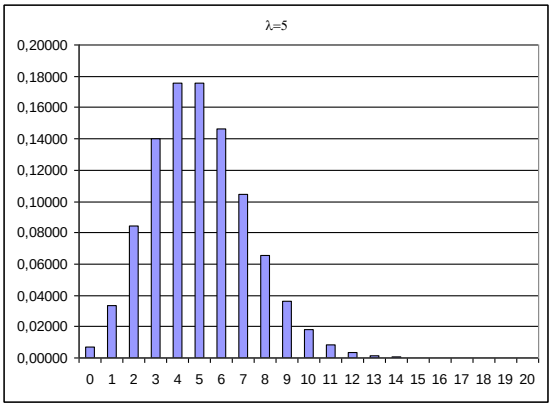
$$E(X|\lambda)=\lambda$$

$$Var(X|\lambda)=\lambda \quad \Rightarrow \qquad d.e.(X|\lambda)=\sqrt{\lambda}$$

Ejemplo 17

Supongamos que, en promedio, ocurren 5 fallas diarias en una línea de montaje y que la cantidad de fallas diarias es una variable aleatoria X que sigue una distribución de Poisson. ¿Cuál es la probabilidad de obtener exactamente tres fallas en un día? ¿Y la probabilidad de obtener cinco fallas?

Para responder, en primer lugar se debe tener en cuenta que la cantidad promedio representa el parámetro, por lo que $\lambda=5$. Luego, simplemente aplicamos la fórmula expuesta:



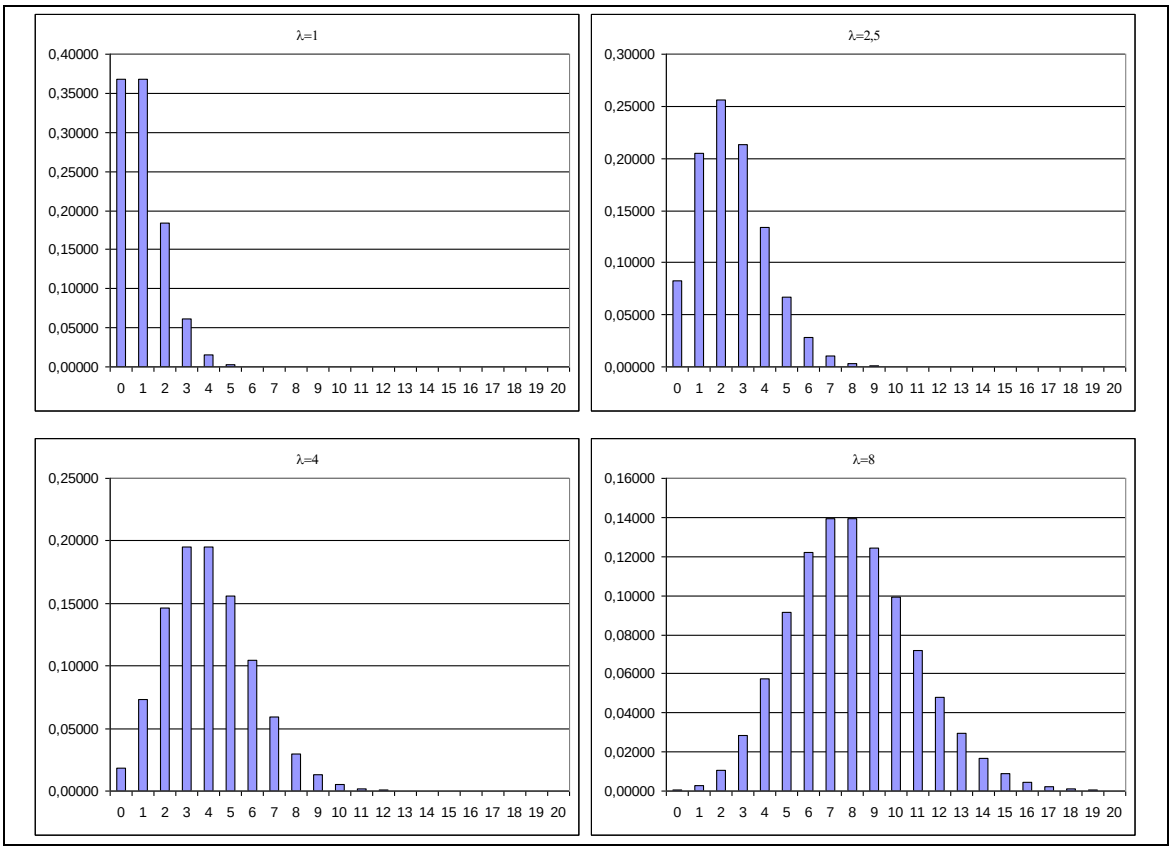
$$P(X=3|\lambda=5)=\frac{\exp(-5)\times 5^3}{3!}\cong 0,14037$$

$$P(X=5|\lambda=5)=\frac{\exp(-5)\times 5^5}{5!}\cong 0,17547$$

La función de probabilidad se observa en la figura anterior.

En el ejemplo previo puede observarse una característica importante de la distribución. Si bien el dominio continúa hasta el infinito, existe un gran sesgo hacia la derecha, y las probabilidades correspondientes a valores altos se hacen despreciables rápidamente.

En la siguiente figura, se compara la forma de la distribución para distintos valores del parámetro λ.



El parámetro de la distribución de Poisson está íntimamente relacionado con la amplitud del intervalo temporal (o espacial) al cual se refiere el fenómeno bajo estudio, y en caso de que se desee la probabilidad referida a otro período (o longitud), simplemente se debe adaptar el parámetro proporcionalmente. Así, en el ejemplo anterior, $\lambda = 5$ para la distribución de fallas diarias. Si se desea la distribución de las fallas semanales, se deberá utilizar $\lambda = 25$ (asumiendo 5 días laborales en la semana). El ejemplo siguiente ilustra esta característica.

Ejemplo 18

Consideremos un torneo de fútbol en el que se convierte un promedio de 2,5 goles por partido. Supongamos que la cantidad de goles sigue una distribución de Poisson. ¿Cuál es la probabilidad de que en 5 partidos se conviertan 10 goles? Si en total se juegan 20 partidos, ¿Cuántos goles se esperan en el torneo? ¿Cuál es el desvío estándar de la variable aleatoria: Cantidad de goles?

La variable aleatoria a analizar es: Cantidad de goles convertidos en un partido, que sigue una distribución de Poisson

Para responder la primera pregunta debemos encontrar el parámetro adecuado. Si el promedio es de 2,5 por partido, entonces, en cinco partidos deberíamos utilizar el parámetro $\lambda_5 = 5 \times 2,5 = 12,5$. Luego, simplemente utilizamos la fórmula:

$$P\left(X_5 = 10 \middle| \lambda_5 = 12,5\right) = \frac{\exp(-12) \times 12^{10}}{10!} \cong 0,09564$$

Si se considera el torneo completo, entonces, el parámetro adecuado para 20 partidos es $\lambda_{20} = 20 \times 2,5 = 50$. Al ser el parámetro igual a la esperanza matemática y a la varianza, tenemos que:

$$E\left(X_{20} \middle| \lambda = 50\right) = 50$$

$$Var\left(X_{20} \middle| \lambda_{20} = 50\right) = 50 \quad \Rightarrow d.e.\left(X_{20} \middle| \lambda_{20} = 50\right) = \sqrt{50} = 5\sqrt{2} \cong 7,0711$$

En el siguiente ejemplo se considera un caso en el cual debe analizarse si resulta apropiado utilizar dicha distribución.

Ejemplo 19

Consideremos la cantidad de goles por equipo y por partido de un campeonato mundial de fútbol. Cada partido corresponde a dos observaciones, una por cada equipo. Nos afirman que la cantidad de goles que marca *cada* equipo en un partido es una variable aleatoria Poisson. Para confirmarlo, tomamos los datos de Alemania 2006, los cuales se observan en el cuadro. Así, hubo 48 equipos que durante un partido terminaron con el marcador en 0, hubo 36 equipos que terminaron un partido con un gol a favor, y así sucesivamente.

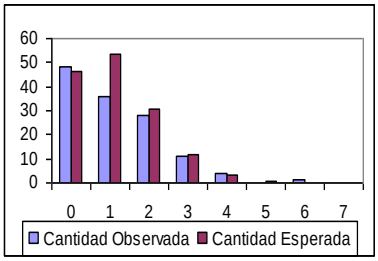
Goles	Frecuencia
0	48
1	36
2	28
3	11
4	4
5	0
6	1
7	0

Con estos datos, se puede calcular el promedio, sumando el producto de cada marcador por la cantidad de observaciones, y luego dividiendo por el total. Esto nos da:

$$\frac{0\times 48+1\times 36+...+7\times 0}{128}=1{,}1484375$$

Es decir, que, en promedio, cada equipo que juega convierte aproximadamente 1,15 goles. Recordemos que en una distribución de Poisson, el parámetro λ es la media de la distribución, por lo que con estos datos, esperamos que la cantidad de goles por equipo por partido sea una variable aleatoria con distribución Poisson con $\lambda=1{,}15$. En la siguiente tabla, en la tercera columna se observa la probabilidad asociada a cada posible marcador (por equipo y por partido), si la misma fuera una Poisson, mientras que en la cuarta se observan las cantidades esperadas en un total de 128 observaciones, calculadas multiplicando este valor por la probabilidad teórica correspondiente. Vemos que los valores observados no difieren sustancialmente de aquéllos esperados si la variable "cantidad de goles por partido y por equipo en un mundial" fuera una Poisson con media 1,15, salvo en el caso de un gol. En consecuencia, resultaría apropiado utilizar esta distribución. El gráfico ilustra mejor esta idea.

Goles	Cantidad Observada	Prob. Teórica	Cantidad Esperada
0	48	0,31664	46,55
1	36	0,36413	53,53
2	28	0,20938	30,78
3	11	0,08026	11,80
4	4	0,02307	3,39
5	0	0,00531	0,78
6	1	0,00102	0,15
7	0	0,00019	0,03



Aproximación de la distribución binomial

Cuando una variable aleatoria presenta las características correspondientes a una distribución binomial, en ciertas ocasiones se pueden aproximar las probabilidades mediante la distribución de Poisson.

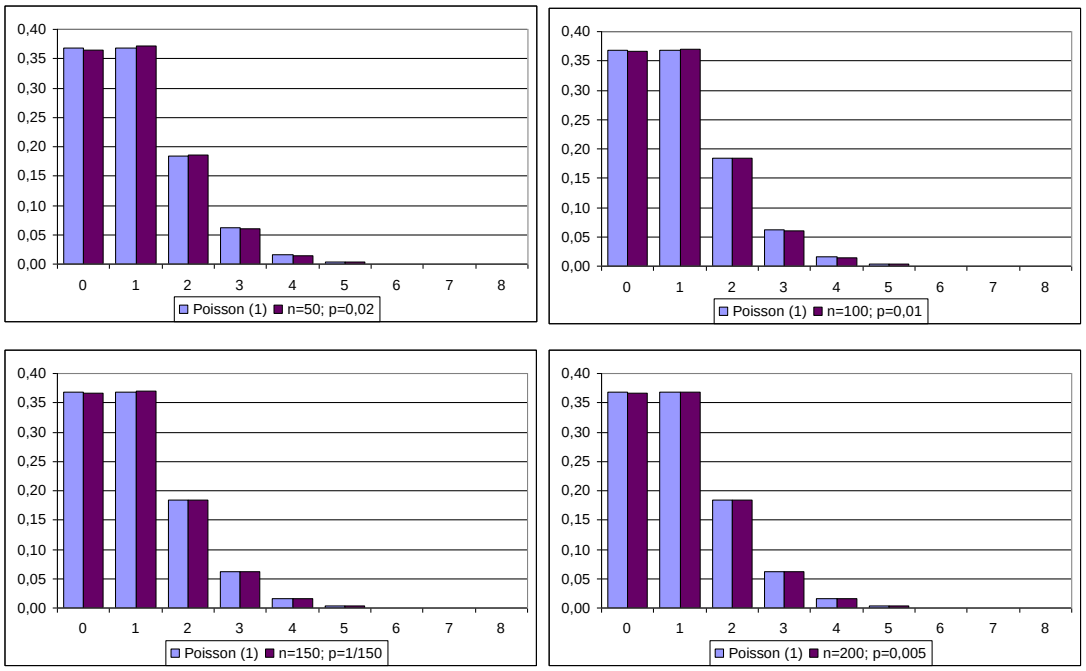
La función de probabilidad Binomial converge a la distribución de Poisson cuando el tamaño de la muestra, n, tiende a infinito y la probabilidad de éxito, p, tiende a cero, manteniéndose constante el producto de ambas $\lambda=n\times p$:

$$\lim_{\substack{n\to\infty\\ p\to 0}} P_B\left(X=k\,|n,p\right)=P_P\left(X=k\,|\lambda=n\times p\right)$$

La demostración de la relación precedente se encuentra en el Apéndice al final de este capítulo.

En las siguientes tabla y figura se comparan las funciones de probabilidad para distintos valores de n y p, manteniendo el producto constante en $\lambda=n\times p=1$.

k	Binomial				Poisson
	n=50	n=100	n=150	n=0,005	λ
	p=0,02	p=0,01	p=0,0067	n=200	1,00
0	0,36417	0,36603	0,36665	0,36696	0,36788
1	0,37160	0,36973	0,36911	0,36880	0,36788
2	0,18580	0,18486	0,18456	0,18440	0,18394
3	0,06067	0,06100	0,06111	0,06116	0,06131
4	0,01455	0,01494	0,01507	0,01514	0,01533
5	0,00273	0,00290	0,00295	0,00298	0,00307
6	0,00042	0,00046	0,00048	0,00049	0,00051
7	0,00005	0,00006	0,00007	0,00007	0,00007
8	0,00001	0,00001	0,00001	0,00001	0,00001
9	0,00000	0,00000	0,00000	0,00000	0,00000



Ejemplo 20

En un control de calidad, diariamente se extraen 50 artículos para ser analizados. Si más de 3 artículos son defectuosos, se deberá detener el proceso y analizar las causas de los defectos para corregirlas. Si en un día determinado el 1% de los artículos son defectuosos, ¿cuál es la probabilidad de detener el proceso?

Lo que se debe calcular en definitiva es la probabilidad de que la variable X, "cantidad de artículos defectuosos", sea mayor a tres. Esta variable puede considerarse binomial, con parámetros $n=50$ y $p=0,01$. La probabilidad de que haya más de tres artículos defectuosos es:

$$\begin{aligned} P_B\left(X \geq 3 \middle| n=50,\, p=0,01\right) &= 1-F_B\left(3 \middle| n=50,\, p=0,01\right) \\ &\cong 1-0,99840 \\ &= 0,00160 \end{aligned}$$

Si aproximáramos esa probabilidad utilizando la distribución de Poisson con parámetro $\lambda = 50 \times 0,01 = 0,5$, obtendríamos:

$$P_P\left(X \geq 3 \middle| \lambda = 0,5\right) = 1-F_P\left(3 \middle| \lambda = 0,5\right) \cong 1-0,99825 = 0,00175$$

Como puede apreciarse ambas probabilidades son muy cercanas. Con lo cual, podríamos afirmar que hay aproximadamente un 0,2% (redondeando 0,0016 y 0,00175) de que el proceso se detenga.

2.5 Distribuciones Continuas

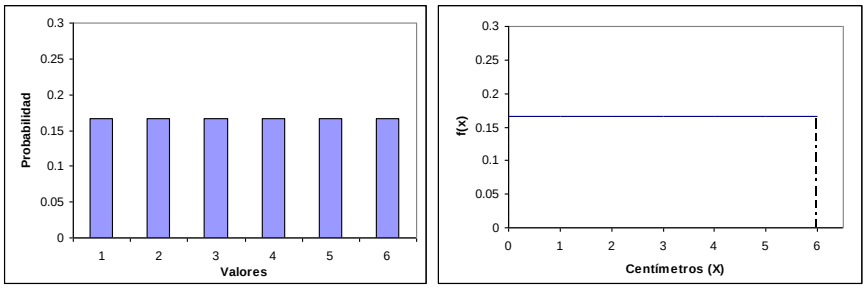
En esta sección analizaremos las principales distribuciones asociadas a variables aleatorias continuas. Para definir una variable continua, debemos caracterizar su dominio y su función de densidad o de distribución (ver Capítulo 2).

Recordemos que este tipo de variable está generalmente asociada a datos que provienen de observaciones originadas en un proceso de medición, en oposición a las variables discretas que, en general, se asocian a un proceso de conteo.

2.5.1 Distribución Uniforme

Como primera aproximación a las funciones continuas podemos describir a la Distribución Uniforme. En este caso, la función de densidad refleja que cada uno de los posibles valores de la variable tiene igual probabilidad de ocurrencia. De la misma manera que al arrojar un dado existe igual probabilidad de obtener cualquiera de los seis posibles valores, al considerar la distribución uniforme cualesquiera dos *intervalos* de igual longitud tienen la misma probabilidad.

Por ejemplo, si el crecimiento diario, en centímetros, de una determinada especie vegetal puede estar comprendido entre 0 y 6 sin predominar un valor sobre otro, la distribución sería uniforme. En la siguiente figura, se comparan la función de probabilidad de un dado y la función de densidad uniforme correspondiente al crecimiento de la planta.



Puede observarse, en la figura anterior, que la función de densidad uniforme es una constante, c, para todo el rango de valores para el cual la variable está definida (esta definición, que puede hacerse sobre cualquier intervalo real, dependerá del significado asignado). Al tratarse de una función de densidad, debe verificarse que:

$$\int_a^b c\ dx = 1$$ siendo a y b los extremos del intervalo considerado

Tal como se detalla en el Apéndice, la constante para la cual es verdadera la condición anterior es $\frac{1}{b-a}$.

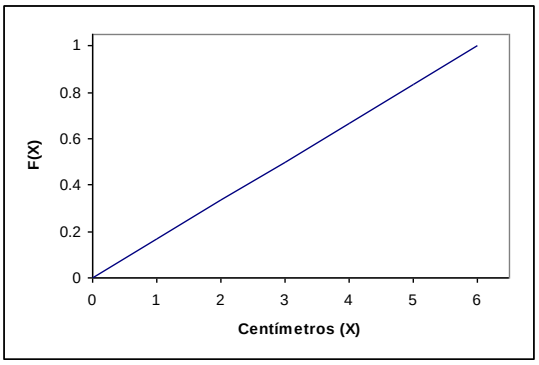
Se considera un evento aleatorio que puede tomar un valor real en el rango $[a\ ;\ b]$, teniendo cualesquiera dos intervalos de igual longitud la misma probabilidad de ocurrencia. Sea X la variable aleatoria representativa de ese evento. Entonces X está **uniformemente distribuida**, con función de densidad:

$$f\left(x|a,b\right) = \frac{1}{b-a} \qquad a \le x \le b$$

Realizando la operación correspondiente, puede verse que la función de distribución es una función lineal de la forma:

$$F\left(x|a,b\right) = \frac{x-a}{b-a} \qquad a \le x \le b$$

Retornando a nuestro ejemplo del crecimiento diario de un vegetal, la gráfica de la función de distribución se observa a continuación:



Dada la simplicidad de esta distribución, es muy sencillo proceder al cálculo de sus momentos.

Sea X una variable aleatoria con distribución **Uniforme** con parámetros a y b. Entonces,

$$E\left(X|a,b\right)=\frac{a+b}{2}$$

$$Var\left(X|a,b\right)=\frac{\left(b-a\right)^2}{12} \qquad \Rightarrow \qquad d.e.\left(X|a,b\right)=\frac{b-a}{\sqrt{12}}$$

Una de las principales aplicaciones de esta distribución está en la generación de números al azar para procedimientos de simulación, muestreo y programación entre otras. En general, los procedimientos de simulación se inician seleccionando aleatoriamente cualquier número entre el 0 y el 1. De manera que cada uno de esos valores tenga igual probabilidad de ocurrencia al momento de realizar la programación, es necesario considerar que se distribuyen uniformemente en el intervalo $[0;1]$.

Ejemplo 21

Tomemos el caso anterior de la simulación para iniciarnos en esta distribución. La función de densidad va a ser de la forma:

$$f\left(x|a=0;b=1\right)=1 \qquad 0\le x\le 1$$

La esperanza de esta variable, conforme a las definiciones dadas en el apartado, es igual a $\frac{0+1}{2}=\frac{1}{2}$ y la varianza $\frac{1}{12}$.

Utilizando una planilla de Microsoft® Excel puede observar este comportamiento. Puede generar una gran cantidad de números aleatorios utilizando la función "ALEATORIO( )"[29] y calcular el promedio entre ellos. Presionando repetidas veces la tecla F9 se generarán nuevos números. Podrá comprobar que el promedio está en todos los casos cercano a $\frac{1}{2}$ y la varianza a $\frac{1}{12}$.

2.5.2 Distribución Exponencial

La distribución exponencial sirve especialmente para fenómenos en los cuales la probabilidad de observar un valor elevado disminuye rápidamente. Esto es, la probabilidad de que la variable esté comprendida en un intervalo cercano a cero, será mayor a la probabilidad de que se encuentre en un intervalo alejado del origen, siempre que los intervalos que se comparen sean del mismo tamaño.

La distribución exponencial tiene como dominio a todos los números reales positivos, y posee un único parámetro asociado a la escala:

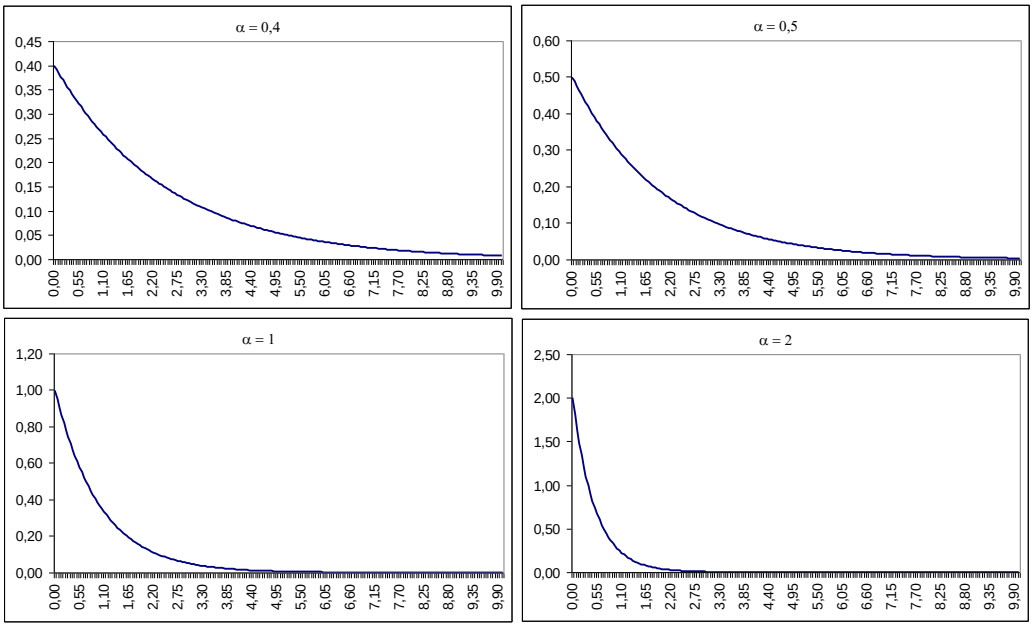
[29] La función correspondiente para la versión en inglés de Microsoft® Excel es "RAND( )".

Dominio	Parámetros
$D=[0;+\infty)$	$\alpha>0$ (escala de la función de densidad)

Una variable aleatoria X tiene **distribución exponencial**, si su función de densidad es:

$$f\left(x|\alpha\right)=\alpha\cdot\exp\left(-\alpha\cdot x\right) \qquad x\geq 0;\quad \alpha>0$$

En la siguiente figura, se ilustra la función de densidad para distintos valores del parámetro. En la misma puede observarse que cuanto mayor es el parámetro α, la distribución se encuentra más concentrada cerca del origen, es decir, que la probabilidad de ocurrencia de valores elevados es muy pequeña.



Una aplicación importante de esta distribución está dada por su relación con la variable Poisson: el tiempo que transcurre entre ocurrencias de un proceso de Poisson con intensidad λ, se distribuye exponencialmente con parámetro $\alpha=\lambda$. De manera más general, la distribución exponencial suele utilizarse para modelar el tiempo que transcurre entre fallas de cierto proceso. El parámetro α es la frecuencia de las fallas, mientras que su inverso, $1/\alpha$, es el tiempo promedio entre fallas (ver Canavos, 1997).

En base a la función de densidad expuesta, se puede obtener la **función de distribución**:

$$F\left(x|\alpha\right)=P\left(X\leq x|\alpha\right)=\int_0^x \alpha\exp\left(-\alpha y\right)dy=1-\exp\left(-\alpha x\right)$$

Ejemplo 22

Suponga que las fallas en una línea de producción ocurren en promedio cada 2 horas. Se desea determinar la probabilidad de que una falla ocurra antes de la primera hora de operación, y la probabilidad de que la producción se realice de manera continua por 4 horas sin ninguna falla.

De acuerdo a las características de este problema, el tiempo entre fallas es exponencial. Siendo el tiempo promedio entre fallas la inversa del parámetro, tenemos que $\alpha=1/2$. Denotaremos con la letra T al tiempo entre fallas. De este modo, la probabilidad de que una falla ocurra antes de la primera hora es:

$$P(T \le 1|\alpha = 1/2) = F_{\exp}(1|\alpha = 1/2)$$
$$= 1 - \exp(-1/2 \times 1)$$
$$\cong 0,3935$$

Por otro lado, la probabilidad de que la producción continúe por 4 hs. sin fallas es bastante baja, 0,1353 aproximadamente:

$$P(T \ge 4|\alpha = 1/2) = 1 - F_{\exp}(4|\alpha = 1/2)$$
$$= 1 - \left[1 - \exp(-1/2 \times 4)\right]$$
$$= \exp(-1/2 \times 4)$$
$$\cong 0,1353$$

La **esperanza** y la **varianza** de la distribución exponencial son:

$$E(X|\alpha) = 1/\alpha$$

$$Var(X|\alpha) = 1/\alpha^2 \qquad \Rightarrow \qquad d.e.(X|\alpha) = 1/\alpha$$

Las fórmulas anteriores plasman los comentarios realizados anteriormente: el tiempo promedio es la inversa del parámetro, y cuanto mayor es el parámetro, más concentración existe en la distribución (más pequeña es la varianza).

Otra aplicación de la distribución exponencial puede darse en análisis de supervivencia.

Ejemplo 23

Suponga que la tasa de mortalidad es independiente de la edad de la persona y la variable T es el tiempo al fallecimiento, entonces, la función de densidad que refleja su distribución es de la forma:

$$f(t) = \alpha \cdot e^{-\alpha \cdot t}$$

Si se sabe que el tiempo promedio al fallecimiento es de cincuenta años, ¿cuál es la probabilidad de fallecer antes de que transcurran sesenta años?

El primer paso para la resolución es deducir el valor del parámetro α. Si $E(X|\alpha) = 1/\alpha$ y se sabe que $E(X|\alpha) = 50 \Rightarrow \alpha = \frac{1}{50} = 0,02$.

Luego, utilizando la función de distribución podemos calcular la probabilidad deseada:

$$F(60|\alpha = 0,02) = 1 - \exp(-0,02 \times 60) = 0,6988$$

2.5.3 Distribución Normal

La distribución Normal es la más conocida y ampliamente utilizada, especialmente, cuando se realizan estudios sobre una población basados en una muestra aleatoria extraída de la misma. La amplia utilización se debe en parte al Teorema Central del Límite que será estudiado en capítulos posteriores. Además, debido a la forma de su función de densidad, esta distribución resulta apropiada para muchos fenómenos sociales, económicos, biológicos, etc., que puedan ser objeto de estudios estadísticos.

El dominio de una variable que posee distribución Normal está dado por todos los números reales, y sus dos parámetros, μ y σ, están asociados a la ubicación y a la escala de la función de densidad, siendo su forma siempre similar a la de una "campana":

Dominio	Parámetros	
$D = \mathbb{R}$	$\mu \in \mathbb{R}$ (posición)	$\sigma > 0$ (escala)

La variable aleatoria X tiene distribución **Norma**l, si su función de densidad es:

$$f\left(x|\mu,\sigma\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad x \in \mathbb{R};\quad \mu \in \mathbb{R}, \sigma > 0$$

Se recuerda que las probabilidades se calculan mediante la integración de la función de densidad. Sin embargo, la integración de la función anterior es sumamente complicada, por lo cual se recurren a aproximaciones de la misma, a través de las cuales se construyeron tablas que permitieron obtener de manera directa los valores. La mayoría de los libros de estadística incluyen apéndices con estas tablas. En esta obra no se incluirán, debido a que las probabilidades pueden calcularse con cualquier software que posea funciones estadísticas (Microsoft® Excel o SPSS, por ejemplo).

Los valores tabulados corresponden a la variable **estandarizada**, es decir, aquélla cuya media es igual a cero y el desvío estándar es igual a uno. Para transformar una variable Normal con media μ y desvío σ en una variable Normal Estándar, simplemente debemos restar la media y dividir por el desvío.

Si X es una variable Normal con media μ y desvío σ, entonces

$$Z = \frac{X-\mu}{\sigma}$$

Es una variable Normal Estándar, es decir, Normal con media 0 y desvío 1.

La función de distribución Normal Estándar, cuyos valores son los que se exponen en las tablas estadísticas, es:

$$F(x|\mu = 0, \sigma = 1) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\exp(-x^2)\,dx \qquad x \in \mathbb{R}$$

De este modo, al momento de buscar la probabilidad acumulada hasta un determinado valor en la tabla, el primer paso será la estandarización de la misma. De esta manera, si la variable X se distribuye con media 25 y desvío estándar 7, entonces la variable $z = (X-25)/7$ se distribuye con media cero y desvío uno. Si quiere conocerse cuál es la probabilidad de que la variable X tome un valor inferior a 20, deberá buscarse en las tablas la probabilidad acumulada correspondiente al valor:

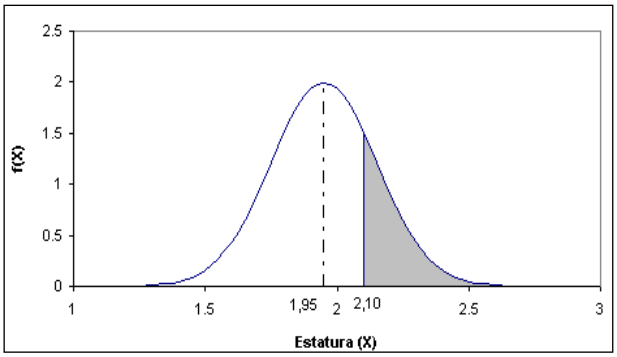
$$z = (20-25)/7 = -5/7$$

Ejemplo 24

Consideremos que la variable aleatoria *X* representa la altura de los jugadores de básquet y se distribuye normalmente con media 1.95 m y desvío estándar de 20cm. ¿Cuál es la probabilidad de que un jugador de básquet tenga una altura superior a 2,10m?

Responder a la pregunta planteada equivale a calcular:

$$P\left(X > 2{,}10\right) = 1 - F\left(2{,}1 \middle| \mu = 1{,}95; \sigma = 0{,}2\right)$$



Utilizando Microsoft® Excel directamente podremos calcular $F\left(2,1|\mu=1,95;\sigma=0,2\right)=0,7734$, con lo cual la probabilidad de que un jugador tenga una altura superior a 2,10 m. es de 0,2266. Al mismo resultado llegamos en caso de recurrir a las tablas. Pero para ello antes debemos estandarizar la variable, de modo tal que:

$$F_X\left(2,1|\mu=1,95;\sigma=0,2\right)=F_Z\left(\frac{2,1-1,95}{0,2}|\mu=0;\sigma=1\right)=0,7734$$

Como se mencionó anteriormente, los parámetros están asociados a la posición y a la escala de la función de densidad. Esta característica de los parámetros se refleja más claramente teniendo en cuenta que los mismos representan la media y el desvío estándar de la variable.

La **esperanza** y la **varianza** de una variable aleatoria con distribución **Normal** están dadas por:

$$E\left(X|\mu;\sigma\right)=\mu$$

$$Var\left(X|\mu;\sigma\right)=\sigma^2 \qquad \Rightarrow \qquad d.e.\left(X|\mu;\sigma\right)=\sigma$$

2.5.4 Distribución Gamma

Antes de introducirnos en la distribución en sí, primeramente, definiremos la "función gamma", ya que la función de densidad de una variable con distribución gamma utiliza dicha función.

Se llama **función gamma** a la siguiente integral, que se lee "gamma de *a*":

$$\Gamma\left(a\right)=\int_0^{+\infty}u^{a-1}\exp\left(-u\right)du$$

Algunas propiedades de la función gamma son: (ver Canavos, 1997):

1. $\Gamma\left(a\right)=\left(a-1\right)!$ $\qquad a$ entero positivo
2. $\Gamma\left(a\right)=\left(a-1\right)\cdot\Gamma\left(a-1\right)$ $\qquad a>0$
3. $\Gamma\left(0,5\right)=\sqrt{\pi}$

La distribución gamma tiene dos parámetros que la caracterizan, uno asociado a la escala y otro asociado a la forma, y el dominio está dado por los reales positivos:

Dominio	**Parámetros**	
$D=[0;+\infty)$	$\alpha>0$ (escala)	$\beta>0$ (forma)

La variable aleatoria x tiene **distribución gamma**, si su función de densidad es:

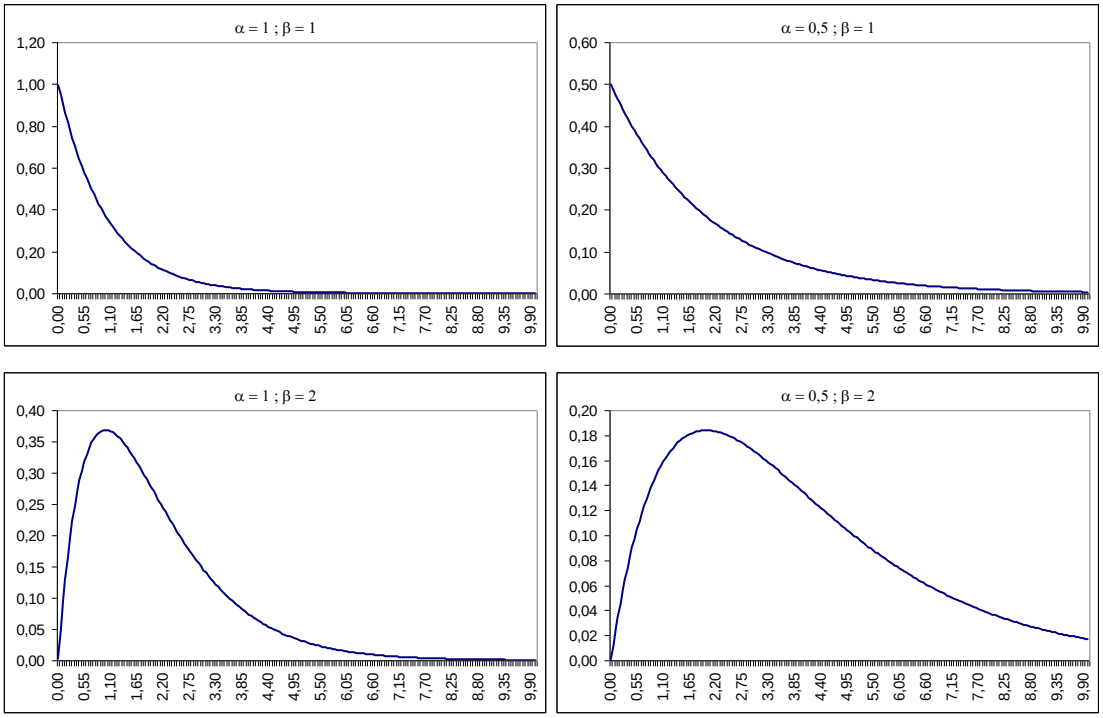
$$f\left(x|\alpha,\beta\right)=\frac{\left(x\alpha\right)^{\beta}}{x\Gamma\left(\beta\right)}\exp\left(-\alpha x\right)\qquad x>0;\ \alpha,\beta>0$$

Los valores de la función de densidad y las probabilidades acumuladas (función de distribución) pueden calcularse con Microsoft® Excel[30]. Sin embargo, en este programa la función de densidad está expresada como:

$$f\left(x|\alpha,\beta\right)=\frac{x^{\alpha_{Excel}-1}}{\beta_{Excel}{}^{\alpha_{Excel}}\Gamma\left(\alpha_{Excel}\right)}\exp\left(-x/\beta_{Excel}\right)=\frac{\left(x/\beta_{Excel}\right)^{\alpha_{Excel}}}{x\Gamma\left(\alpha_{Excel}\right)}\exp\left(-x/\beta_{Excel}\right)$$

Por lo tanto, puede apreciarse que $\alpha_{Excel}=\beta$ y $\beta_{Excel}=1/\alpha$.

En la siguiente figura se observa la distribución gamma para distintos valores de los parámetros. El parámetro β es de forma, y el parámetro α de escala. En la figura, puede observarse la similitud de las dos primeras gráficas con la distribución exponencial expuesta en la Sección 2.5.2. En realidad, como puede apreciarse en la definición, **cuando $\beta=1$ la distribución gamma se reduce a la exponencial**.



En general, si $\beta\leq1$ la función decae rápidamente, mientras que cuando $\beta>1$ la función presenta un pico que ocurre cuando $x=(1/\alpha)\times(\beta-1)$. Cuando $\beta>1$, este último valor de x es donde la función de densidad alcanza su máximo, es decir, es la **moda** de la variable aleatoria. Cuando $\beta\leq1$ la variable no posee moda.

[30] La definición de la función de densidad en Microsoft® Excel está realizada de manera distinta a la expuesta aquí. Por lo tanto, para el uso de la función, deberá utilizarse como primer parámetro

Ejemplo 24

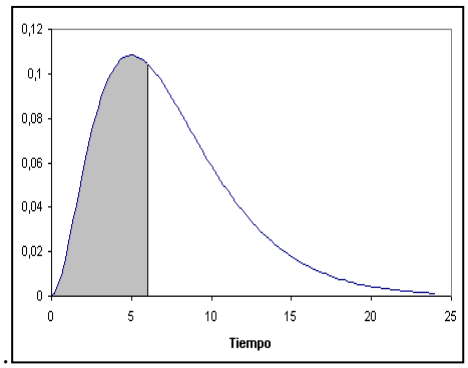
Una interesante aplicación de esta distribución se relaciona con la variable T_i, definida como el tiempo que media hasta la i-ésima ocurrencia de un evento que sigue un proceso de Poisson a lo largo del tiempo[31]. Asumamos que el evento consiste en la aparición de una falla en la línea de montaje, y deseamos analizar la variable T_3: "tiempo hasta la ocurrencia de la tercera falla". Suponga que el tiempo, en horas, hasta la aparición de la segunda falla sigue la distribución Gamma, con $\alpha = 0,4$ y $\beta = 3$. Es decir que:

$$f_{T_3}(t) = \frac{1}{2t} e^{-0,4t} (0,4t)^3$$

La probabilidad de que antes de 6 hs. ocurra la tercera falla es de aproximadamente:

$$\int_0^6 \frac{1}{2t} e^{-0,4t} (0,4t)^3 \, dt = 0,4303$$

Esta probabilidad se calculó con Microsoft® Excel utilizando $\alpha_{Excel} = 3$ y $\beta_{Excel} = 1/0,4$, y corresponde al área sombreada en la siguiente figura:



La **esperanza** y la **varianza** de una variable aleatoria con distribución **gamma** están dadas por:

$$E(X|\alpha,\beta) = \frac{\beta}{\alpha}$$

$$Var(X|\alpha,\beta) = \frac{\beta}{\alpha^2} \qquad \Rightarrow \qquad d.e.(X|\alpha,\beta) = \frac{\sqrt{\beta}}{\alpha}$$

Ejemplo 26

Si continuamos el ejemplo anterior, el tiempo esperado hasta la ocurrencia de la tercera falla se calcula como:

$$E(T_3|\alpha = 0,4; \beta = 3) = \frac{3}{0,4} = 7,5$$

Esto implica que, en promedio, la tercera falla ocurrirá a las 7 horas y media de iniciado el proceso de producción. La varianza y el desvío estándar son:

$$Var(T_3|\alpha = 0,4; \beta = 3) = \frac{3}{0,4^2} = 18,75 \quad \Rightarrow \quad d.e.(T_3|\alpha = 0,4; \beta = 3) = 4,33$$

[31] Ver Landro (2002)

Puede observarse que los desvíos respecto a la media son claramente significativos siendo el desvío estándar superior a la mitad del valor medio. Es decir, que el coeficiente de variación es mayor al 50%:

$$c.v.(T_3) = \frac{4,33}{7,5} = 0,5774$$

2.6 Anexo: Demostraciones

2.6.1 Deducción de la Función de Probabilidad de una Variable Hipergeométrica

La deducción de la función de probabilidad de una variable hipergeométrica surge de la definición clásica de probabilidad, donde se define la probabilidad de un evento como el cociente entre los casos favorables a la ocurrencia y el total de posibles casos.

El evento que se representa mediante una variable hipergeométrica consiste en que en una muestra (sin reposición) de m elementos de una población de tamaño h, x de esos elementos tengan una determinada característica, sabiendo que del total de la población sólo la presentan una cantidad a.

Los casos posibles para este evento definido están dados por todas las posibles muestras de m elementos, en donde lo que resulta relevante no es el orden en el cual se escojan a los mismos, sino los elementos que conforman cada grupo. Es necesario entonces recurrir a las reglas de conteo que se describen en el Capítulo 1: estamos frente a una **combinatoria de *h* elementos tomados de a *m***. En consecuencia:

$$Casos\ posibles = \binom{h}{m} = \frac{h!}{m!(h-m)!}$$

Determinar los casos favorables puede resultar un poco más complejo, pero, de todas formas, también obedece a la consideración de las reglas de conteo vistas. Las muestras que serán favorables a nuestra variable son aquéllas que contengan x éxitos y exactamente $(m-x)$ fracasos. Esto implica que estamos frente a la ocurrencia conjunta de esos dos eventos. Para calcular la cantidad de posibles combinaciones, debemos considerar que cualquier conjunto de x éxitos puede combinarse con cualquier conjunto de $(m-x)$ fracasos: la cantidad total surge del producto:

$$Cant.de\ grupos\ con\ "x"\ éxitos\ \times\ Cant.de\ grupos\ con\ "m-x"\ fracasos$$

Los grupos con x éxitos se pueden seleccionar del total de **a** elementos que presentan la característica en la población (una combinatoria de x elementos tomados de a a). De manera análoga, los grupos con $(m-x)$ fracasos se pueden seleccionar del conjunto de elementos que no presentan la característica favorable al evento, los cuales comprenden un total de $(h-a)$. En consecuencia,

$$Casos\ favorables = \binom{a}{x} \cdot \binom{h-a}{m-x} = \frac{a!}{x!(a-x)!} \times \frac{(h-a)!}{(m-x)!\left[(h-a)-(m-x)\right]!}$$

Claramente, entonces, se deduce la función de probabilidad que se había introducido en la Sección 4.4.4 del presente capítulo.

$$P(X = x|m,a,h) = \frac{\binom{a}{x}\binom{h-a}{m-x}}{\binom{h}{m}}$$

2.6.2 Relación Binomial - Poisson

Se definió en el presente capítulo que: "La función de probabilidad Binomial converge a la Poisson cuando el tamaño de la muestra, n, tiende a infinito y la probabilidad de éxito, p, tiende a cero, manteniéndose constante el producto de ambas $\lambda = n \times p$:

$$\lim_{\substack{n\to\infty \\ p\to 0}} P_B\left(X=k\middle|n,p\right)=P_P\left(X=k\middle|\lambda=n\times p\right)\text{"}$$

A continuación, demostraremos la expresión anterior. Recordemos que:

$$P_B\left(X=k\middle|n,p\right)=\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}$$

Vamos a expresar la media de la población, que suponemos constante, como $\lambda=n\cdot p$, con lo cual la expresión toma la forma:

$$P_B\left(X=k\middle|n,p\right)=\frac{n!}{k!(n-k)!}\cdot\left(\frac{\lambda}{n}\right)^k\cdot\left(1-\frac{\lambda}{n}\right)^{n-k}$$
$$=\frac{n!}{k!(n-k)!}\cdot\left(\frac{\lambda}{n}\right)^k\cdot\left(1-\frac{\lambda}{n}\right)^{n}\cdot\left(1-\frac{\lambda}{n}\right)^{-k}$$

Considerando ahora el límite de esta probabilidad cuando el tamaño de la población tiende a infinito, se tiene que:

$$\lim_{n\to\infty} P_B(X=k;n,p)=\lim_{n\to\infty}\frac{n!}{k!\,(n-k)!}\times\left(\frac{\lambda}{n}\right)^k\times\left(1-\frac{\lambda}{n}\right)^n\times\left(1-\frac{\lambda}{n}\right)^{-k}$$
$$=\lim_{n\to\infty}\frac{n\cdot(n-1)\ldots(n-k+1)}{k!}\times\frac{\lambda^k}{n^k}\times\left(1-\frac{\lambda}{n}\right)^n\times\left(1-\frac{\lambda}{n}\right)^{-k}$$
$$=\frac{\lambda^k}{k!}\times\lim_{n\to\infty}\underbrace{\frac{n\cdot(n-1)\ldots(n-k+1)}{n^k}}\times\left(1-\frac{\lambda}{n}\right)^n\times\left(1-\frac{\lambda}{n}\right)^{-k}\times$$

Tiende a 1 al ser dos polinomios de orden k con coeficiente principal igual a 1

$$\times\underbrace{\left(1-\frac{\lambda}{n}\right)^n\times\left(1-\frac{\lambda}{n}\right)^{-k}}$$

Tiende a $e^{-\lambda}$(propiedad de límite) $\to 1$

$$=\frac{\lambda^k}{k!}\times e^{-\lambda}$$
$$=P_P(X=k;\ \lambda=n\times p)$$

2.6.3 Constante en la Distribución Uniforme

Esta demostración es muy sencilla y se expone a fines ilustrativos con la intención de reflejar la importancia de que, al definir una función de densidad, la integral para todo el domino sea igual a uno.

Se ha observado, en la sección correspondiente, que, al considerar una distribución uniforme, cualesquiera de los valores del dominio de la variable aleatoria tienen igual probabilidad de ocurrencia. En consecuencia, se observó que la función de densidad se corresponde con una constante cuyo valor es igual a $\frac{1}{b-a}$ donde a y b son lo límites de integración del intervalo. Esto surge de la restricción impuesta para la función de densidad. Sea $f(x)=c$ la función de densidad, tenemos que, por definición:

$$\int_a^b c\,dx=1$$

En consecuencia, el valor de c surge de:

$$1 = \int_a^b c \; dx$$

$$1 = c \cdot x\Big|_a^b$$

$$1 = c \cdot (b - a) \Rightarrow \boxed{c = \frac{1}{b-a}}$$

En el capítulo anterior se presentaron los conceptos relacionados con las variables aleatorias y sus distribuciones de probabilidad. Asimismo, se expusieron las principales medidas que caracterizan a la distribución de una variable aleatoria (la media, la varianza, el coeficiente de asimetría, el coeficiente de curtosis, etc.).

En este capítulo se presentarán las principales distribuciones que suelen utilizarse en la práctica, las cuales surgen debido a ciertas características que presentan los fenómenos bajo estudio. En este sentido, existen ciertas funciones de probabilidad (discretas) o de densidad (continuas) que representan familias de distribuciones que dependen de los valores de ciertos parámetros. Así, cada familia resulta útil en distintos contextos, y el valor de los parámetros permite conocer no sólo la distribución de la variable aleatoria, sino también las principales medidas descriptivas de la misma.

Por ejemplo, la siguiente función de densidad corresponde a la familia de distribuciones exponenciales (más adelante se analizará en detalle esta familia), y la forma particular que adopte depende del valor del parámetro θ:

$$f(x) = \theta \cdot e^{-\theta x} \qquad (x \geq 0)$$

En el Ejemplo 22 del capítulo anterior, se presentó esta función en el caso particular en que $\theta = 0,5$.

Aquí supondremos que los parámetros que caracterizan a cada familia son conocidos. En capítulos posteriores se analizarán distintos métodos para calcular los valores de los mismos, o más precisamente “estimar” sus valores a partir de observaciones muestrales.

3 Descripción de Datos

Dario Bacchini
Lara Vazquez
Matías Larrá
Juana Llamas

La **Estadística Descriptiva** se utiliza para describir un conjunto de datos referidos a un fenómeno.

En este capítulo se realizará una descripción de los datos a través de ciertas medidas que resumen las principales características del conjunto de datos bajo estudio.

Si bien en este capítulo no se tratará la inferencia, el cálculo de las medidas numéricas que describen un conjunto de datos será fundamental cuando, en base a una muestra, deseemos inferir ciertas características de una población.

La descripción numérica de un conjunto de datos brinda gran información relacionada con la distribución de sus valores. Existen medidas que proporcionan una idea de la ubicación de la distribución, es decir, en torno a qué valor se encuentran distribuidos los datos. Estas medidas se conocen con el nombre de **Medidas de Posición**.

Por otro lado, están las **Medidas de Dispersión**, las cuales brindan información respecto a qué tan diseminados se encuentran los datos en relación con su ubicación central.

Finalmente, las **Medidas de Forma** indican, precisamente, la forma que tiene la distribución. Estas medidas permiten saber si hay tendencia a que los valores se agrupen a algunos de los lados de los valores centrales (simetría) y si los valores centrales tienen gran probabilidad de ocurrir o no (curtosis). Algunos autores incluyen estas medidas en las Medidas de Dispersión, lo cual es adecuado ya que indican cómo están dispersos los datos respecto de su centro. Sin embargo, aquí las expondremos en un apartado distinto para dar mayor sencillez a la exposición.

Las medidas numéricas descriptivas más importantes que estudiaremos se encuadran dentro de lo que se conoce como **Momentos** de una distribución de datos. A continuación, expondremos las fórmulas de cálculo de los mismos, las cuales quizás le resulten un tanto complejas.

El **Momento Absoluto (o Respecto del Origen) de Orden "r"**, ma_r , se calcula de la siguiente manera:

$$ma_r = \frac{1}{n}\sum_{i=1}^{M} x_i^r \cdot f_i$$

Donde x_i representa el i-ésimo valor observado, f_i la frecuencia del mismo, M es la cantidad de valores distintos observados y n es la cantidad total de valores observados contando repeticiones.

El **Momento Centrado (o Respecto de la Media) de Orden "r"**, mc_r , se calcula de la siguiente manera:

$$mc_r = \frac{1}{n}\sum_{i=1}^{M} \left(x_i - ma_1\right)^r \cdot f_i$$

Siendo ma_1 el momento absoluto de orden 1, y las demás variables como antes.

En el caso de contar con **datos agrupados**, x_i representará la marca de clase, y el cálculo de los Momentos será aproximado.

Antes de avanzar vamos a volver sobre algunos conceptos vistos en el **Prefacio**.

Población: Desde un punto de vista estadístico, una Población (o Universo) es la totalidad de individuos u objetos que se desea estudiar. O, más ampliamente, podemos pensar en toda la información disponible referida a un fenómeno.

Muestra: Es una parte de la Población que se ha seleccionado para ser analizada con el fin de obtener conclusiones respecto de la totalidad de los elementos de la misma.

Datos: Son la materia prima de cualquier estudio estadístico y, por ello, es fundamental la calidad de los mismos. Si los datos que utilizamos no son confiables, los resultados que obtengamos a partir de los mismos tampoco lo serán, aun utilizando las técnicas estadísticas más avanzadas.

Tipos de Datos

Datos Cualitativos: Son aquellos datos que no son intrínsecamente numéricos, no pueden ser sometidos a cuantificación y arrojan respuestas categóricas referidas a atributos de los elementos de la muestra o población.

Nominales: Los números identificadores de cada categoría son asignados sin poseer ningún tipo de jerarquía entre sí. Es decir, que la asignación es totalmente arbitraria.

Ordinales: Los números en la codificación de categorías se asignan de acuerdo a un orden que contiene información sobre la intensidad del atributo.

Datos cuantitativos: En este caso, los datos son intrínsecamente numéricos. Es decir que surgen de un proceso de medición o de conteo, y de acuerdo con ello, podemos separarlos en dos subcategorías:

Discretos: Surgen, generalmente, de un proceso de conteo. Los valores que puede asumir la variable pertenecen a un conjunto finito o infinito, pero numerable.

Continuos: Se obtienen mediante un proceso de medición, y pueden tomar infinitos valores dentro de un intervalo.

3.1 Distribuciones de Frecuencia

En primer lugar, nos interesará saber cuántas observaciones de cada valor se obtuvieron. Para ello, utilizaremos las frecuencias absolutas y relativas de cada uno de los valores de la variable, que indican la cantidad y el porcentaje de datos que se halló para cada valor obtenido.

La **frecuencia absoluta**, f_i, correspondiente a un valor x_i de la variable estudiada, es la cantidad de observaciones del mismo dentro del total de datos.

La **frecuencia relativa**, fr_i, de cada valor x_i, se obtiene dividiendo la correspondiente frecuencia absoluta, f_i, por el número total de observaciones, n, e indica la proporción de observaciones correspondientes a dicho valor.

$$f_{ri} = \frac{f_i}{n}$$

De manera general, se denominará **Distribución de Frecuencias** al conjunto de los valores de las variables y sus respectivas frecuencias, ya sean estas absolutas o relativas: $(x_i; f_i)$ ó $(x_i; f_{ri})$.

En el siguiente ejemplo, ilustramos los conceptos definidos precedentemente.

Ejemplo 1

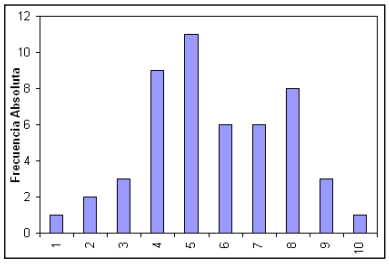
Consideremos las notas de los exámenes finales de la materia Estadística de un curso hipotético de la Universidad de Buenos Aires.

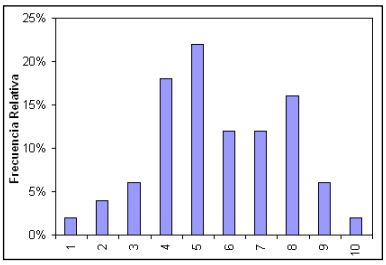
Datos de Notas				
1	4	5	6	8
2	4	5	6	8
2	4	5	7	8
3	4	5	7	8
3	4	5	7	8
3	5	5	7	8
4	5	6	7	9
4	5	6	7	9
4	5	6	8	9
4	5	6	8	10

Los valores que puede tomar la variable son los números enteros del 1 al 10, con lo cual se trata de una variable cuantitativa discreta. La frecuencia absoluta (o simplemente frecuencia) es la cantidad de veces que se repite cada valor y la frecuencia relativa será la cantidad de veces que se repite cada valor dividido por el número total de datos, que en este caso es 50. Con la información de los "Datos de Notas" podemos armar la Distribución de Frecuencias.

Valor x(i)	Frecuencia n(i)	Frec. Relativa f(i)
1	1	2,0%
2	2	4,0%
3	3	6,0%
4	9	18,0%
5	11	22,0%
6	6	12,0%
7	6	12,0%
8	8	16,0%
9	3	6,0%
10	1	2,0%

Si graficamos cada valor observado en el eje de abscisas y la frecuencia correspondiente (absoluta o relativa) en el eje de ordenadas, obtenemos los gráficos de barras representativos de la distribución de los datos.





Como podemos observar en el ejemplo anterior, los gráficos de frecuencias relativas y absolutas tienen exactamente la misma forma, salvo por los valores del eje de ordenadas. En el caso de frecuencias absolutas, el eje mencionado representa cantidades, mientras que en el gráfico de frecuencias relativas se representan porcentajes.

De acuerdo con las definiciones expuestas anteriormente, la suma de todas las frecuencias absolutas es igual al número total de datos, mientras que la suma de todas las frecuencias relativas es igual al 100% (o la unidad). En símbolos:

$$\sum_{i=1}^{M} f_i = n \qquad\qquad \sum_{i=1}^{M} f_{ri} = 1$$

Donde M indica la cantidad de valores distintos observados y n la cantidad total de datos contando las repeticiones.

Ejemplo 2

Considerando el ejemplo anterior, vemos que $M = 10$ pues hay diez notas distintas observadas, mientras que $n = 50$ es el total de alumnos evaluados.

Si sumamos las frecuencias absolutas obtenemos el número total de alumnos calificados:

$$\sum_{i=1}^{10} f_i = 1+2+3+9+11+6+6+8+3+1 = 50$$

A su vez, si sumamos las frecuencias relativas, obtenemos el 100%:

$$\sum_{i=1}^{10} f_{ri} = 0.02+0.04+0.06+0.18+0.22+0.12+0.12+0.16+0.06+0.02 = 1$$

Ejemplo 3

El siguiente cuadro compara la cantidad de explotaciones agropecuarias de cada provincia en los años 1998 y 2002 en base a los Censos Nacionales Agropecuarios realizados por el INDEC (Instituto Nacional de Estadísticas y Censos de la República Argentina).

En este caso, se observa que la variable de interés es la provincia, mientras que la frecuencia está relacionada con la cantidad de explotaciones agropecuarias que posee cada una.

En base a la información de la tabla anterior, la cual podría interpretarse como una distribución de frecuencias absolutas, construiremos las frecuencias relativas y los diagramas de barra correspondientes.

Para hacer más sencillo el análisis e ilustrar claramente los conceptos, trabajaremos solamente con los totales de cada año (primera columna) y utilizaremos únicamente las seis provincias con mayor cantidad de EAP[32]. Teniendo en cuenta estas restricciones, construimos la tabla de frecuencias.

Explotaciones agropecuarias (EAP) por tipo de delimitación, según provincia. Total del país. Años 1988 y 2002

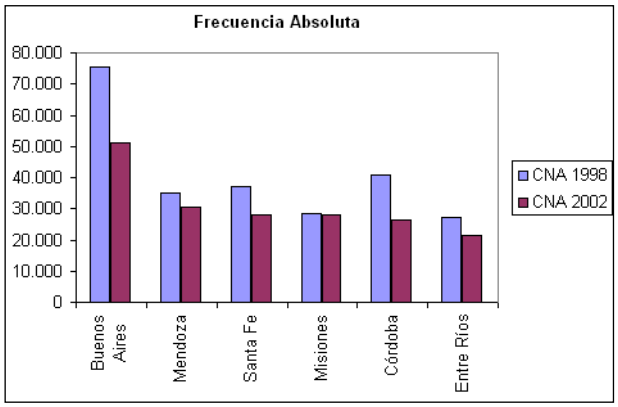
Provincia	Explotaciones agropecuarias					
	Censo Nacional Agropecuario 1988			Censo Nacional Agropecuario 2002		
	Total	Con límites definidos	Sin límites definidos	Total	Con límites definidos	Sin límites definidos
Total del país	**421,221**	**378,357**	**42,864**	**333,533**	**297,425**	**36,108**
Buenos Aires	75,531	75,479	52	51,116	51,107	9
Mendoza	35,221	33,249	1,972	30,656	28,329	2,327
Santa Fe	37,029	36,884	145	28,103	28,034	69
Misiones	28,566	27,517	1,049	27,955	27,072	883
Córdoba	40,817	40,061	756	26,226	25,620	606
Entre Ríos	27,197	27,134	63	21,577	21,577	-
Santiago del Estero	21,122	11,532	9,590	20,949	10,830	10,119
Chaco	21,284	17,595	3,689	16,898	15,694	1,204
Corrientes	23,218	22,070	1,148	15,244	14,673	571
Salta	9,229	4,798	4,431	10,297	5,575	4,722
Formosa	12,181	9,582	2,599	9,962	8,994	968
Tucumán	16,571	15,998	573	9,890	9,555	335
Catamarca	9,538	6,988	2,550	9,138	6,694	2,444
Jujuy	8,526	4,286	4,240	8,983	4,061	4,922
San Juan	11,001	10,300	701	8,509	7,927	582
La Rioja	7,197	5,374	1,823	8,116	5,852	2,264
La Pampa	8,718	8,632	86	7,775	7,774	1
Río Negro	9,235	7,709	1,526	7,507	7,035	472
Neuquén	6,641	2,530	4,111	5,568	2,198	3,370
San Luis	6,962	5,974	988	4,297	4,216	81
Chubut	4,241	3,484	757	3,730	3,574	156
Santa Cruz	1,114	1,102	12	947	944	3
Tierra del Fuego	82	79	3	90	90	-

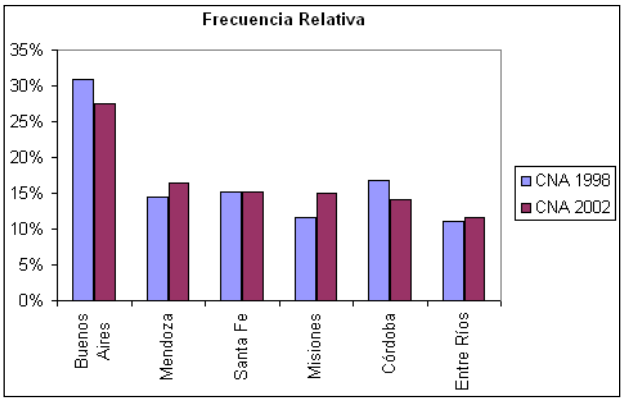
Fuente: INDEC, Censo Nacional Agropecuario 1988 y 2002.

TABLA DE FRECUANCIAS				
Explotaciones agropecuarias (EAP)				
	CNA 1998		CNA 2002	
Total	244.361	100%	185.633	100%
	Frec.	**Frec. Rel.**	**Frec.**	**Frec. Rel.**
Buenos Aires	75.531	30,91%	51.116	27,54%
Mendoza	35.221	14,41%	30.656	16,51%
Santa Fe	37.029	15,15%	28.103	15,14%
Misiones	28.566	11,69%	27.955	15,06%
Córdoba	40.817	16,70%	26.226	14,13%
Entre Ríos	27.197	11,13%	21.577	11,62%

[32] Las frecuencias relativas y el total de observaciones se calcularán en base a estas seis provincias solamente.

Con la información precedente, podemos construir un gráfico comparativo de las frecuencias relativas y absolutas de los dos años.





Comparando los dos gráficos, puede observarse que las barras de 1998 tienen la misma forma en ambos, y lo mismo ocurre con las barras del 2002 (ya hemos mencionado que las figuras de las frecuencias absolutas y relativas tienen la misma forma). Sin embargo, al comparar la relación que hay entre 1998 y 2002 las comparaciones en términos relativos y absolutos pueden ser distintas.

Por ejemplo, la cantidad de EAP de Mendoza en 2002 es menor que en 1998 (ver gráfico Frecuencia Absoluta), pero esa cantidad representa un porcentaje mayor del total de EAP analizadas (ver gráfico Frecuencia Relativa). Lo mismo sucede con las provincias de Misiones y Entre Ríos.

Este fenómeno se debe a que, si bien disminuye la frecuencia absoluta de un valor, n_i, también baja la cantidad total de observaciones, N, por lo cual el cociente entre ambos valores puede ser mayor. Dependiendo del análisis que queramos hacer, nos será más útil un gráfico que otro.

Continuando con el análisis de la distribución de frecuencias de un conjunto de datos, además de la frecuencia correspondiente a cada valor, es útil la información relacionada con la frecuencia de valores menores o iguales a una determinada observación, es decir, la frecuencia acumulada para un valor dado.

La **Frecuencia Acumulada** correspondiente al valor x_i es la suma de las frecuencias de todos los valores menores o iguales a x_i:

$$F_i = \sum_{k=1}^{i} f_k$$

La **Frecuencia Relativa Acumulada** correspondiente al valor x_i es la suma de las frecuencias relativas de todos los valores menores o igual a x_i:

$$F_{ri} = \sum_{k=1}^{i} f_i$$

Alternativamente, esta última se puede calcular como la frecuencia absoluta acumulada, F_i, dividida entre el total de observaciones n.

$$F_{ri} = \frac{F_i}{n}$$

Ilustremos las definiciones con los datos de las notas de los exámenes finales del Ejemplo 2 de más arriba.

Ejemplo 4

En base a las frecuencias calculadas para cada una de las notas, podemos obtener las frecuencias acumuladas simplemente sumando todas las frecuencias anteriores. Veamos algunos ejemplos de los valores con los cuales se construye la tabla de frecuencias acumuladas.

La frecuencia acumulada absoluta de 2, es la suma de las frecuencias de 1 y 2. La frecuencia relativa acumulada se obtienen dividiendo el valor calculado entre el número total de alumnos evaluados, es decir, 50.

$$F_2 = f_1 + f_2 = 1 + 2 = 3$$

$$F_{r2} = \frac{F_2}{50} = \frac{3}{50} = 0,06$$

La frecuencia acumulada de 4 es la suma de las frecuencias de 1, 2, 3 y 4. La frecuencia relativa acumulada se obtiene dividiendo el valor calculado entre 50.

$$F_4 = f_1 + f_2 + f_3 + f_4 = 1 + 2 + 3 + 9 = 15$$

$$F_{r4} = \frac{F_4}{50} = \frac{15}{50} = 0,30$$

La Tabla de Frecuencia Simples y Acumuladas es:

Valor x(i)	Frecuencia n(i)	Frecuencia Acumulada	Frec. Relativa f(i)	Frec. Relativa Acumulada
1	1	1	2,0%	2,0%
2	2	3	4,0%	6,0%
3	3	6	6,0%	12,0%
4	9	15	18,0%	30,0%
5	11	26	22,0%	52,0%
6	6	32	12,0%	64,0%
7	6	38	12,0%	76,0%
8	8	46	16,0%	92,0%
9	3	49	6,0%	98,0%
10	1	50	2,0%	100,0%

El cuadro anterior nos permite ver directamente la cantidad de alumnos que resultaron desaprobados en el examen, la cual está representada por la frecuencia acumulada del valor $x = 3$. Es decir que 6 alumnos resultaron insuficientes en el examen[33].

A su vez, observando en la columna correspondiente a la frecuencia relativa acumulada, podemos ver qué porcentaje obtuvo una nota inferior o igual a 3: 12%.

Como es lógico, siempre el mayor valor observado, x_M , tiene una frecuencia absoluta acumulada igual al número total de datos y una frecuencia relativa acumulada igual a la unidad.

$$F(x_M) = n$$

$$F_{ri}(x_M) = 1$$

Además, podemos observar que la frecuencia acumulada de un valor puede obtenerse sumando la frecuencia acumulada hasta el valor anterior y la frecuencia correspondiente al valor en cuestión:

$$F_i = F_{i-1} + f_i$$

Además, directamente de la fórmula anterior, vemos que la frecuencia de un valor determinado se puede obtener mediante la resta de la frecuencia acumulada hasta el mismo y la frecuencia acumulada hasta el anterior:

$$f_i = F_i - F_{i-1}$$

Ejemplo 5

Consideremos el ejemplo anterior para ilustrar las fórmulas. La frecuencia del valor 5 puede obtenerse mediante la resta de la frecuencia acumulada correspondiente a dicho valor y aquélla correspondiente al valor 4:

$$f_5 = F_5 - F_4 = 26 - 15 = 11$$

[33] En la Facultad de Ciencias Económicas de la UBA, los exámenes finales se aprueban con una nota superior o igual a 4 (cuatro).

La frecuencia acumulada de 9 puede obtenerse sumando la frecuencia acumulada hasta 8 y la frecuencia correspondiente a 9:

$$F_9 = F_8 + f_9 = 46 + 3 = 49$$

Veamos ahora qué sucede con las frecuencias acumuladas cuando las variables son cualitativas. Si el análisis se realiza sobre variables cualitativas cuyo ordenamiento no es obvio, la frecuencia acumulada pierde sentido ya que no podemos definir claramente cuándo un valor es menor que otro.

Considere, por ejemplo, el caso estudiado más arriba en relación a la cantidad de explotaciones agropecuarias de las distintas provincias de la República Argentina. Siendo los "valores" en dicho ejemplo los nombres de las provincias, no tiene sentido hablar de la frecuencia acumulada de Córdoba, porque no podemos decir que Buenos Aires sea menor o mayor que Córdoba, y lo mismo ocurriría con la frecuencia acumulada de cualquier otra provincia. Esto se debe a que la variable analizada es Cualitativa Nominal (ver Prefacio), cuyos valores son "etiquetas" que no pueden ordenarse sin recurrir a la arbitrariedad.

Una observación final al respecto: ¡No confunda la variable con su frecuencia! La variable estudiada es la provincia ("Buenos Aires", "Córdoba", etc.) y la frecuencia es la cantidad de EAP que posee la misma. Las frecuencias sí pueden ordenarse, pero la frecuencia acumulada se calcula para cada valor de la variable y esta última no posee ordenamiento.

Estas observaciones respecto a las variables cualitativas no se presentan cuando las mismas se pueden ordenar sin arbitrariedad. Como ilustración de esto consideremos el año que cursa un determinado alumno en la carrera de Economía. Esta variable es de tipo Cualitativa Ordinal (ver Prefacio). En este caso, sí se puede analizar la frecuencia acumulada hasta un determinado valor. Por ejemplo, nos puede interesar la frecuencia relativa acumulada hasta tercer año. Es decir, el porcentaje de alumnos que cursa primero, segundo y tercer año, o lo que es lo mismo, el porcentaje de alumnos que no ha alcanzado aún el cuarto año.

Cuando se analizan variables cuantitativas continuas (su valor puede ser cualquier número real), la frecuencia absoluta de cada uno de los valores será muy pequeña e incluso existirán muchísimos valores para los cuales no existan observaciones. Por ejemplo, supongamos que estamos analizando las estaturas de un grupo de 30 personas, y realizamos las mediciones con una precisión de 3 decimales. Es muy probable que todas las observaciones arrojen un valor distinto. En este caso, es conveniente agrupar los datos correspondientes a un intervalo de valores, para que la visualización de las distribuciones sea más adecuada. En el apartado siguiente, trataremos estos casos.

3.1.1 Distribuciones de Frecuencia con Datos Agrupados

Si deseamos analizar la estatura de los alumnos del curso de Estadística, en primer lugar mediremos a cada uno de los alumnos. Luego, cuando calculemos las frecuencias absolutas de cada valor, veremos que la mayoría de las observaciones son únicas y su cálculo no nos brinda ninguna idea respecto de la distribución de las estaturas. En este caso, resulta conveniente agrupar los datos en intervalos, y asignar una frecuencia (absoluta y relativa, simple o acumulada) a cada intervalo en lugar de a cada valor observado.

Los intervalos en los cuales se agrupan los datos se denominan **intervalos de clase.**

Cada intervalo tiene un **límite superior** y un **límite inferior**, asignándose al mismo todas las observaciones mayores o iguales al límite inferior e inferiores al límite superior.

Se denomina **marca de clase** al punto medio de cada intervalo de clase, es decir, al promedio simple entre el límite superior y el límite inferior.

Cada uno de los intervalos tendrá su **frecuencia de clase**, absoluta y relativa.

Finalmente, se denomina **amplitud** de un intervalo de clase a la diferencia entre el límite superior y el límite inferior. Es decir:

$$amplitud = w_i = L_{sup} - L_{inf}$$

Ejemplo 6

Considere los siguientes datos de las estaturas, en metros, de los alumnos del Ejemplo 1.

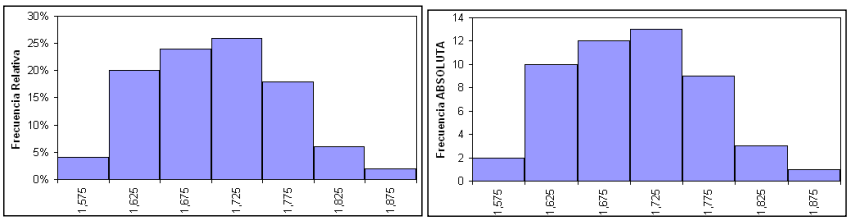
Datos de Estatura				
1,594	1,640	1,680	1,729	1,760
1,594	1,642	1,687	1,731	1,763
1,612	1,652	1,687	1,737	1,781
1,614	1,652	1,691	1,738	1,787
1,622	1,653	1,702	1,738	1,796
1,624	1,658	1,704	1,738	1,797
1,633	1,660	1,704	**1,740**	1,801
1,635	1,675	1,705	**1,752**	1,817
1,640	1,679	1,715	1,753	1,818
1,640	1,680	1,717	1,753	1,859

Si consideramos intervalos de 5 cm. de amplitud cada uno, podemos construir el siguiente cuadro de datos agrupados.

Clase		Marca	Frecuencia	Frecuencia	Frec. Relativa	Frec. Relativa
LI	LS	y(j)	n(j)	Acumulada	f(i)	Acumulada
1,55	1,60	1,575	2	2	4,0%	4,0%
1,60	1,65	1,625	10	12	20,0%	24,0%
1,65	1,70	1,675	12	24	24,0%	48,0%
1,70	**1,75**	**1,725**	**13**	**37**	26,0%	74,0%
1,75	1,80	1,775	9	46	18,0%	92,0%
1,80	1,85	1,825	3	49	6,0%	98,0%
1,85	1,90	1,875	1	50	2,0%	100,0%

Los gráficos de frecuencias absolutas y relativas para datos datos agrupados por intervalos se llaman histogramas

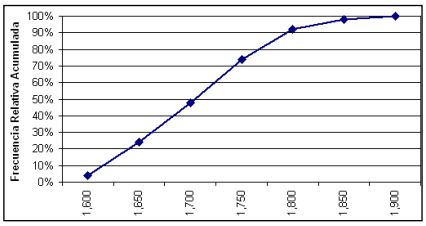
El **Histograma** de un conjunto de datos es un gráfico que representa en el eje de ordenadas la frecuencia (absoluta o relativa) y en el eje de abscisas los valores de la variable. Es decir, es el gráfico de la distribución de frecuencias.

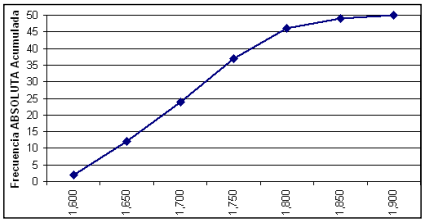


A su vez, las frecuencias acumuladas se grafican mediante la ojiva.

El **Polígono de Frecuencias Acumuladas u Ojiva** se construye marcando en primer lugar cada uno de los pares $(L_s; F_i)$ y, luego, uniendo los puntos marcados mediante líneas rectas.

El **Polígono de Frecuencias Relativas Acumuladas** es igual al anterior, pero utilizando los pares $(L_s; F_{ri})$.





Observamos que en los gráficos de frecuencias simples del ejemplo anterior se utilizaron las marcas de clase en el eje de abscisas, mientras que en los gráficos de frecuencias acumuladas se utilizaron los límites superiores. Esto se debe a que el punto medio de una clase no acumula toda la frecuencia de esa clase.

Consideremos por ejemplo la segunda clase, siendo el límite inferior 1,60m., el límite superior 1,65m., y la marca de clase el promedio entre estos dos últimos valores, 1,625m. La frecuencia acumulada de esta clase es 12, sin embargo, no hay 12 observaciones menores o iguales a 1,625m., sino que en realidad son menores a 1,65m.

En ocasiones, de acuerdo a la variable que se estudie, hay intervalos que poseen una gran cantidad de observaciones, mientras que otros quedarían vacíos distorsionando la distribución de frecuencias, en esos casos es conveniente utilizar intervalos de amplitudes diferenciales.

Un típico ejemplo de una variable que conviene agrupar con intervalos de distinta longitud es el ingreso familiar. Veamos un ejemplo.

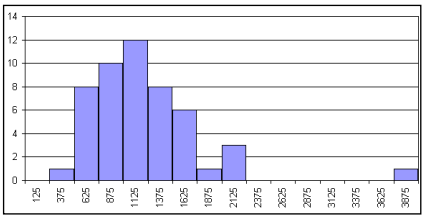
Ejemplo 7

Consideremos los siguientes datos correspondientes al nivel de Ingreso que percibe por su trabajo un grupo de 50 personas.

Datos de Ingresos				
$ 476,36	$ 757,57	$ 1.064,75	$ 1.241,23	$ 1.582,23
$ 525,98	$ 779,97	$ 1.067,14	$ 1.273,37	$ 1.587,23
$ 566,48	$ 828,98	$ 1.082,39	$ 1.297,33	$ 1.626,68
$ 592,74	$ 835,74	$ 1.141,95	$ 1.351,68	$ 1.700,32
$ 619,73	$ 844,95	$ 1.187,00	$ 1.414,08	$ 1.730,47
$ 624,75	$ 874,51	$ 1.195,29	$ 1.434,53	$ 1.751,05
$ 666,10	$ 879,40	$ 1.195,81	$ 1.441,13	$ 2.041,14
$ 669,69	$ 895,00	$ 1.219,68	$ 1.455,31	$ 2.094,69
$ 689,61	$ 945,62	$ 1.219,90	$ 1.497,00	$ 2.207,91
$ 756,76	$ 1.013,76	$ 1.236,16	$ 1.562,25	$ 3.794,98

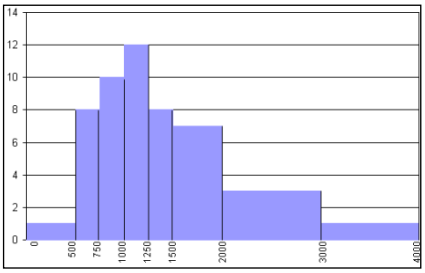
Es obvio que necesitamos agrupar los mismos, ya que no hay ningún valor repetido. Si consideramos intervalos de amplitud $ 250, obtenemos la siguiente tabla de frecuencias y el gráfico correspondientes.

Clase		Marca	Frecuencia	Frecuencia
LI	LS	x(i)	n(i)	Acumulada
0	250,00	125	0	0
250,00	500,00	375	1	1
500,00	750,00	625	8	9
750,00	1000,00	875	10	19
1000,00	1250,00	1125	12	31
1250,00	1500,00	1375	8	39
1500,00	1750,00	1625	6	45
1750,00	2000,00	1875	1	46
2000,00	2250,00	2125	3	49
2250,00	2500,00	2375	0	49
2500,00	2750,00	2625	0	49
2750,00	3000,00	2875	0	49
3000,00	3250,00	3125	0	49
3250,00	3500,00	3375	0	49
3500,00	3750,00	3625	0	49
3750,00	4000,00	3875	1	50



Puede notarse que muchas clases quedaron sin observaciones en el extremo superior, mientras que se observa una gran concentración en el extremo inferior de la distribución. Esto indica que quizás resulte más conveniente construir intervalos de distinta amplitud. Por ejemplo, podemos construir la siguiente tabla de frecuencias y su correspondiente gráfico.

Clase		Marca	Frecuencia	Frecuencia
LI	LS	x(i)	n(i)	Acumulada
0	500,00	250	1	1
500,00	750,00	625	8	9
750,00	1000,00	875	10	19
1000,00	1250,00	1125	12	31
1250,00	1500,00	1375	8	39
1500,00	2000,00	1750	7	46
2000,00	3000,00	2500	3	49
3000,00	4000,00	3500	1	50



3.2 Medidas de Posición

Aquí presentamos medidas relacionadas con la ubicación de los datos observados. Por ejemplo, si medimos la estatura de un grupo de estudiantes varones, es razonable pensar que las observaciones estarán en torno a los 1,75 m. Sin embargo, si medimos las estaturas de los miembros de un equipo de básquet, es más probable que las observaciones se agrupen cerca de los 2 m. de altura. De esta manera, estas medidas de posición nos indican dónde está ubicada, aproximadamente, la distribución de las observaciones.

Dentro de estas medidas de posición, podemos diferenciar entre las de **tendencia central** (indican cuál es el centro de la distribución de frecuencias) y las de **tendencia no central** (brindan información respecto de la ubicación de ciertos valores, no necesariamente centrales).

Las medidas de tendencia central que estudiaremos aquí son la **Media**, la **Mediana** y la **Moda**. En cuanto a las medidas de posición no central, se estudiarán los **Percentiles**.

3.2.1 Media Aritmética

La **media aritmética**[34] de un conjunto de datos se calcula como:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{M} x_i \cdot f_i$$

En el caso de que no existan datos repetidos, la frecuencia de cada valor es igual a 1, es decir que $f_i = 1$ para cualquier observación. Además, la cantidad de observaciones distintas coincide

[34] La media aritmética es el promedio simple entre los datos, y se lo suele denominar de distintas maneras: media, promedio o valor medio.

con la cantidad total de datos, es decir, que $M = N$. De esta manera, **si no hay observaciones repetidas**, la media aritmética es:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$$

Esta adaptación, cuando $f_i = 1$ y $M = N$, puede realizarse en varias fórmulas que se presentarán en lo que resta del capítulo. Sin embargo, no expondremos más la misma, para evitar ser reiterativos en la exposición.

Cabe mencionar que en realidad esta última forma se puede utilizar siempre, nada más que hay que prestar atención de sumar todas las repeticiones de un mismo valor. Podemos ver que el cálculo de la media aritmética de este modo es más intuitivo: simplemente sumamos todos los valores y dividimos por la cantidad de observaciones. La fórmula de más arriba, al utilizar las frecuencias, es más operativa, ya que en lugar de sumar f_i veces un mismo número, simplemente multiplicamos al mismo por dicho valor, el cual indica justamente la cantidad de veces que se repite. Veamos un ejemplo.

Ejemplo 8

Consideremos los datos referidos a las notas de los exámenes finales, las cuales se reproducen en la tabla.

Si sumamos todos los valores y, luego, dividimos dicha suma por la cantidad total de observaciones, $n = 50$, obtendremos el cálculo de la media aritmética sin utilizar las frecuencias:

Datos de Notas				
1	4	5	6	8
2	4	5	6	8
2	4	5	7	8
3	4	5	7	8
3	4	5	7	8
3	5	5	7	8
4	5	6	7	9
4	5	6	7	9
4	5	6	8	9
4	5	6	8	10

$$\overline{x} = \frac{1}{50}\sum_{i=1}^{50} x_i = \frac{1+2+2+3+3+3+...+9+9+9+10}{50} = \frac{284}{50} = 5,68$$

Sin embargo, utilizando las frecuencias la fórmula se reduce mucho (¡no hacen falta puntos suspensivos!), ya que en lugar de sumar varias veces un mismo número, simplemente los multiplicamos por la cantidad de veces que se repite, es decir, por su frecuencia. Utilizando esta técnica, el cálculo sería:

$$\overline{x} = \frac{1}{50}\sum_{i=1}^{50} x_i \cdot f_i = \frac{1\times1+2\times2+3\times3+4\times9+5\times11+6\times6+7\times6+8\times8+9\times3+10\times1}{50} = \frac{284}{50} = 5,68$$

Veamos ahora un ejemplo con datos agrupados.

Ejemplo 9

Consideremos los datos de estatura del Ejemplo 6 y el cálculo de las frecuencias realizado.

Si consideramos los datos sin agrupar de la primera tabla, al haber muy pocas observaciones repetidas, la utilización de las frecuencias en el cálculo no brinda ningún atajo. Por ello, simplemente sumamos todos los valores y dividimos por la cantidad de observaciones.

Datos de Estatura				
1,594	1,640	1,680	1,729	1,760
1,594	1,642	1,687	1,731	1,763
1,612	1,652	1,687	1,737	1,781
1,614	1,652	1,691	1,738	1,787
1,622	1,653	1,702	1,738	1,796
1,624	1,658	1,704	1,738	1,797
1,633	1,660	1,704	1,740	1,801
1,635	1,675	1,705	1,752	1,817
1,640	1,679	1,715	1,753	1,818
1,640	1,680	1,717	1,753	1,859

$$\overline{x} = \frac{1}{50}\sum_{i=1}^{50} x_i = \frac{1,594+1,594+1,612+\ldots+1,817+1,818+1,859}{50} = \frac{85,179}{50} = 1,704$$

Consideremos ahora los datos agrupados como en el Ejemplo 7. Tenemos 7 clases distintas, es decir que $M = 7$ (ver tabla). Si calculamos el producto de cada marca de clase por su frecuencia, obtenemos los valores de la última columna de la siguiente tabla. Luego, sumando dichos productos obtenemos el total ilustrado al final de la última columna.

Clase		Marca	Frecuencia	y(j) * n(j)
LI	LS	y(j)	n(j)	
1,55	1,60	1,575	2	3,150
1,60	1,65	1,625	10	16,250
1,65	1,70	1,675	12	20,100
1,70	1,75	1,725	13	22,425
1,75	1,80	1,775	9	15,975
1,80	1,85	1,825	3	5,475
1,85	1,90	1,875	1	1,875
			Suma =	85,250

$$\sum_{i=1}^{7} y_i \cdot f_i = 1,575\times 2 + 1,625\times 10 + \ldots + 1,875\times 1 = 3,150 + 16,250 + \ldots + 1,875 = 85,250$$

Finalmente, la suma calculada es divida por la cantidad total de datos, $n = 50$, para obtener la media aritmética aproximada:

$$\overline{y} = \frac{1}{50}\sum_{i=1}^{7} y_i \cdot f_i = \frac{85,250}{50} = 1,705$$

En el ejemplo se puede observar que la media calculada con datos agrupados no coincide exactamente con el promedio simple (calculado con los datos sin agrupar). Esto se debe a que al agrupar los datos, algo de información estamos perdiendo. Sin embargo, los cálculos son más sencillos cuando se trabaja con los datos agrupados. De esta manera, hay un intercambio entre sencillez y exactitud.

En la medida que la agrupación se realice de manera tal de reflejar la distribución de los datos, los cálculos realizados con datos agrupados estarán próximos a los valores verdaderos provenientes de todas las observaciones.

Finalmente, de manera alternativa, podemos calcular la media en términos de la frecuencia relativa. Esta última se definía como:

$$f_i = \frac{f_{ri}}{n}$$

Si remplazamos esta expresión en el cálculo de la media, obtenemos lo siguiente:

$$\overline{x} = \sum_{i=1}^{M} x_i \cdot f_{ri}$$

Antes de continuar, haremos una observación. Al inicio de esta sección hemos definido los momentos, mencionando que serían útiles para calcular las medidas de tendencia central, de dispersión y de forma. Podemos observar que la media aritmética definida más arriba es simplemente el momento absoluto de orden 1, ya que si remplazamos en la fórmula expuesta al principio de esta sección el valor "r" por un 1, obtenemos la fórmula del promedio. Por lo tanto, podemos escribir:

$$\overline{x} = ma_1$$

Luego, con esta observación podemos expresar los momentos centrados de la siguiente forma:

$$mc_r = \frac{1}{n}\sum_{i=1}^{M}\left(x_i - \overline{x}\right)^r \cdot f_i$$

Ya volveremos sobre estos últimos cuando veamos medidas de dispersión y de forma. Pero antes, veamos otras medidas de tendencia central.

3.2.2 Mediana

La media aritmética es la medida más ampliamente utilizada para indicar el centro de una distribución. Sin embargo, el cálculo de la misma es muy sensible a los valores extremos, entendiendo por éstos a aquéllos muy pequeños o muy grandes.

Consideremos, por ejemplo, la observación de las siguientes estaturas (en metros) de cuatro personas: 1.70, 1.72, 1.73 y 2.10. El promedio de estas observaciones es 1.81, pero seguramente esperaríamos un valor central cercano a 1.72 ó 1.73. La distorsión se debe a la observación de la estatura 2.10, la cual es muy grande en comparación con las demás.

Para superar este inconveniente, la mediana es una medida que utiliza los valores centrales de los datos ordenados para indicar el centro de la distribución.

La **Mediana** de un conjunto de datos, X_{me}, es el valor central cuando los datos están ordenados de manera creciente o decreciente.

Si la cantidad de datos es impar, simplemente ordenamos los datos y nos fijamos cuál queda en el medio. Si la cantidad de datos es par, no hay un único valor central, por lo cual la mediana será el promedio simple entre los dos valores centrales.

En la definición anterior vemos que la mediana es el valor de los datos que deja a la misma cantidad de datos por encima y por debajo. Es decir, que tiene una frecuencia acumulada de $\frac{n}{2}$ y una frecuencia relativa acumulada del 50%.

Ejemplo 10

Supongamos que contamos con 5 datos referidos a la estatura de los jugadores de un equipo de básquet. La mediana será simplemente el valor "del medio" cuando los datos estén ordenados. Si las estaturas observadas (ordenadas) son:

1,869	1,908	1,912	1,978	2,072

Entonces, la mediana es la tercera observación, es decir que $X_{me} = 1{,}912$.

En caso de que la cantidad de observaciones sea par, deberemos calcular el promedio entre los valores centrales. Si incluimos a los suplentes del equipo, las observaciones son:

1,869	1,901	1,912	1,978	2,072
1,889	1,908	1,921	1,994	2,075

Luego, la mediana será el promedio entre los dos valores centrales, es decir, entre la observación 5 y 6 en el conjunto de datos ordenados:

$$X_{me} = \frac{1{,}912 + 1{,}921}{2} = 1{,}9165$$

De manera más formal, podemos volver a definir la mediana en términos matemáticos.

Si contamos con n **observaciones ordenadas**, entonces, si n es impar, la mediana es:

$$X_{me} = x_{\frac{n+1}{2}}$$

O bien, cuando n es par:

$$X_{me} = \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)$$

Ejemplo 11

Tomemos en cuenta el ejemplo anterior. En primer lugar, pueden considerarse sólo los titulares del equipo siendo $n = 5$, y al ser impar debemos calcular $\frac{n+1}{2} = 3$. Luego, la mediana es igual a la tercera observación:

$$X_{me} = x_3 = 1{,}912$$

Luego, cuando se consideraron los suplentes y los titulares, $n=10$. Siendo el total de observaciones par, hay dos valores centrales dados por las posiciones $\frac{n}{2}=5$ y $\frac{n}{2}+1=6$ de los datos ordenados. De esta manera, la mediana es el promedio entre la quinta y la sexta observación:

$$X_{me}=\frac{x_5+x_6}{2}=\frac{1,912+1,921}{2}=1,9165$$

Consideremos un ejemplo con una mayor cantidad de datos, como los "Datos de Estatura" utilizados en la sección anterior.

Ejemplo 12

Los datos ordenados se observan en la tabla. Luego, al ser $n=50$ un número par, la mediana será el promedio entre las ubicaciones $\frac{n}{2}=25$ y $\frac{n}{2}+1=26$, las cuales están resaltadas en la tabla. Es decir:

$$X_{me}=\frac{x_{25}+x_{26}}{2}=\frac{1,702+1,704}{2}=1,703$$

Datos de Estatura				
1,594	1,640	1,680	1,729	1,760
1,594	1,642	1,687	1,731	1,763
1,612	1,652	1,687	1,737	1,781
1,614	1,652	1,691	1,738	1,787
1,622	1,653	**1,702**	1,738	1,796
1,624	1,658	**1,704**	1,738	1,797
1,633	1,660	1,704	1,740	1,801
1,635	1,675	1,705	1,752	1,817
1,640	1,679	1,715	1,753	1,818
1,640	1,680	1,717	1,753	1,859

Cuando solamente disponemos de los **datos agrupados**, la mediana estará dentro del primer intervalo que acumule una frecuencia mayor o igual a $\frac{n}{2}$. El cálculo de la misma es:

$$X_{me}=L_i+\frac{\frac{n}{2}-F_{j-1}}{f_j}\cdot w_j$$

Donde L_i es el límite inferior del intervalo que contiene la mediana, n es la cantidad total de observaciones, F_{j-1} es la frecuencia acumulada hasta el intervalo anterior, f_j es la frecuencia del intervalo, y w_j es la amplitud del intervalo.

Ejemplo 13

Los datos agrupados de estatura, del ejemplo 6, con sus respectivas frecuencias simples y acumuladas, se observan en la tabla.

	Clase		Marca	Frecuencia	Frecuencia
j	LI	LS	y(j)	n(j)	Acumulada
1	1,55	1,60	1,575	2	2
2	1,60	1,65	1,625	10	12
3	1,65	1,70	1,675	12	24
4	**1,70**	**1,75**	**1,725**	**13**	**37**
5	1,75	1,80	1,775	9	46
6	1,80	1,85	1,825	3	49
7	1,85	1,90	1,875	1	50

El cuarto intervalo es el primero que acumula $\frac{n}{2} = 25$ o más. La mediana entonces es:

$$X_{me} = L_{i4} + \frac{\frac{n}{2} - F_3}{f_4} \cdot w_4 = 1,70 + \frac{\frac{50}{2} - 24}{13} \cdot 0,05 = 1,7038$$

3.2.3 Moda

Tanto la media como la mediana se utilizan exclusivamente con datos cuantitativos, ya que no nos basta con conocer la categoría en la cual se encuentra una observación, sino que tenemos que conocer el valor de cada observación para realizar el cálculo. La moda tiene la principal ventaja de poder calcularse con datos cualitativos y en distribuciones que son relativamente simétricas, indicará la posición central de las observaciones.

La **Moda** o el **Modo** de un conjunto de datos, X_{mo}, es el valor más frecuente. Es decir, es aquel valor que tiene mayor frecuencia (tanto absoluta como relativa).

Puede presentarse el caso en que dos (o más) valores tienen la máxima frecuencia. En este caso decimos que la distribución es **bimodal** (o **multimodal**).

Ejemplo 14

Los datos de las notas de los alumnos, con sus respectivas frecuencias simples, son los expuestos en la tabla.

Valor	Frecuencia
x(i)	n(i)
1	1
2	2
3	3
4	9
5	**11**
6	6
7	6
8	8
9	3
10	1

La moda, es simplemente el valor más observado (el que tiene frecuencia máxima), es decir que $X_{mo} = 5$.

Cuando contamos con **datos agrupados**, existe una **clase modal**, la cual es la que posee mayor frecuencia.

La moda se calcula de la siguiente manera:

$$X_{mo} = L_i + \frac{d_1}{d_1 + d_2} \cdot w_i$$

Donde L_i es el límite inferior de la clase modal, d_1 es la diferencia entre la frecuencia del intervalo modal y la frecuencia del intervalo anterior y d_2 la diferencia entre la frecuencia del intervalo modal y la frecuencia del intervalo posterior.

Ejemplo 15

Los datos agrupados de estatura, con sus respectivas frecuencias simples, se observan en la tabla. La clase modal (aquélla con mayor frecuencia) es la cuarta. Por lo cual la moda es:

	Clase		Marca	Frecuencia
j	LI	LS	y(j)	n(j)
1	1,55	1,60	1,575	2
2	1,60	1,65	1,625	10
3	1,65	1,70	1,675	12
4	**1,70**	**1,75**	**1,725**	**13**
5	1,75	1,80	1,775	9
6	1,80	1,85	1,825	3
7	1,85	1,90	1,875	1

$$X_{mo} = L_{i4} + \frac{d_1}{d_1 + d_2} \cdot w_i = 1,70 + \frac{(13-12)}{(13-12)+(13-9)} \cdot 0,05 = 1,71$$

Observación: Si los intervalos tuviesen amplitudes diferenciales no serían comparables, ya que en un intervalo de mayor amplitud es más probable tener una mayor frecuencia, no porque los datos sean más frecuentes, sino que porque el intervalo tiene un rango mayor.

En estos casos se debe pensar a la frecuencia como el área del rectángulo correspondiente a ese intervalo de clase en el histograma, para luego calcular la altura h_i de cada uno usando la fórmula:

$$h_i = \frac{f_i}{w_i}$$

Finalmente el intervalo modal será el de mayor h_i, ocupando esta medida el lugar de las frecuencias en la fórmula de cálculo anteriormente planteada.

Pasamos ahora a analizar la medida de posición no central que estudiaremos aquí.

3.2.4 Percentiles

Cuando tenemos un conjunto de observaciones ordenadas de una variable cuantitativa, podemos calcular fácilmente el porcentaje de observaciones que se encuentran por debajo de un valor determinado, simplemente observando la frecuencia relativa acumulada hasta el mismo. En base a esta idea, se definen los percentiles, deciles y cuartiles de la distribución de frecuencias.

Los **Percentiles** dividen la distribución de frecuencias en cien partes iguales. El primero, P_1, acumula $\frac{n}{100}$ (1%), el segundo, P_2, $\frac{2 \cdot n}{100}$ (2%), y así hasta el último, P_{99}, que acumula $\frac{99 \cdot n}{100}$ (99%).

Ejemplo 16

Los datos de estatura ordenados se observan en la tabla, ordenados de menor a mayor.

Datos de Estatura				
1,594	1,640	1,680	1,729	1,760
1,594	1,642	1,687	1,731	1,763
1,612	1,652	1,687	1,737	1,781
1,614	1,652	1,691	1,738	1,787
1,622	1,653	1,702	1,738	1,796
1,624	1,658	1,704	1,738	1,797
1,633	1,660	1,704	**1,740**	1,801
1,635	1,675	1,705	**1,752**	1,817
1,640	1,679	1,715	1,753	1,818
1,640	1,680	1,717	1,753	1,859

Siendo $n = 50$, el percentil 10 es el valor que acumula $\frac{10 \cdot n}{100} = 5$ observaciones. Es decir que $P_{10} = 1,622$.

El percentil 75 acumula $\frac{75 \cdot n}{100} = 37,5$ observaciones. Al igual que con la mediana, calculamos el promedio simple entre la observación 37 y 38, obteniendo como resultado:

$$P_{75} = \frac{1,740 + 1,752}{2} = 1,746$$

Cuando solamente disponemos de los **datos agrupados**, el percentil estará dentro del primer intervalo que acumule una frecuencia mayor o igual a $\frac{n \cdot k}{100}$. El cálculo de la misma es:

$$P_k = L_i + \frac{\frac{n \cdot k}{100} - F_{j-1}}{f_j} \cdot w_j$$

Donde L_i es el límite inferior del intervalo que contiene al percentil, n es la cantidad total de observaciones, F_{j-1} es la frecuencia acumulada hasta el intervalo anterior, f_j es la frecuencia del intervalo, y w_j es la amplitud del intervalo.

Ejemplo 17

Los datos agrupados de las estaturas de los alumnos, con sus respectivas frecuencias simples y acumuladas, se observan en la tabla.

	Clase		Marca	Frecuencia	Frecuencia
j	LI	LS	y(j)	n(j)	Acumulada
1	1,55	1,60	1,575	2	2
2	1,60	1,65	1,625	10	12
3	1,65	1,70	1,675	12	24
4	**1,70**	**1,75**	**1,725**	**13**	**37**
5	1,75	1,80	1,775	9	46
6	1,80	1,85	1,825	3	49
7	1,85	1,90	1,875	1	50

Calculo de la altura que supera al 60% de los alumnos, P_{60} :

El cuarto intervalo es el primero que acumula $\frac{n \cdot k}{100} = \frac{50 \cdot 60}{100} = 30$ o más.

El percentil 60 es:

$$P_{60} = L_i + \frac{\frac{n \cdot k}{100} - F_{j-1}}{f_j} \cdot w_j = 1.70 + \frac{\frac{50 \cdot 60}{100} - 24}{13} 0.05 = 1.702$$

Los **cuartiles**, dividen la distribución de frecuencias en cuatro partes iguales. El primer cuartil, tiene una frecuencia acumulada de $\frac{n}{4}$, 25%, el segundo $\frac{2 \cdot n}{4}$, 50%, y el tercero $\frac{3 \cdot n}{4}$, 75%.

Los **deciles** dividen la distribución en diez partes iguales, el primero acumula $\frac{n}{10}$ (10%) observaciones, el segundo $\frac{2 \cdot n}{10}$ (20%), y así sucesivamente, hasta que el noveno y último decil acumula $\frac{9 \cdot n}{10}$ (90%).

De las definiciones anteriores, podemos ver que los cuartiles son los percentiles 25, 50 y 75, mientas que los deciles son los percentiles 10, 20,…, 90. De lo que se deduce que los últimos son un caso particular de los primeros.

Finalmente, observamos que la Mediana es el percentil 50, el quinto decil y el segundo cuartil.

En el siguiente apartado, presentamos las medidas numéricas de dispersión, las cuales brindan una idea respecto de la concentración de las observaciones en torno a sus medidas de tendencia central.

3.2.5 Medidas de Dispersión

Si bien las medidas de posición brindan información valiosa en relación a la ubicación de los datos, es importante también conocer qué tan dispersas están las observaciones.

Consideremos por ejemplo la observación de la estatura dos grupos de tres personas cada uno. Las observaciones del primer grupo son 1,70m., 1,71m. y 1,72m., mientras que las del segundo son 1,55m., 1,71m. y 1,87m. Los dos grupos tienen la misma media y mediana, la cual es de 1,71m. Sin embargo, claramente las observaciones son muy distintas, poseyendo el segundo grupo mucha más dispersión que el primero.

3.2.6 Rango

La medida de dispersión más sencilla es el rango o recorrido, y nos indica la longitud del intervalo en el cual están situadas todas las observaciones del conjunto de datos. Cuanto mayor es el rango, más dispersos estarán los datos.

El **Rango (o Recorrido)** de un conjunto de datos es la diferencia entre el máximo y el mínimo valor observado:

$$\text{Rango} = x_{\max} - x_{\min} .$$

Ejemplo 17

Utilizando los dos grupos ejemplificados al inicio de esta sección, podemos ver que el rango del primer grupo es $1,72 - 1,70 = 0,02$ (2cm.), mientras que el rango del segundo grupo es $1,87 - 1,55 = 0,32$ (32cm.).

3.2.7 Rango Intercuartílico

Esta medida es muy sencilla, y es muy similar a la anterior. La diferencia entre ambas es que el **Rango Intercuartílico** calcula la longitud del intervalo en el cual se encuentra el **50% de las observaciones centrales**, mientras el **Rango** calcula la longitud del intervalo en el cual se encuentran **todas las observaciones**.

El **Rango Intercuartílico** es la diferencia entre el tercer y el primer cuartil, o lo que es lo mismo, entre el percentil 75 y el percentil 25:

$$\text{RIC} = Q_{75} - Q_{25} .$$

Ejemplo 18

Si consideramos las estaturas del Ejemplo 16 puede calcularse el Rango Intercuartílico. Para hacerlo, deben de conocerse los percentiles 75 y 25. El percentil 75, conforme a lo desarrollado en el ejemplo al cual nos referimos, es 1,746.

El percentil 25 corresponde a la observación $\frac{25 \cdot n}{100} = \frac{25 \cdot 50}{100} = 12,5$. En consecuencia, se requiere de calcular el promedio entre las observaciones 12 y 13:

$$Q_{25} = \frac{1,642 + 1,652}{2} = 1,647 .$$

El Rango Intercuartílico de las alturas observadas es entonces:

$$\text{RIC} = 1,746 - 1,647 = 0,099 \text{ (9,9 cm.).}$$

3.2.8 Varianza

La medida más ampliamente usada para representar la dispersión de los datos es la varianza, la cual es una medida de los desvíos que existen entre las observaciones y su media.

Para calcularla, se promedian las desviaciones cuadráticas de cada uno de los datos respecto de la media aritmética. Es decir, que cada desviación se eleva al cuadrado y luego, se promedian estos valores.

La **Varianza muestral**, S^2, de un conjunto de datos es:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{M}(x_i - \bar{x})^2 \cdot f_i$$

Ejemplo 19

Considere los dos grupos de estaturas mencionados al inicio de esta sección.

La varianza muestral del primer grupo es:

$$S_1^2 = \frac{1}{2}\left[(1,70-1,71)^2 + (1,71-1,71)^2 + (1,72-1,71)^2\right] = \frac{0,0002}{2} = 0,0001$$

La varianza muestral del segundo es:

$$S_2^2 = \frac{1}{2}\left[(1,55-1,71)^2 + (1,71-1,71)^2 + (1,87-1,71)^2\right] = \frac{0,0512}{2} = 0,0256$$

Como era de esperar, la varianza del segundo grupo es mayor que la del primero.

Podemos desarrollar las fórmulas correspondientes a la varianza, obteniendo las siguientes **formas de cálculo** más sencillas:

$$S^2 = \frac{n}{n-1}(ma_2 - \bar{x}^2)$$

En general, cualquier momento centrado puede expresarse en términos de momentos absolutos[35]. En el Apéndice de este capítulo se presenta la deducción de la fórmula anterior y se demuestra, además, cómo expresar cualquier momento centrado utilizando los momentos absolutos.

3.2.9 Desvío Estándar

Quizás se pregunte por qué elevar al cuadrado las desviaciones en lugar de promediarlas directamente. La respuesta es que, si se promedian directamente las desviaciones respecto de la media, aquéllas positivas se compensarán con las negativas y el resultado final será cero, no

[35] Recordemos que la Varianza es el momento centrado de orden 2 y la media el momento absoluto de orden 1.

aportando ninguna noción respecto de la dispersión de los datos. En el Apéndice se demuestra que para cualquier conjunto de datos el momento centrado de orden 1 es siempre igual cero.

Un pequeño inconveniente que presenta la varianza es que su valor está expresado en unidades al cuadrado. Por ejemplo, si la variable que se mide son metros, la varianza indicará metros cuadrados. Por ello, en lugar de utilizar la varianza como medida de dispersión se suele utilizar el desvío estándar.

El **Desvío Estándar muestral** de un conjunto de datos, S, es la raíz cuadrada de la Varianza muestral.

3.2.10 Medidas de Forma

Finalmente, las **Medidas de Forma** constituyen el último grupo de medidas que permite caracterizar un conjunto de datos. Particularmente, como su nombre lo indica, permiten caracterizar la forma que tiene la distribución de frecuencias.

3.2.11 Coeficiente de Asimetría

Cuando en un conjunto de observaciones encontramos las mismas desviaciones por encima que por debajo de la media, decimos que la distribución de frecuencias es **simétrica**.

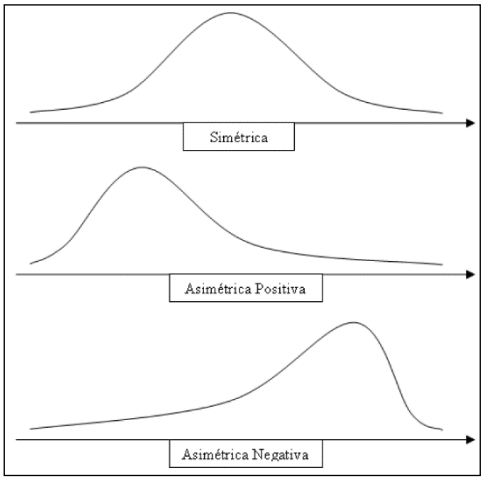
Si las desviaciones positivas (observaciones mayores a la media) son más importantes que las desviaciones negativas, decimos que la distribución es **asimétrica positiva (o con sesgo a la derecha)**, mientras que si ocurre lo contrario diremos que la distribución posee **asimetría negativa (o con sesgo a la izquierda)**. En la figura, vemos la caracterización gráfica de los tres casos mencionados.

Una manera sencilla de determinar si una distribución es simétrica o no, es comparando la media y la mediana y, posiblemente, también la moda. Las relaciones son:

$X_{mo} < X_{me} < \overline{x} \quad \Rightarrow$ Asimétrica Positiva (Sesgo a la Derecha).

$X_{mo} = X_{me} = \overline{x} \quad \Rightarrow$ Simétrica (Sesgo Nulo).

$X_{mo} > X_{me} > \overline{x} \quad \Rightarrow$ Asimétrica Negativa (Sesgo a la Izquierda).



El Coeficiente de Asimetría es una medida numérica que permite determinar qué tipo de asimetría posee una distribución.

El **Coeficiente de Asimetría** es:

$$As = \frac{\frac{1}{n}\sum_{i=1}^{M}(x_i - \bar{x})^3 \cdot f_i}{S^3} = \frac{mc_3}{\left(\sqrt{\frac{n}{n-1} \cdot mc_2}\right)^3}$$

La relación es:

$As = 0 \quad \Rightarrow$ Simétrica

$As > 0 \quad \Rightarrow$ Asimétrica Positiva

$As < 0 \quad \Rightarrow$ Asimétrica Negativa

En el numerador del coeficiente, el exponente de los desvíos es impar, por lo cual si hay desvíos más importantes hacia la derecha de la media (más cantidad o de mayor tamaño) que a la izquierda, el coeficiente resultará positivo. Lo contrario ocurrirá cuando los desvíos más importantes se encuentren a la izquierda de la media, resultando la distribución asimétrica negativa.

Ejemplo 20

Si tomamos los datos de las estaturas de los dos grupos de tres personas utilizados para ilustrar las medidas de dispersión, podremos ver qué los datos son simétricos, ya que el coeficiente de asimetría es nulo.

El momento centrado de orden dos, para cada uno de los grupos, ya se había calculado y es, respectivamente, 0,000067 y 0,17067. A continuación, mostramos el cálculo del momento centrado de orden 3 y el correspondiente coeficiente de asimetría:

Grupo 1:

$$\left.\begin{aligned} mc_3 &= 1/3 \times \left[(1,70-1,71)^3 + (1,71-1,71)^3 + (1,72-1,71)^3\right] \\ &= 1/3 \times \left[(-0,01)^3 + (0,01)^3\right] = 0 \end{aligned}\right\} \Rightarrow As_1 = 0$$

Grupo 2:

$$\left.\begin{aligned} mc_3 &= 1/3 \times \left[(1,55-1,71)^3 + (1,71-1,71)^3 + (1,87-1,71)^3\right] \\ &= 1/3 \times \left[(-0,16)^3 + (0,16)^3\right] = 0 \end{aligned}\right\} \Rightarrow As_2 = 0$$

Claramente, entonces, se observa que el coeficiente de asimetría es igual a cero dado que las desviaciones por encima y por debajo de la media son igual de importantes en ambos casos.

Veamos un ejemplo un poco más complejo de datos sesgados.

Ejemplo 21

Consideremos los datos de ingresos, cuyos valores reproducimos en la tabla.

Datos de Ingresos				
$ 476,36	$ 757,57	$ 1.064,75	$ 1.241,23	$ 1.582,23
$ 525,98	$ 779,97	$ 1.067,14	$ 1.273,37	$ 1.587,23
$ 566,48	$ 828,98	$ 1.082,39	$ 1.297,33	$ 1.626,68
$ 592,74	$ 835,74	$ 1.141,95	$ 1.351,68	$ 1.700,32
$ 619,73	$ 844,95	$ 1.187,00	$ 1.414,08	$ 1.730,47
$ 624,75	$ 874,51	$ 1.195,29	$ 1.434,53	$ 1.751,05
$ 666,10	$ 879,40	$ 1.195,81	$ 1.441,13	$ 2.041,14
$ 669,69	$ 895,00	$ 1.219,68	$ 1.455,31	$ 2.094,69
$ 689,61	$ 945,62	$ 1.219,90	$ 1.497,00	$ 2.207,91
$ 756,76	$ 1.013,76	$ 1.236,16	$ 1.562,25	$ 3.794,98

El promedio es $ 1210,77, mientras que la varianza es 320.667,85.

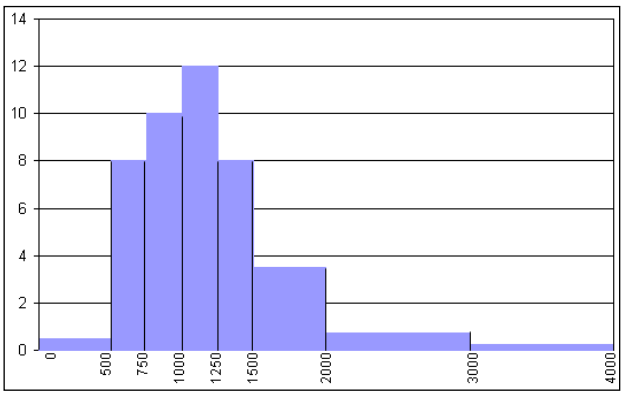
El momento centrado de orden tres es:

$$mc_3 = \frac{1}{50}\sum_{i=1}^{50}(x_i - 1210,77)^3 = \frac{1}{50}\left[(476,36-1210,77)^3 + (525,98-1210,77)^3 + ... + (3794,98-1210,77)^3\right] = 351.335.654,81$$

Finalmente, el coeficiente de asimetría es:

$$As = \frac{351.335.654,81}{\left(\sqrt{320.667,85}\right)^3} = 1,935$$

En el gráfico de este grupo de datos, realizado en el Ejemplo 7 y reproducido a continuación, se puede ver la asimetría de la distribución.



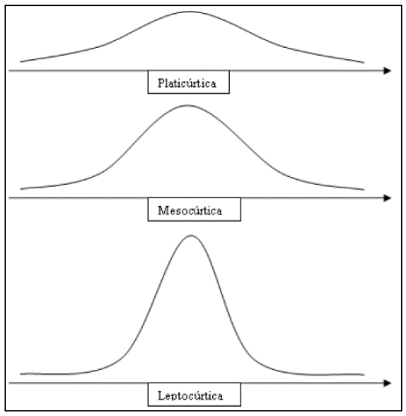
3.2.12 Curtosis

La última medida que estudiaremos aquí, y que está relacionada con la forma de la distribución, indica qué tan puntiaguda es la distribución.

Si la distribución es muy chata, diremos que la distribución es **platicúrtica**; si es muy puntiaguda, diremos que es **leptocúrtica**; finalmente, si no es ni muy chata ni muy puntiaguda, diremos que es **mesocúrtica**.

Lógicamente, si es muy chata o puntiaguda debe determinarse en comparación a alguna distribución de referencia. Esta distribución es la llamada Normal o Gaussiana, estudiada en el Capítulo 2.

En la figura, se ilustran los tres tipos de distribuciones mencionados.



El **Coeficiente de Curtosis** es:

$$K = \frac{\frac{1}{N}\sum_{i=1}^{M}(x_i - \bar{x})^4 \cdot n_i}{S^4} = \frac{mc_4}{\left(\frac{n}{n-1}mc_2\right)^2}$$

La relación es:

$K = 3$ $\Rightarrow$ Mesocúrtica

$K > 3$ $\Rightarrow$ Leptocúrtica

$K < 3$ $\Rightarrow$ Platicúrtica

El número 3 que aparece en la definición anterior no es un error ni una casualidad. Lo que sucede es que nuestra distribución de referencia, la **Normal**, tiene un Coeficiente de Curtosis igual a 3. Por ello, las comparaciones se hacen contra este número.

Ejemplo 22

Consideremos los datos del ejemplo anterior. El momento centrado de orden cuatro es:

$$mc_4 = 1/50 \times \sum_{i=1}^{50}(x_i - 1210,77)^4 = 1/50 \times \left[(476,36-1210,77)^4 + (525,98-1210,77)^4 + ... + (3794,98-1210,77)^4\right]$$
$$= 970.538.478.593,57$$

Luego, ya que la varianza es 320.667,85 (ver Ejemplo 21) el coeficiente de curtosis es:

$$K = \frac{970.538.478.593,57}{(320.667,85)^2} = 9,4384$$

Siendo este valor mayor que 3, podemos afirmar que la distribución es leptocúrtica.

3.3 Apéndice: Demostraciones

3.3.1 Cálculo de la varianza

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{M}(x_i - \bar{x})^2 \cdot f_i$$

$$= \frac{n}{n-1}\sum_{i=1}^{M}(x_i - \bar{x})^2 \cdot \frac{f_i}{n}$$ Multiplico y divido por n

$$= \frac{n}{n-1}\sum_{i=1}^{M}(x_i - \bar{x})^2 \cdot f_{ri}$$ Por definición de frecuencia relativa

$$= \frac{n}{n-1}\sum_{i=1}^{M}({x_i}^2 - 2 \cdot x_i \cdot \bar{x} + \bar{x}^2) \cdot f_{ri}$$ Al desarrollar cuadrados

$$= \frac{n}{n-1}\left(\sum_{i=1}^{M}{x_i}^2 \cdot f_{ri} - 2 \cdot \bar{x} \cdot \sum_{i=1}^{M} x_i \cdot f_{ri} + \bar{x}^2 \sum_{i=1}^{M} f_{ri}\right)$$

$$= \frac{n}{n-1}(ma_2 - 2 \cdot \bar{x}^2 + \bar{x}^2)$$

$$S^2 = \frac{n}{n-1}(ma_2 - \bar{x}^2)$$

3.3.2 Momento centrado de primer orden:

Se recuerda, a continuación, la definición de momento centrado de orden r:

$$mc_r = \frac{1}{N}\sum_{i=1}^{M}\left(x_i - ma_1\right)^r \cdot n_i$$

Puede demostrarse, entonces, que el momento centrado de orden 1 es igual a cero, independientemente del grupo de datos sobre el cual se trabaje.

$$mc_1 = \frac{1}{N}\sum_{i=1}^{M}\left(x_i - ma_1\right) \cdot n_i$$

$$= \sum_{i=1}^{M}\left(x_i - ma_1\right) \cdot f_i$$

$$= \underbrace{\sum_{i=1}^{M} x_i \cdot f_i}_{ma_1} - ma_1 \cdot \sum_{i=1}^{M} f_i$$

$$= ma_1 - ma_1$$

$$\boxed{mc_1 = 0}$$

4 Distribuciones de muestreo y Estimación

Dario Bacchini
Lara Vazquez
Andrea Lepera

Una población, estadísticamente hablando, contiene a todos los elementos relacionados con un fenómeno que se pretende estudiar. Por ejemplo, si analizamos el nivel de ingreso de los ciudadanos económicamente activos de Argentina, entonces, todos y cada uno de los ingresos de cada ciudadano activo constituyen la población bajo estudio.

Generalmente, el estudio de toda la población resulta muy costoso, o poco práctico o hasta puede resultar imposible), y por ello se utilizan muestras. Una muestra constituye una selección relativamente reducida de elementos de la población. Es importante que la obtención de la muestra sea técnicamente buena para que la misma represente a la población, y así poder realizar estudios sobre la muestra y extrapolar los resultados a la población. Lo que se busca al trabajar con muestras es conseguir lo que sería una población en miniatura, más manipulable y con un costo de adquisición menor. Sin embargo, por más que se tomen todos los recaudos necesarios, siempre al trabajar con muestras se enfrenta el riesgo de que la misma no represente fielmente a la población.

Uno de los principales objetivos de la estadística aplicada es lograr realizar afirmaciones generales respecto de ciertas características de una población sobre la base de observaciones realizadas en una muestra de la misma. La muestra es el "camino" a través del cual se pueden obtener conclusiones relacionadas con la población. Este procedimiento de obtener conclusiones de una población en base a datos de una muestra es lo que hemos denominado "Inferencia Estadística" en el Capítulo 1. Por ejemplo, las encuestas políticas a un grupo de personas sirven para sacar conclusiones respecto de la intención de voto de toda la población, la medición del rating televisivo en ciertos hogares es usado para estimar la cantidad de personas de la población que mira cada programa, la selección de algunas lamparitas para medir su vida útil se usa para inferir la duración de toda la producción, etc.

En algunos casos, el análisis de toda la población bajo estudio resulta destructivo, como en el caso del control de calidad: si se utilizan todos los productos fabricados para testearlos, ¡no quedaría ninguno para vender! En otros casos, el análisis de la población resulta costoso, como en los estudios de mercado: el costo de encuestar a toda la población en relación a un producto puede eliminar las ganancias que genere la venta del mismo.

Hemos mencionado que siempre existe el riesgo de que la muestra no represente adecuadamente a la población y, por ello, es importante mencionar que para realizar Inferencia Estadística, además de sacar conclusiones en relación a la población, es necesario brindar alguna medida del riesgo de éstas. Esta idea se comprenderá con mayor precisión cuando se analicen en el capítulo siguiente los intervalos de confianza.

Por otra parte, al tomar *diferentes muestras de una misma población* se obtendrán diferentes características muestrales (media, desvío estándar, etc.). De este modo, **las medidas muestrales que se analizan para inferir características poblacionales constituyen variables aleatorias**, porque su valor es incierto y depende de la muestra extraída de la población (depende del "resultado del experimento"). Por ello, antes de pasar de lleno a la realización de inferencias, en este capítulo, se analizará la distribución de probabilidades de algunas características muestrales, lo cual permitirá no sólo realizar estimaciones, sino también indicar el nivel de riesgo que poseen las mismas.

4.1 Muestreo Aleatorio: Técnicas

Es importante que la muestra sea representativa de la población, y si bien nunca tendremos la certeza de que dicha condición se cumpla, existen algunas técnicas que permiten seleccionar "mejores" muestras.

Como primera medida, hay que evitar que exista sesgamiento en la recolección de datos. Por ejemplo, si deseamos estudiar la cantidad de simpatizantes de cada club de fútbol que hay en el país, no sería muy adecuado realizar encuestas en la puerta del estadio de Boca antes de un partido. Por ello, es necesario "extender" lo más posible el proceso de recolección de datos para cubrir mejor a la población.

Por otra parte, el tamaño de la muestra es importante, ya que cuantos más datos contenga la misma, mayor información relacionada con la población tendrá, y las inferencias serán más confiables. Sin embrago, el mayor tamaño implicará un mayor costo, y deberá buscarse un equilibrio entre estos factores.

Cuando se eligen los elementos que conformarán una muestra se pueden utilizar dos métodos bien diferenciados: se eligen arbitrariamente o se realiza una selección al azar. El primer caso se conoce como "muestreo de juicio" (o no aleatorio) y se utiliza cuando el observador tiene gran conocimiento del objeto de estudio o cuando no se desea dejar fuera de la muestra algunos elementos considerados importantes para el estudio. En el segundo caso, denominado "muestreo aleatorio" (o probabilístico), cada elemento de la población tiene una probabilidad determinada de ser incluido en la muestra.

Supongamos que se diseña un experimento y se lo realiza para medir una propiedad X, o bien, se selecciona un elemento determinado para observar en el mismo la propiedad X [36]. En el primer "ensayo" se obtiene la observación X_1, en el segundo (realizado idealmente en las mismas condiciones) se obtiene la observación X_2, y así sucesivamente hasta obtener n observaciones $\{X_1; X_2; ...; X_n\}$ de la propiedad X. Entonces $\{X_1; X_2; ...; X_n\}$ forman un conjunto de variables iid que constituye una *muestra aleatoria* de la propiedad X.

Nótese que al decir que los ensayos se realizan "en las mismas condiciones", estamos asumiendo:

- Que se trata de una **población infinita**, o puede tratarse como tal por su gran tamaño, o bien,
- Que se realiza el **muestreo con reposición en una población finita** (ver Capítulo 1).

Si el muestreo se realiza sin reposición en una población finita, entonces las variables aleatorias $\{X_1; X_2; ...; X_n\}$ no serán independientes (ver Canavos, 1997).

Es importante destacar que cada miembro X_i de la muestra aleatoria constituye una variable aleatoria *antes de que se extraiga la muestra*, cuya distribución es idéntica a la de la población. Una vez que se llevan a cabo las observaciones, obtendremos una *realización* $\{x_1; x_2; ...; x_n\}$ de la muestra aleatoria.

Ejemplo 1

Supongamos que se desea conocer la cantidad de aparatos de TV que tiene cada hogar en la ciudad de Buenos Aires, para lo cual se extraerá una muestra aleatoria de 50 hogares de la población. Cada hogar es un elemento de la población, y la propiedad a observar es la "cantidad de aparatos de TV".

Antes de extraer la muestra, cada observación es una variable aleatoria que podría tomar valores entre cero y, digamos, veinte[37]. Si realizamos las siguientes observaciones en 50 hogares, tenemos *una* realización de la muestra aleatoria:

5	4	2	3	1	2	2	5	2	5
1	1	2	2	3	2	1	3	0	1
2	3	3	4	0	1	3	3	4	4
2	2	3	6	1	3	2	1	3	5
3	5	2	2	3	1	3	2	2	2

Estos valores son el resultado de *una* realización de la muestra aleatoria. Si volvemos a realizar el proceso de selección, obtendremos otra muestra que seguramente será distinta a la anterior.

El proceso de obtención de una muestra aleatoria puede realizarse mediante el empleo de distintas técnicas, las cuales se analizan en mayor detalle a continuación.

Simple

Realizamos un "muestreo aleatorio simple" cuando todos los elementos de la población tienen la misma probabilidad de ser incluidos en la muestra. Es decir, que si se toma una muestra de una población con N elementos en total, cada uno tendrá una probabilidad de $1/N$ de ser incluido

[36] El primer caso corresponde, en general, a fenómenos físicos: se construyen experimentos y se los repite bajo las mismas condiciones. El segundo está más relacionado con las ciencias económico-sociales: se seleccionan ciertos objetos tangibles de una población determinada.
[37] Suponiendo que ningún hogar tendrá más de 20 TVs.

en la muestra. El hecho de que sea “aleatorio” implica que cada elemento tiene una probabilidad de ser escogido, y que sea “simple” implica que la probabilidad es la misma para todos los elementos. Además, cada una de las posibles muestras de tamaño n tiene la misma posibilidad de ser seleccionada (ver “Sistemático”, más adelante).

Para realizar la selección pueden utilizarse tablas de números aleatorios y una codificación de los elementos de la población. Para mayores detalles, ver Berenson y Levine (1996).

En lo que sigue de esta obra supondremos que contamos con una muestra aleatoria simple, ya que este tipo de muestreo es la base de las técnicas de Inferencia Estadística que se desarrollarán en los capítulos posteriores. Sin embargo, por una cuestión de completitud, se expondrán, a continuación, otras técnicas que suelen utilizarse para obtener muestras, sin ahondar en sus detalles.

Sistemático

Realizamos un “muestreo aleatorio sistemático”, cuando todos los elementos de la población están ordenados y elegimos al azar el primer elemento, e incluimos en la muestra a los que aparecen en la lista cada k elementos. El tamaño de la muestra será de aproximadamente N/k, donde N es el tamaño de la población.

Ejemplo 2

Consideremos una empresa de telefonía celular que tiene 300.000 abonados y supongamos que tenemos un listado ordenado de los mismos. Deseamos realizar un sistemático con $k = 1.000$, y seleccionamos al azar el primer miembro de la muestra, que resulta el cliente 512. Entonces, el segundo miembro de la muestra será el cliente 1.512, el tercero el cliente número 2.512, y así, sucesivamente. En total, al llegar al cliente 299.512, habremos seleccionado 300.000/1.000 = 300 clientes.

La sistematización del proceso de muestreo puede referirse también a una cuestión temporal o espacial, y no sólo al orden. Por ejemplo, realizaríamos un muestreo sistemático temporal, si todos lo lunes observáramos la cantidad de vehículos que pasan por un peaje. Por otra parte, si medimos el nivel de precipitaciones en un día determinado cada 100 km., estaríamos realizando un muestreo sistemático espacial. Se debe tener mucho cuidado cuando se realizan muestreos de este tipo, ya que se podrían generar errores. Por ejemplo, es posible que los días lunes pasen más vehículos por los peajes de ingreso a la ciudad de Buenos Aires, y sería un error realizar la medición dicho día si se desea estimar el promedio diario.

Si bien cada elemento tiene igual probabilidad de ser incluido en la muestra, al igual que en el muestreo simple, en el caso del muestreo sistemático cada muestra de tamaño n no tiene la misma probabilidad de ser seleccionada. Por ejemplo, consideremos el caso de la encuesta a abonados: de todas las muestras de tamaño 300 que podrían elegirse haciendo un muestro simple, utilizando el muestreo sistemático no podría seleccionarse nunca una muestra que contenga los clientes 2 y 3 simultáneamente. De este modo, usando el muestreo sistemático, hay muestras que quedan excluidas.

Estratificado

Si en una población existen grupos muy homogéneos entre sí, pero bastante diferenciados entre ellos, resulta conveniente realizar un “muestreo estratificado”. Cada grupo relativamente homogéneo constituye un estrato, y para extraer la muestra se lleva a cabo un muestreo simple en cada uno de los estratos, de manera que se obtiene una sub-muestra por cada estrato.

El tamaño de cada sub-muestra puede ser proporcional al tamaño relativo del estrato respecto de la población, o bien, todas las sub-muestras pueden ser del mismo tamaño n/h, donde n es el tamaño de la muestra y h la cantidad de estratos. Por ejemplo, si se desea analizar el precio en dólares por metro cuadrado de las propiedades en la Ciudad de Buenos Aires, podrían considerarse como estratos los distintos barrios, pues los valores en cada uno serán similares. De este modo, en la muestra se incluirían algunas propiedades de cada barrio (la cantidad elegida podría ser, por ejemplo, proporcional a la cantidad de habitantes en cada barrio).

Por conglomerados

Si varios elementos de la población forman una unidad inseparable a los efectos del estudio, entonces, decimos que tenemos la población separada en conglomerados. Realizamos un "muestreo por conglomerados" cuando seleccionamos al azar algunos conglomerados (hacemos un "muestreo aleatorio" entre los conglomerados) e incluimos en la muestra a todos los elementos individuales que pertenecen a cada conglomerado elegido.

En general este muestreo se utiliza cuando los conglomerados son bastante similares entre sí, y hay mucha diversidad dentro de cada uno de ellos. Al existir similitud entre los conglomerados, se supone que cada uno es representativo de la población total. Por ejemplo si se desea estudiar la proporción de fumadores en la Ciudad de Buenos Aires, podrían considerarse como conglomerados los distintos barrios, presuponiéndose que en todos los barrios la proporción de fumadores es similar. Así se considerarían en la muestra algunos barrios al azar.

Si bien en los muestreos "estratificado" y "por conglomerado", se divide a la población total en grupos, en el primer caso hay mucha homogeneidad en cada grupo y diferencia entre los grupos, mientras que en el segundo hay gran homogeneidad entre los grupos, pero mucha diversidad dentro de cada uno.

Cabe señalar que tanto en el muestreo estratificado como en este, usamos los barrios de la Ciudad de Buenos Aires, pero con distintos criterios. En el primer caso, se considera que el valor del metro cuadrado de una propiedad es similar dentro de cada barrio pero diferente de un barrio a otro. En el muestreo por conglomerados, en cambio, suponemos que no hay diferencias entre la proporción de fumadores entre un barrio y otro, por lo que cualquier barrio elegido, ofrecería información similar.

4.2 Distribuciones de Estadísticos

En la sección anterior, hemos visto que cuando se extraen muestras aleatorias, las observaciones pueden diferir de una extracción a otra (de hecho, esto es lo que hace "aleatoria" a la muestra). Por ello, al ser aleatoria la muestra, también lo será cualquier medida que se calcule con la misma, como el promedio, la varianza o la mediana. A continuación, analizaremos esta cuestión.

4.2.1 Estadísticos y Parámetros

La utilidad de las muestras radica en su utilización para obtener conclusiones referidas a cierta población. Se denominan "estadísticos" a las características observadas en la muestra. Éstos permiten realizar inferencias referidas a las características correspondientes poblacionales denominadas "parámetros". De modo general, podemos decir que un "estadístico" es una característica de la "muestra", mientras que un "parámetro" es una característica de la "población". Los primeros permiten estimar los valores de los segundos.

El término "parámetro" está asociado a la utilización de los modelos de probabilidad estudiados en el Capítulo 2 para la realización de inferencias. Por ejemplo, si la población analizada se distribuye normalmente, necesitaremos inferir los valores de μ y σ, si la población es exponencial, necesitamos el parámetro de escala α, etc. En general, desde el enfoque clásico no bayesiano, los parámetros se consideran constantes fijas cuyo valor es desconocido.

Un **parámetro** es un valor que describe (parcial o totalmente) a una distribución de probabilidades de una propiedad poblacional de interés.

Para poder realizar cualquier afirmación probabilística, necesitaremos conocer los valores de los parámetros, o al menos obtener estimaciones de los mismos a partir de una muestra aleatoria. Para ello se utilizan los "estadísticos".

Un **estadístico** es una función de los elementos de una muestra aleatoria que no posee valores desconocidos.

Puede notarse que la definición anterior es bastante general, y no se limita solamente a las medidas descriptivas expuestas en el Capítulo 2 (media, mediana, varianza, etc.).

Ejemplo 3

Supongamos que se extraerá una muestra aleatoria de 4 elementos de una población, $\{X_1; X_2; X_3; X_4\}$ Los siguientes son estadísticos:

$$E_1 = \frac{X_1 + X_4}{2}$$

$$E_2 = X_1.X_2.X_3.X_4$$

$$E_3 = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

Si deseamos estimar la media poblacional ¿Cuál utilizaríamos? Bastante intuitivamente, creremos que todo el mundo respondería E_3, pero ¿por qué?

El ejemplo ilustra que, cuando se desea realizar inferencias relacionadas con los parámetros poblacionales, se debe seleccionar un estadístico adecuado para dicha tarea. La selección estará basada en ciertas propiedades que poseen algunos estadísticos, las cuales serán analizadas en el capítulo siguiente, donde se estudia más a fondo la estimación de parámetros.

4.2.2 Distribución de Muestreo

Las características que se calculan con una muestra (media, varianza, etc.), o más generalmente los "estadísticos", son variables aleatorias, porque son funciones de los elementos de una muestra aleatoria de la población. De este modo, para cada realización de la muestra aleatoria, se obtendrá un valor distinto del "estadístico", una *realización* del mismo.

Ejemplo 4

Consideremos los estadísticos E_1, E_2 y E_3 el Ejemplo 3. Supongamos que se extrajo la siguiente muestra, que constituye *una* realización de la muestra aleatoria:

$$\{x_1; x_2; x_3; x_4\}_1 = \{1; 2; 2; 5\}$$

Entonces, la realización correspondiente de cada estadístico es[38]:

$$e_{1;1} = \frac{1+5}{2} = 3; \quad e_{2;1} = 1 \times 2 \times 2 \times 5 = 20; \quad e_{3;1} = \frac{1+2+2+5}{4} = 2,5$$

Si repetimos el proceso de muestreo y obtenemos otra realización:

$$\{x_1; x_2; x_3; x_4\}_2 = \{2; 3; 2; 4\}$$

Entonces, tendremos las correspondientes realizaciones de los estadísticos:

$$e_{1;2} = \frac{2+4}{2} = 3; \quad e_{2;2} = 2 \times 3 \times 2 \times 4 = 48; \quad e_{3;2} = \frac{2+3+2+4}{4} = 2,75$$

De manera general, el valor particular que tome un estadístico, depende de los valores $\{x_1; x_2; ...; x_n\}$ que tome la muestra aleatoria $\{X_1; X_2; ...; X_n\}$ [39]. En el ejemplo previo, el tamaño de cada muestra era $n = 4$, y las variables $\{X_1; X_2; X_3; X_4\}$ se referían a los valores a observar, *desconocidos antes de extraer la muestra*. Una vez que se extrae una muestra particular, estas variables toman valores específicos, y permiten calcular un valor particular del estadístico, una *realización* del mismo. En el ejemplo, 2,5 y 2,75 son dos *realizaciones* del estadístico E_3.

[38] El segundo subíndice lo utilizamos para indicar el número de muestra.

[39] Como siempre, utilizamos letras mayúsculas para denotar variables aleatorias y letras minúsculas para las *realizaciones* de las mismas.

Como se obtienen distintos valores de acuerdo a la muestra que se seleccione, podríamos pensar en una distribución de probabilidades para los valores que puede tomar el estadístico. Es más, si pudiésemos tomar todas las muestras posibles de un tamaño n determinado, y con cada una de ellas calcular el valor del estadístico E, observando las frecuencias de cada valor, obtendríamos lo que se denomina distribución de muestreo del estadístico.

La **distribución de muestreo** de un estadístico es la distribución de probabilidad del mismo que se obtendría si se toman infinitas muestras aleatorias independientes (o todas las muestras posibles) de tamaño n.

Nótese la analogía con la noción de una distribución de probabilidad.

- Si tomamos *todos los elementos* de la población (los habitantes de Buenos Aires, por ejemplo) y medimos cierta característica (la estatura), observando las frecuencias de cada valor obtendremos la distribución de probabilidades de la característica observada.
- Si tomamos *todas las muestras* posibles de tamaño n de la población y calculamos un estadístico determinado, observando las frecuencias de cada valor *del estadístico* obtendremos la distribución de muestreo del mismo.

Todas las distribuciones de probabilidad pueden describirse, al menos parcialmente, por sus momentos. En particular, la media y la varianza de un estadístico nos proporcionarán una importante información respecto de su distribución de muestreo. En este punto, es importante remarcar que, en general, la distribución de muestreo de un estadístico no coincide con la distribución de la población, por lo que posiblemente tampoco lo hagan su media y varianza. Más adelante volveremos sobre esta cuestión.

Ejemplo 5

Supongamos que en una fábrica de lamparitas el departamento de control de calidad desea estudiar la vida útil de cada producto. Para ello, extraemos una muestra de 4 lamparitas diarias durante dos semanas y calculamos el promedio. Los resultados obtenidos se ilustran en la siguiente tabla. Podemos ver que el promedio depende de la muestra extraída y, en este caso, tiene un rango que va de 616,11 a 778,16 horas.

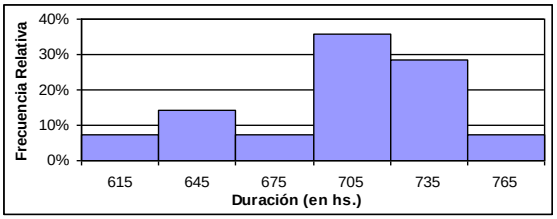
Con este procedimiento, tenemos 14 realizaciones de la muestra aleatoria y, consecuentemente, 14 realizaciones del estadístico "media muestral". Con estas realizaciones, podemos utilizar las técnicas del Capítulo 2 para construir la distribución de frecuencias y el correspondiente histograma de las medias muestrales, como se observa en la siguiente figura. Si pudiésemos seguir extrayendo muestras de manera indefinida y conseguir infinitas realizaciones de la muestra aleatoria (cosa que no es posible en términos prácticos), entonces, la distribución que obtendríamos sería la distribución de muestreo del promedio muestral.

Muestra #	1	2	3	4	5	6	7
	651	773	867	617	624	654	741
	532	840	691	710	679	623	683
	573	741	685	680	679	825	758
	709	759	727	784	557	650	794
Promedio	**616,11**	**778,16**	**742,59**	**697,77**	**634,72**	**688,19**	**743,92**

Muestra #	8	9	10	11	12	13	14
	795	780	710	680	753	647	624
	765	643	703	753	744	785	700
	728	743	740	749	552	765	692
	691	673	690	683	793	713	612
Promedio	**744,73**	**709,83**	**710,74**	**716,05**	**710,69**	**727,72**	**656,90**

Datos de 14 muestras con $n = 4$

Clase LI	Clase LS	Marca x(i)	Frecuencia n(i)	Frec. Relat. f(i)
600,00	630,00	615	1	7,14%
630,00	660,00	645	2	14,29%
660,00	690,00	675	1	7,14%
690,00	720,00	705	5	35,71%
720,00	750,00	735	4	28,57%
750,00	780,00	765	1	7,14%



Distribución de frecuencias e histograma de las medias muestrales

Al utilizar el proceso de muestreo para calcular estadísticos y, con ellos, estimar parámetros poblacionales, siempre se acarrea un riesgo y, por ello, las inferencias estadísticas están sujetas a la posibilidad de error. El desvío estándar de una distribución de muestreo es una medida de la dispersión del estadístico y, por tanto, del error que se pueda cometer en las conclusiones que se obtienen. Por ello, el desvío estándar, en este contexto, suele denominarse **error estándar** del estadístico.

En lo que resta del capítulo analizaremos las distribuciones de algunos estadísticos particulares, a saber: media muestral, varianza muestral y proporción muestral.

4.3 Distribución de $\bar{X}$ (media muestral): varianza poblacional conocida

Si bien ya hemos trabajado en varias ocasiones con la media muestral, y sabemos que es un estadístico que se utiliza para realizar inferencias en relación a la media poblacional, definámosla formalmente.

Sea $\{X_1; X_2; ...; X_n\}$, una muestra aleatoria. Entonces, el siguiente estadístico se denomina **media muestral**:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$= (X_1 + X_2 + ... + X_n)/n$$

Dado que el valor que tome el estadístico depende de la realización que se obtenga de una muestra aleatoria, $\bar{X}$ constituye una variable aleatoria, ya que su valor es incierto antes de la extracción de la muestra.

Ejemplo 6

Supongamos que se desea estimar la estatura (en metros) de los alumnos de la Facultad de Ciencias Económicas de la Universidad de Buenos Aires. Para ello, se tomará una muestra aleatoria de 60 alumnos (30 varones y 30 mujeres): $\{V_1; V_2; ...; V_{30}\}$ y $\{M_1; M_2; ...; M_{30}\}$. Consideremos las siguientes observaciones:

Varones				
1,68	1,72	1,70	1,71	1,57
1,82	1,86	1,87	1,77	1,65
1,70	1,79	1,72	1,98	1,77
1,93	1,69	1,73	1,80	1,72
1,66	1,73	1,59	1,94	1,74
1,79	1,73	1,65	1,79	1,76

Mujeres				
1,60	1,85	1,76	1,82	1,59
1,68	1,63	1,54	1,62	1,64
1,65	1,55	1,65	1,69	1,70
1,78	1,79	1,63	1,72	1,69
1,58	1,62	1,59	1,56	1,61
1,89	1,63	1,72	1,63	1,62

Podemos ver que la realización de la media muestral, correspondiente a esta realización de las muestras aleatorias es:

$$\overline{v}_1 = 1,75 \qquad \overline{m}_1 = 1,67$$

Supongamos ahora que tomamos otra muestra de 60 alumnos (obtenemos otra realización de la muestra aleatoria), siendo las observaciones, con sus respectivas medias, las siguientes:

Varones				
1,75	1,75	1,86	1,79	1,85
1,79	1,82	1,90	1,77	1,81
1,69	1,72	1,76	1,84	1,89
1,71	1,89	1,93	1,78	1,64
1,51	1,81	1,64	1,81	1,90
1,81	1,91	1,78	1,77	1,69

Mujeres				
1,64	1,71	1,66	1,76	1,75
1,62	1,63	1,63	1,62	1,87
1,66	1,65	1,64	1,61	1,83
1,83	1,63	1,81	1,70	1,63
1,59	1,53	1,81	1,77	1,83
1,66	1,68	1,78	1,58	1,76

$$\overline{v}_2 = 1,79 \qquad \overline{m}_2 = 1,70$$

Si siguiéramos el proceso y seleccionáramos todas las muestras posibles de 30 varones y 30 mujeres, obtendríamos todas las posibles realizaciones de los estadísticos $\overline{V}$ y $\overline{M}$, respectivamente. Con todos estos valores, construiríamos la distribución de muestreo de los mismos.

Generalmente, se toma una única muestra (o pocas) con la cual se realizan inferencias. Por ello, es importante conocer cómo es la distribución del estadístico, sin tener que recurrir a la extracción todas las muestras posibles (lo cual, de hecho, en general resulta imposible).

La esperanza matemática y la varianza son dos características que representan (al menos parcialmente) a toda distribución de probabilidades. Estas características, referidas a la media muestral, son analizadas a continuación. Luego se analizará la forma completa de la distribución de muestreo de $\overline{X}$ en ciertos casos especiales, y la manera de aproximarla cuando se carece de alguna información.

4.3.1 Esperanza y Varianza de la media muestral

Sea $\{X_1; X_2; ...; X_n\}$, una muestra aleatoria de variables iid con media μ. Entonces,

$$E(\overline{X}) = \mu$$

Lo anterior quiere decir que si tomamos *todas las muestras posibles de tamaño n* y calculamos la media muestral de cada una, el promedio de todas estas medias muestrales será la media poblacional. El hecho de que el valor esperado coincida con la media poblacional es una importante propiedad, que será analizada con mayor detalle en el capítulo siguiente.

Ejemplo 7

Supongamos que se desea analizar una población de 5 elementos, donde los valores de la característica de interés son 10, 12, 14, 16 y 18. Entonces $\mu = 14$. Obviamente, en este caso, no es necesario utilizar muestreo, pero lo haremos para ilustrar la proposición anterior.

Supongamos que se toman todas las posibles muestras de tamaño 2 y se calcula el promedio en cada una. Entonces, el promedio de todas estas medias muestrales será igual a la media poblacional, es decir 14. En la siguiente tabla, se observan los cálculos[40]:

[40] Por lo expuesto en la sección 1 de este capítulo, el muestreo es realizado con reposición, y de acuerdo a lo visto en el Capítulo 1 hay $n^r = 5^2$ muestras posibles.

Muestra #	X1	X2	Media muestral (i)
1	10	10	11
2	10	12	12
3	10	14	13
4	10	16	14
5	10	18	12
6	12	10	13
7	12	12	14
8	12	14	12
9	12	16	13
10	12	18	15
11	14	10	10
12	14	12	11
13	14	14	15

Muestra #	X1	X2	Media muestral (i)
14	14	16	14
15	14	18	16
16	16	10	13
17	16	12	14
18	16	14	15
19	16	16	16
20	16	18	17
21	18	10	14
22	18	12	15
23	18	14	16
24	18	16	17
25	18	18	18
Promedio medias muestrales =			**14**

Si se realizara lo mismo con todas las muestras de tamaño 3, se obtendría el mismo resultado.

Generalmente, las poblaciones son mucho más grandes, o incluso infinitas, y no es posible tomar todas las muestras de un tamaño determinado. Además, esto no sería muy práctico si se desean hacer inferencias, ya que tomar a toda la población sería más sencillo que tomar todas las muestras posibles de un tamaño determinado. Por ejemplo, en el caso anterior hay 5 elementos en la población, y 25 muestras posibles de tamaño 2 (tomadas con reposición). El Ejemplo 7 fue utilizado solamente con fines ilustrativos, ya que en la práctica nunca tomaremos todas las muestras posibles de cierto tamaño.

Sea $\{X_1; X_2; ...; X_n\}$, una muestra aleatoria de variables iid con desvío estándar σ. Entonces, el error estándar de $\overline{X}$ es:

$$Var(\overline{X}) = \frac{\sigma^2}{n} \qquad \Rightarrow \qquad \text{error estándar} = d.e.(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

La fórmula anterior implica que a medida que se aumenta el tamaño muestral, *n*, el error estándar del estadístico se reduce. Si el desvío es más pequeño, las observaciones estarán más concentradas en torno a la esperanza, por lo que es de esperar que las realizaciones de $\overline{X}$ estén bastantes próximas a la media poblacional, μ, que es precisamente la esperanza del estadístico analizado.

El hecho de que el desvío disminuya cuando crece el tamaño muestral es otra propiedad importante de un estadístico, que será analizada en el próximo capítulo.

Las demostraciones de las proposiciones anteriores son muy sencillas y se encuentran en el Apéndice del final del capítulo.

Como resulta obvio por la fórmula de cálculo, el desvío de la media muestral (error estándar de $\overline{X}$) resulta inferior al desvío poblacional. ¿Por qué ocurre esto? Cuando se selecciona un individuo de la población, éste puede resultar en una observación "extrema" (muy bajo o muy alto, por ejemplo), y si tomamos *individualmente* todos los elementos de la población, las observaciones pueden estar muy dispersas. En cambio, cuando tomamos una muestra de un tamaño determinado, 30 por ejemplo, y calculamos el promedio, es menos probable que este valor sea "extremo", ya que los valores altos se promedian con los bajos, obteniendo un valor más próximo al centro de la distribución. De este modo, si tomamos muchas muestras de tamaño 30 y calculamos el promedio de cada una de ellas, la dispersión de las medias muestrales resultará inferior a la dispersión de la población.

Además, cuanto más grande el tamaño muestral, menor fluctuación tendrá la media muestral, ya que las observaciones "muy altas" se estarán *promediando* con más cantidad de valores bajos o medios, haciendo menos notorio su efecto. Veamos algunos ejemplos para clarificar las ideas.

Ejemplo 8

Consideremos el ejemplo anterior. El desvío estándar poblacional puede calcularse y resulta igual a (ver Capítulo 2):

$$\sigma = \sqrt{\frac{1}{5}\sum_{i=1}^{5}(x_i - \mu)^2} \cong 2,8284.$$

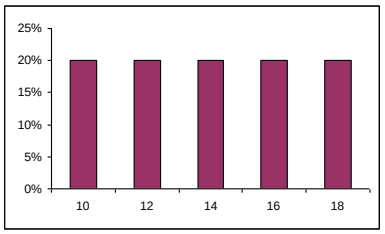
Ahora, calculemos el error estándar de la media muestral para un tamaño de muestra $n = 2$. Como hemos visto, la media de la "media muestral" coincide con la media poblacional, es decir $\mu = \mu_{\bar{X}} = 14$. Además, hemos visto que se tienen 25 posibles muestras de tamaño 2, por lo que:

$$\sigma_{\bar{X}} = \sqrt{\frac{1}{25}\sum_{i=1}^{25}\left(\bar{x}_i - \mu_{\bar{X}}\right)^2} = 2$$

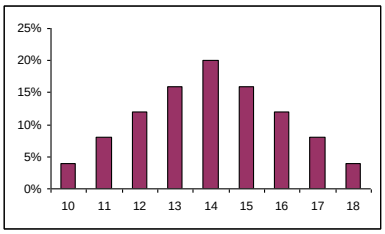
Podemos apreciar que el error estándar de la media muestral es menor que el desvío estándar poblacional y, además, que se cumple la relación establecida en la proposición anterior:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cong \frac{2,8284}{\sqrt{2}} \cong 2$$

En la siguiente figura, y utilizando los datos del ejemplo anterior, graficamos la distribución poblacional y la distribución de $\bar{X}$ para $n = 2$. Aunque hay bastante dispersión en la distribución de la media muestral, podemos apreciar que la misma es menor que en la población.



Distribución poblacional



Distribución de $\bar{X}$ con $n = 2$

Al igual que en el caso anterior, el ejemplo previo es solamente ilustrativo, ya que en la práctica nunca tomaremos todas las muestras posibles.

Ejemplo 9

Consideremos el ejemplo de las estaturas expuesto anteriormente. Supongamos, además, que se sabe que el desvío estándar poblacional es de 10 cm (o sea 0,10 m) para los varones y 8 cm (0,08 m) para las mujeres. Con esta información, podemos calcular el error estándar de la media muestral cuando se toman 30 observaciones:

$$d.e.\left(\bar{V}\right) = \frac{0,10}{\sqrt{30}} \cong 0,018 \qquad d.e.\left(\bar{M}\right) = \frac{0,08}{\sqrt{30}} \cong 0,015$$

Si en lugar de tomar 30 observaciones, se tomaran 20 o 10, entonces los errores estándares correspondientes serían:

$$d.e.\left(\bar{V}\right) = \frac{0,10}{\sqrt{20}} \cong 0,022 \qquad d.e.\left(\bar{M}\right) = \frac{0,08}{\sqrt{20}} \cong 0,018$$

$$d.e.\left(\bar{V}\right) = \frac{0,10}{\sqrt{10}} \cong 0,032 \qquad d.e.\left(\bar{M}\right) = \frac{0,08}{\sqrt{10}} \cong 0,025$$

Como ya hemos mencionado, podemos observar que al disminuir el tamaño muestral, se incrementa el error estándar de la "media muestral".

En este punto, es importante destacar que la disminución en el error estándar cuando se aumenta el tamaño de la muestra es decreciente. En el ejemplo previo, se puede observar que la disminución en el error estándar al aumentar una muestra de 10 a 20 es muy superior a la disminución que se obtiene cuando se aumenta de 20 a 30. Por esta razón, es importante evaluar la conveniencia de aumentar el tamaño de la muestra: *una muestra mayor reduce el error estándar, pero resulta más costosa*. En términos prácticos, si la muestra es mayor a 30, la reducción del error estándar al aumentar la muestra no es demasiado significativa.

Los resultados de esta sección son válidos **sin importar la forma que tenga la distribución poblacional**. En lo que sigue analizaremos la forma de la distribución completa de $\overline{X}$ (no sólo su media y su varianza) cuando la población es Normal, y cómo se puede aproximar la misma cuando la población no es Normal.

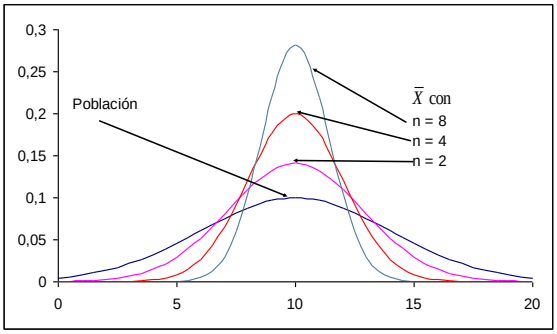
4.3.2 Poblaciones Normales

Cuando una muestra aleatoria es extraída de una población con distribución Normal, podemos conocer de manera exacta la distribución de la media muestral.

La distribución de muestreo de $\overline{X}$ cuando se muestrea una población Normal con media μ y desvío σ (ambos conocidos) es también Normal con media μ y desvío $\sigma/\sqrt{n}$. Es decir, $\overline{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}})$.

La demostración de la afirmación anterior excede el alcance del presente libro. El lector interesado puede consultar Novales Cinca (1997) o Canavos (1997).

Como ya se ha mencionado, el error estándar disminuye cuando aumenta el tamaño muestral, por lo que, en este caso, tendremos que la distribución de muestreo de $\overline{X}$ tiene la forma de campana tradicional de una variable Normal, pero cuando aumenta n la distribución estará más concentrada en torno a su valor medio. Esto puede apreciarse en la figura.



Ahora analicemos un ejemplo de aplicación.

Ejemplo 10

Supongamos que el nivel de ingreso medio de cierta localidad del país tiene una distribución Normal, con media 1.500 y desvío 500. Si se extrae una muestra aleatoria de tamaño $n = 20$, ¿cuál será la probabilidad que el *ingreso medio* de la misma esté entre 1.400 y 1.600?

Para responder esta pregunta, en primer lugar, debemos conocer la distribución de probabilidades del ingreso medio, la cual de acuerdo a lo visto es Normal con:

$$\mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}} = \sigma/\sqrt{n}$$
$$= 1.500 \qquad = 500/\sqrt{20} \cong 111,80$$

Luego, utilizando lo visto en el Capítulo 2 podemos calcular:

$$P(1400 \le \overline{X} \le 1600) = P(\overline{X} \le 1600) - P(\overline{X} \le 1400)$$
$$= P\left(\frac{\overline{X} - 1500}{111,80} \le \frac{1600 - 1500}{111,80}\right) - P\left(\frac{\overline{X} - 1500}{111,80} \le \frac{1400 - 1500}{111,80}\right)$$
$$= P(Z \le 0,8944) - P(Z \le -0,8944)$$
$$= 0,8145 - 0,1855$$
$$= 0,6290$$

Es decir, que al extraer una muestra de tamaño 20, hay un 62,90% de probabilidad de que el ingreso medio se encuentre entre 1400 y 1600. Podemos comparar este valor con la probabilidad análoga considerando una extracción individual. Es decir, responder a la pregunta ¿Cuál será la probabilidad

que el *ingreso de una persona* extraída al azar esté entre 1.400 y 1.600? Calculando este valor, tenemos que:

$$\begin{aligned} P(1400 \le X \le 1600) &= P(X \le 1600) - P(X \le 1400) \\ &= P\left(\frac{X - 1500}{500} \le \frac{1600 - 1500}{500}\right) - P\left(\frac{\overline{X} - 1500}{500} \le \frac{1400 - 1500}{500}\right) \\ &= P(Z \le 0,2) - P(Z \le -0,2) \\ &= 0,5793 - 0,4207 \\ &= 0,1586 \end{aligned}$$

Por lo tanto, si extraemos un elemento individual de la población, hay una probabilidad de 15,86% de que el ingreso esté entre 1400 y 1600. Esto ilustra claramente la mayor concentración que posee la distribución de la media muestral en relación a la distribución poblacional. Asimismo, cuando se aumenta el tamaño de la muestra, la concentración es aún mayor. Por ejemplo, si tomamos una muestra de tamaño $n = 40$, la probabilidad aumenta a 79,41% (¡compruébelo!).

En este ejemplo, se supuso que la población era Normal. Sin embargo, en la práctica, generalmente, se desconoce la forma de distribución poblacional, y ello elimina la posibilidad de realizar un análisis como el anterior. Afortunadamente, utilizando un teorema estadístico que veremos en la sección siguiente, por más que se desconozca la distribución de la población, en ciertos casos podremos obtener una aproximación de la distribución de muestreo de la media muestral.

4.3.3 Poblaciones No Normales: Teorema Central del Límite

En el capítulo 2 y en éste, hemos visto que la distribución Normal tiene gran importancia en muchas aplicaciones, y ello se debe a que la suma de variables aleatorias independientes con igual distribución (sin importar cuál sea ella), converge a una distribución Normal, cuando la cantidad de variables involucradas en la suma crece. Es decir, que si se suman muchas variables aleatorias con igual distribución, la distribución de la variable resultante será aproximadamente Normal.

Los comentarios intuitivos mencionados en el párrafo anterior se formalizan en el siguiente teorema, cuya demostración excede el alcance de esta obra. El lector interesado puede consultar Canavos (1997).

Teorema Central del Límite (v. 1): Sean $\{X_i; i = 1,2,...,n\}$, variables aleatorias independientes e idénticamente distribuidas (iid), siendo la media y varianza de cada una μ y σ^2, respectivamente. Entonces, cuando la cantidad de variables crece indefinidamente ($n \to \infty$), la distribución de la variable:

$$Z_n = \frac{\sum_{i=1}^{n} X_i - \mu n}{\sigma\sqrt{n}}$$

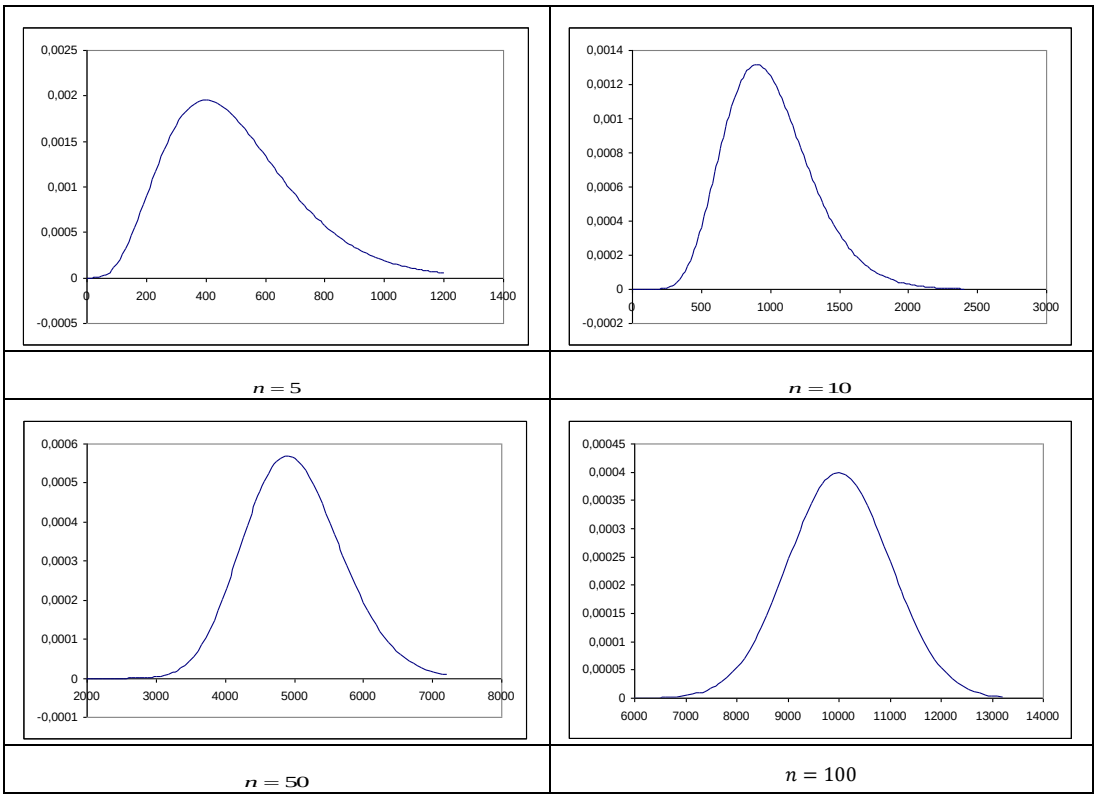
Converge a una distribución Normal Estándar. Es decir:

$$\lim_{n\to\infty} Z_n = Z \sim N(0,1)$$

Recordamos del Capítulo 2 que si se suman n variables aleatorias iid con media μ y varianza σ^2, la esperanza de la suma será $n\mu$ y la varianza será $n^2\sigma^2$. Por lo tanto, la variable Z_n que aparece en el teorema, es una variable estandarizada, y por ello su media y varianza son 0 y 1, respectivamente. Por lo tanto, en el teorema sería equivalente decir:

$$Y_n = \sum_{i=1}^{n} X_i \quad \Rightarrow \quad \lim_{n\to\infty} \frac{Y_n - n\mu}{n\sigma} = Y \sim N(0,1)$$

A fines de ilustrar este teorema, veremos lo que sucede al considerar la suma de variables con distribución exponencial con parámetro $\alpha = 0,01$. A medida que incrementemos el número n de variables que sumemos, vemos la convergencia a la distribución normal:



Primero veamos un ejemplo del Teorema, y luego veremos su aplicación en las distribuciones de muestreo.

Ejemplo 11

En una zona productiva dada, existen numerosas hilanderías. El promedio diario de hilados que produce cada una de ellas sigue una distribución Exponencial con media igual a 100m. Se desea saber cuál es la distribución de la media diaria total si existen 50 empresas en la zona.

Recordamos del Capítulo 2 que la media de la distribución exponencial es igual a $1/\alpha$, con lo cual:

$$1/\alpha = 100 \qquad \Rightarrow \qquad \alpha = 0,01$$

El desvío estándar en la distribución exponencial coincide con la media, por lo tanto, es igual a 100m. Por el *Teorema Central del Límite*, la variable aleatoria "media diaria total de la zona" tendrá una distribución aproximadamente Normal cuya media será igual a $n \times \mu = 50 \times 100 = 5000$, y el desvío estándar será $\sigma\sqrt{n} = 100\sqrt{50} \cong 707,11$. Observar los gráficos anteriores: en el caso de n=500, el centro de la campana de Gauss está en 5000 (media de la variable suma) y los puntos de inflexión en 5000±707,11.

Es importante destacar que en este ejemplo consideramos que la media para cada empresa seguía una distribución exponencial. Sin embargo, mientras que el promedio de cada empresa tenga la misma distribución (aunque la desconozcamos) y sean independientes es de aplicación el enunciado de este teorema.

Volviendo a las distribuciones de muestreo, el TCL es de aplicación cuando se realizan muestreos sobre poblaciones cuya distribución se desconoce, y no se puede afirmar que sea

Normal. Primero, enunciaremos el teorema de otra manera[41], y después, haremos algunos comentarios al respecto, y ejemplificaremos.

Teorema Central del Límite (v. 2): Cuando se muestrea una población cualquiera con media μ y desvío σ (o equivalentemente varianza σ^2), la variable:

$$Z_n = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}, \qquad \text{siendo } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

tiene una distribución que *tiende a una Normal estándar cuando* n *tiende a infinito*.

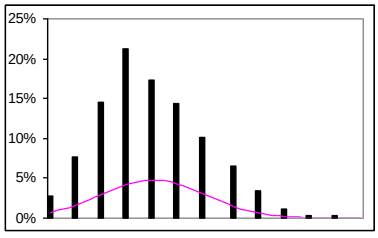
Es decir, que la distribución de $\lim_{n\to\infty} \frac{\bar{X}-\mu}{\sigma/n} = Z \sim N(0,1)$

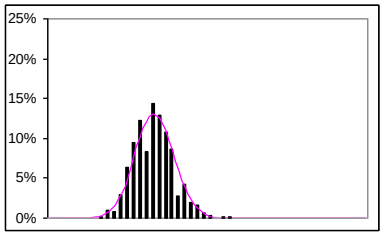
La variable Z_n es exactamente la misma que la expuesta en la versión anterior del teorema. Para obtener esta última expresión de Z_n, nada más dividimos el numerador y el denominador de la versión anterior por n.

De este modo, no importa la forma de la distribución poblacional, siempre que tomemos una muestra lo suficientemente grande, la distribución de muestreo de $\bar{X}$ podrá aproximarse por una distribución Normal. Así, aunque no conozcamos la distribución de la población, para un valor de n suficientemente grande, podremos al menos aproximar la distribución de muestreo de la media muestral (siempre que conozcamos la media y el desvío poblacional). En términos prácticos, para un tamaño muestral mayor a 30 la aproximación es bastante buena, sin importar lo "no Normal" que sea la distribución.

Ejemplo 12

En la siguiente figura, se ilustra la distribución de frecuencias simulada de la media muestral para una distribución de Poisson. Para construirla, se extrajeron 5000 muestras aleatorias de tamaño $n=4$ y 5000 de tamaño $n=30$ de una población con distribución de Poisson con $\lambda=1$. Con cada muestra obtenida se calculó la media muestral, obteniendo así 5000 realizaciones de media muestral con 4 observaciones, y 5000 realizaciones de media muestral con 30 observaciones. Luego, se graficó la distribución de frecuencias de las realizaciones, la cual aproximaría la distribución de muestreo "real". El teorema anterior estipula que cuanto mayor sea n, la distribución de muestreo se acercará más a una distribución Normal. Finalmente, para corroborar el teorema, se superpuso a cada gráfico la distribución Normal correspondiente a cada una, con media $\mu_{\bar{X}} = \lambda$ y desvío $\sigma_{\bar{X}} = \sqrt{\lambda}/\sqrt{n}$. Por lo tanto, la media es 1 en ambos caso, mientras que el error estándar depende del tamaño muestral: es $\sigma_{\bar{X}} = \sqrt{1}/\sqrt{4} = 0{,}5$ para $n=4$ y $\sigma_{\bar{X}} = \sqrt{1}/\sqrt{30} \cong 0{,}1826$ para $n=30$.





n = 4 | **n = 30**

Distribución de la media muestral. Extracción de 5000 muestras de una Poisson con $\lambda = 1$.

En la figura se observa que la distribución de muestreo de $\bar{X}$ se aproxima mucho a la Normal cuando $n=30$. Por otra parte, para $n=4$ se observa gran diferencia entre las dos distribuciones, la Normal y la de muestreo, además de observarse una marcada asimetría en esta última que no se percibe en el otro

[41] Las dos versiones que presentamos dicen exactamente lo mismo, pero con otras palabras.

caso. Asimismo, tal cual se vio en el apartado anterior, cuando aumenta el tamaño muestral, la dispersión se reduce.

Analicemos, ahora, una aplicación del Teorema a un caso práctico de control de producción.

Ejemplo 13

Una empresa que produce y envasa dulce de leche está interesada en medir el tamaño medio de cada uno de sus envases de 500 gr., ya que si el paquete dice que posee más dulce del que realmente contiene se estaría defraudando al consumidor, y ello podría ocasionar sanciones legales.

Para realizar el análisis se extraerá una muestra de 40 potes y se calculará la media muestral, conociendo que el desvío estándar poblacional es de 20 gr.

Ya que se desean evitar sanciones legales, se desea saber si la media poblacional es inferior a los 500 gr. Como se tomará una muestra para realizar el estudio, se corre el riesgo de que la media muestral sobreestime la media real. Para medir este posible error, se desea conocer la probabilidad de que la media observada en la muestra sea significativamente mayor a la real. Concretamente, la empresa desea saber: ¿Cuál es la probabilidad de que la media muestral sea 5 gr. mayor a la media real? O, en otras palabras, ¿cuál es la probabilidad de que la media muestral menos la media real supere los 5 gr.? Es decir, que se desea calcular: $P\left(\bar{X}-\mu>5gr\right)$.

Para realizar el cálculo, reordenaremos la expresión, y utilizaremos el teorema central del límite, ya que el tamaño muestral es mayor a 30. Los datos que conocemos son $n=40$ y $\sigma=20$, por lo tanto:

$$P\left(\bar{X}>\mu+5\right)=P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}>\frac{\mu+5-\mu}{\sigma/\sqrt{n}}\right)=P\left(Z>\underbrace{\frac{5}{20/\sqrt{40}}}_{1,5811}\right)$$

$$=1-P\left(Z\le 1,5811\right)=1-0,943$$

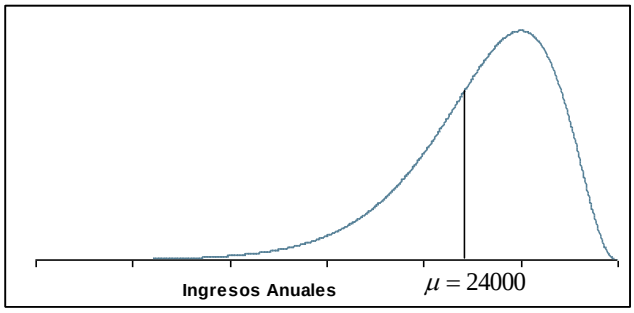
$$=0,057$$

Es decir, que la probabilidad de que la media observada sea 5gr. mayor a la media real es aproximadamente 5,7%.

Ejemplo 14

Supongamos que la distribución de ingresos de los habitantes de cierta localidad está sesgada de manera negativa, como se ilustra en la figura, siendo la media 24.000 y el desvío estándar 5.000. ¿Cuál es la probabilidad de que si se extraen 40 personas de manera aleatoria, el promedio de ingresos de éstas, esté entre 23.000 y 27.000?

Sabemos que, como el número de la muestra es mayor a 30, podemos aproximar la distribución del promedio muestral, $\bar{X}$, mediante una Normal con media $\mu_{\bar{X}}=24000$ y $\sigma_{\bar{X}}=\sigma/\sqrt{n}=5000/\sqrt{40}\cong 790,57$. Así, la probabilidad buscada es:



$$P\left(23000 \le \bar{X} \le 27000\right) = P\left(\frac{23000-24000}{5000/\sqrt{40}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{27000-24000}{5000/\sqrt{40}}\right)$$
$$\cong P\left(-1,2649 \le Z \le 3,7947\right)$$
$$= P\left(Z \le 3,7947\right) - P\left(Z \le -1,2649\right)$$
$$= 0,9999 - 0,1030$$
$$= 0,8970$$

Es decir, que hay casi un 90% de probabilidad de que la media muestral esté en el intervalo propuesto.

4.4 Distribución de $\bar{p}$ (proporción muestral)

En muchos estudios, es importante analizar la proporción de la población que cumple cierta característica: el porcentaje de habitantes que está a favor de cierto candidato político, la proporción de productos fallados en un día de operación, el porcentaje de alumnos que estudia Economía respecto del total de la facultad, etc. Para realizar inferencias estadísticas, en este caso, se utilizará la proporción muestral $\bar{p}$, que es simplemente la cantidad de casos favorables ("éxitos") que se obtuvo en la muestra, dividido por el tamaño de la muestra.

Consideremos una muestra aleatoria de n variables aleatorias iid, cada una con distribución de Bernoulli. Entonces la variable "cantidad de éxitos en la muestra", X, sigue una distribución Binomial, y el siguiente estadístico se denomina **proporción muestral**:

$$\bar{p} = X / n$$

Recordamos del Capítulo 2, que una distribución de Bernoulli es una variable aleatoria que puede tomar los valores 1 ó 0, con probabilidades p y $1-p$, respectivamente. A su vez, si se considera la suma de n variables aleatorias independientes de Bernoulli, se obtiene una variable aleatoria Binomial, que indica la cantidad de "éxitos" que se obtiene cuando se extrae una muestra aleatoria. Además, se vio en dicho capítulo que cuando n tiende a infinito (es lo *suficientemente* grande) la distribución Binomial puede aproximarse con una Normal. Esta aproximación es bastante buena cuando $np \ge 5$ y $n(1-p) \ge 5$ (ver Novales Cinca, 1997). Entonces, recordando que la esperanza de una variable Binomial es np y el desvío estándar $\sqrt{np(1-p)}$, la distribución de la siguiente variable es aproximadamente Normal Estándar si n es grande:

$$Y = \frac{X - np}{\sqrt{np(1-p)}}$$

La variable X ("cantidad de éxitos en una muestra") puede convertirse fácilmente en la "proporción de éxitos en una muestra", simplemente dividiéndola por el tamaño muestral n. Por lo tanto, si en la variable estandarizada anterior dividimos numerador y denominador por n podemos realizar la siguiente afirmación.

Consideremos el estadístico $\bar{p}$. La distribución de muestreo de la siguiente variable es aproximadamente Normal Estándar cuando el tamaño muestral n tiende a infinito:

$$Y = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

La afirmación anterior no es más que una aplicación del Teorema Central del Límite, ya que la variable X es la suma de las variables Bernoulli con componen la muestra, y al dividirla por el tamaño muestra se obtiene la "media muestral" de éstas. Es decir que $\bar{p}$ es simplemente una media muestral, en el caso particular en que cada una de las variables que compone la muestra es una variable de Bernoulli:

$$X = B_1 + B_2 + ... + B_n \qquad \Rightarrow \qquad \bar{p} = \frac{B_1 + B_2 + ... + B_n}{n}$$

Debido a esto, todo lo visto en el apartado anterior es válido aquí.

Ejemplo 15

Supongamos que en una empresa, la proporción de productos fallados en la producción diaria es de 0,10. ¿Cuál es la probabilidad de que si se extraen 100 artículos, la proporción muestral de artículos fallados supere sea menor al 6%?

En este caso, $n=100$ y $p=0,10$, por lo que $np=100\times 0,10=10\geq 5$ y $n(1-p)=100\times 0,90=90\geq 5$, por lo que podemos utilizar la distribución Normal. Entonces,

$$\begin{aligned} P(\bar{p}<0,06) &= P\left(\frac{\bar{p}-p}{\sqrt{p(p-1)/n}}<\frac{0,06-0,10}{\sqrt{0,10\times 0,9/100}}\right)=P\left(\frac{\bar{p}-p}{\sqrt{p(p-1)/n}}<\frac{-0,04}{0,03}\right) \\ &\cong P(Z<-1,3333) \\ &= 0,0912 \end{aligned}$$

Es decir, que hay menos del 10% de probabilidad de que la proporción muestral de artículos fallados sea menor al 6%.

Es importante hacer una observación más, antes de proseguir. Para calcular la varianza poblacional, es necesario conocer el parámetro p, y el mismo generalmente es desconocido (por eso, tomamos muestras y utilizamos el estadístico $\bar{p}$). En estos casos, existen dos caminos alternativos para solucionar este inconveniente, cuya utilización depende del caso práctico particular que se analice:

- Se remplaza $\bar{p}$ por p en el cálculo de la varianza.
- Se utiliza $p=0,5$, porque es el valor que maximiza la varianza de $\bar{p}$.
- Se emplea alguna aproximación previa basada en la opinión de un experto.

Ejemplo 16

Dos candidatos a presidente, de los partidos políticos A y B, llegaron a la segunda vuelta (ballotage) para definir quién gobernará la Nación los próximos 4 años. Para analizar las posibilidades que tiene de ganar, el partido A encarga a una consultora la realización de una encuesta a 500 votantes (es importante que la muestra sea aleatoria, ya que si, por ejemplo, se encuestan solamente a personas de la ciudad de Buenos Aires, la muestra estaría sesgada).

Los analistas, sabiendo que existe una probabilidad de error en los resultados que se obtengan en la muestra, desean informar al partido no sólo el resultado de la encuesta, sino también una medida del riesgo de que los resultados sean erróneos. Para ello, pueden calcular la probabilidad de que la proporción muestral observada supere en 5 puntos porcentuales a la proporción real. Es decir, $P(\bar{p}-p>0,05)$.

Utilizando lo visto en la sección anterior, y lo mencionado en ésta, sabemos que la distribución de muestreo de $\bar{p}$ puede aproximarse con una Normal. Además, como desconocemos el valor poblacional de p, carecemos de información para calcular la varianza de $\bar{p}$. En estos casos, para cubrirnos, utilizamos la máxima varianza posible[42], que ocurre cuando $p=0,5$. Por lo tanto, la probabilidad deseada puede calcularse como:

[42] Si la varianza es mayor, habrá mayor probabilidad de que una realización de $\bar{p}$ difiera p. Por lo que eligiendo la mayor varianza posible, obtendremos la mayor probabilidad de que $\bar{p}-p>0,05$.

$$P(\bar{p}-p>0,05)=P(\bar{p}>p+0,05)=P\left(\frac{\bar{p}-p}{\sqrt{p(1-p)/500}}>\frac{p+0,05-p}{\sqrt{p(1-p)/500}}\right)$$

$$\cong P\left(Z>\frac{0,05}{\sqrt{p(1-p)/500}}\right)=1-P\left(Z\le\frac{0,05}{\sqrt{0,5^2/500}}\right)$$

$$=1-P(Z\le 2,2361)$$

$$=1-0,9873$$

$$=0,0127$$

Puede observarse entonces que, como máximo, hay aproximadamente un 1,27% de probabilidad de que la proporción muestral observada en una muestra aleatoria de 500 votantes supere en un 5% a la proporción real. Por lo tanto, en principio podríamos decir si la proporción muestral de votantes favorables al partido A supera el 55%, el partido tendría una probabilidad de 98,73% (como mínimo) de ganar la elección. En este punto, es muy importante destacar que la muestra debe ser aleatoria, ya que en caso contrario se podrían llegar a conclusiones erradas.

4.5 Distribución de s^2 (varianza muestral) en poblaciones Normales

Sea $\{X_1; X_2; ...; X_n\}$, una muestra aleatoria. Entonces, el siguiente estadístico se denomina **varianza muestral**:

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

El **desvío estándar muestral** es la raíz cuadrada de la varianza muestral.

s^2 es una variable aleatoria, ya que es un estadístico cuya realización depende de la realización de una muestra aleatoria. Por lo tanto, cuando extraemos *una* muestra y calculamos la varianza muestral, lo que tenemos es *una realización* del estadístico s^2, y si extraemos distintas muestras, obtendremos distintas realizaciones de s^2.

La distribución de muestreo de s^2 será la distribución de probabilidades que se obtendría, si se seleccionaran todas las muestras aleatorias posibles; en cada una de ellas se calcula el estadístico y se observan las frecuencias relativas de cada realización. Esto no es posible en términos prácticos y, por ello, al igual que utilizamos la Normal y utilizaremos la t de Student (ver Sección 6) para la media muestral, recurriremos a algunas distribuciones teóricas para la distribución de muestreo de la varianza muestral, a saber: la distribución **Chi-cuadrado** y la distribución **F de Fischer- Snedecor**. Antes de analizar estas distribuciones teóricas, veamos un ejemplo para clarificar los conceptos.

Ejemplo 17

Consideremos el Ejemplo 6 de este capítulo, en el cual se extrajeron dos muestras de estaturas de 30 varones y 30 mujeres. La primera realización de la muestra era:

Varones				
1,68	1,72	1,70	1,71	1,57
1,82	1,86	1,87	1,77	1,65
1,70	1,79	1,72	1,98	1,77
1,93	1,69	1,73	1,80	1,72
1,66	1,73	1,59	1,94	1,74
1,79	1,73	1,65	1,79	1,76

Mujeres				
1,60	1,85	1,76	1,82	1,59
1,68	1,63	1,54	1,62	1,64
1,65	1,55	1,65	1,69	1,70
1,78	1,79	1,63	1,72	1,69
1,58	1,62	1,59	1,56	1,61
1,89	1,63	1,72	1,63	1,62

Por lo tanto, la realización de la varianza muestral, y el correspondiente desvío muestral, es:

$$s_1^2(\text{varones})=0,009355 \quad \Rightarrow \quad s_1(\text{varones})=0,096719$$

$$s_1^2(\text{mujeres})=0,008046 \quad \Rightarrow \quad s_1(\text{mujeres})=0,089702$$

La segunda muestra era:

Varones				
1,75	1,75	1,86	1,79	1,85
1,79	1,82	1,90	1,77	1,81
1,69	1,72	1,76	1,84	1,89
1,71	1,89	1,93	1,78	1,64
1,51	1,81	1,64	1,81	1,90
1,81	1,91	1,78	1,77	1,69

Mujeres				
1,64	1,71	1,66	1,76	1,75
1,62	1,63	1,63	1,62	1,87
1,66	1,65	1,64	1,61	1,83
1,83	1,63	1,81	1,70	1,63
1,59	1,53	1,81	1,77	1,83
1,66	1,68	1,78	1,58	1,76

Siendo las varianzas y desvíos resultantes las siguientes:

$$s_2^2(\text{varones}) = 0,008364 \quad \Rightarrow \quad s_2(\text{varones}) = 0,091457$$
$$s_2^2(\text{mujeres}) = 0,008029 \quad \Rightarrow \quad s_2(\text{mujeres}) = 0,089606$$

Si siguiéramos el proceso y seleccionáramos todas las muestras posibles de 30 varones y 30 mujeres, obtendríamos todas las posibles realizaciones de los estadísticos $s^2(\text{varones})$ y $s^2(\text{mujeres})$, respectivamente. Con todas estas realizaciones y sus frecuencias correspondientes podríamos construir la distribución de muestreo de los estadísticos mencionados. Como esto no es posible en la práctica, se recurren a modelos de probabilidad teóricos.

En los apartados siguientes, en primer lugar expondremos las funciones de densidad teóricas que utilizaremos, con sus correspondientes esperanza y varianza, y, luego, veremos su aplicación a las distribuciones de muestreo.

4.5.1 Distribución de varianza muestral: Chi-cuadrado

La distribución Chi-cuadrado tiene importantes aplicaciones para realizar **inferencias relacionadas con la varianza** de una población. La función de densidad de esta variable es un caso particular de la distribución gamma, y surge de remplazar en la misma el parámetro de escala por el número 0,5 ($\alpha = 1/2$) y el parámetro de forma por $\beta = \eta/2$ (ver Capítulo 2).

La variable X tiene **distribución Chi-cuadrado** con η grados de libertad, si su función de densidad es:

$$f(x|\eta) = \frac{(x/2)^{\eta/2}}{x\Gamma(\eta/2)}\exp(-x/2) \qquad x>0;\ \eta>0$$

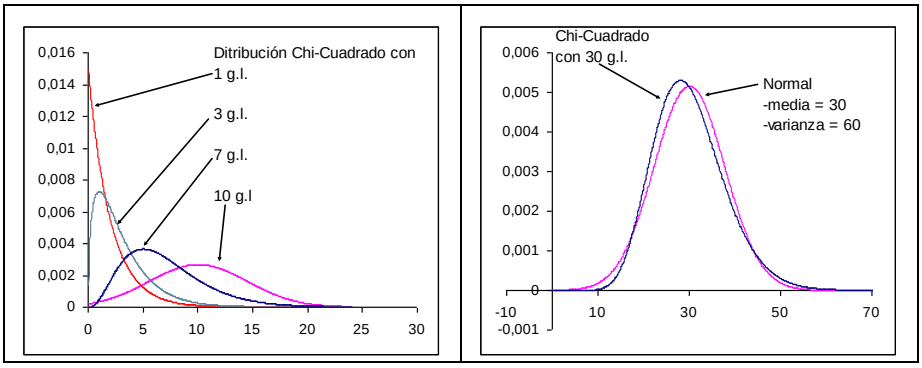
Entonces, escribimos $X \sim \chi^2_\eta$.

La **esperanza** y la **varianza** de una variable aleatoria con distribución Chi-cuadrado están dadas por:

$$E(X|\eta) = \eta \qquad Var(X|\eta) = 2\eta$$

Al igual que con las distribuciones vistas en el Capítulo 2, la función de distribución de la variable Chi-cuadrado se encuentra tabulada en muchos libros de texto, o bien puede calcularse utilizando Microsoft® Excel.

En el primer gráfico de la siguiente figura, se ilustra la función de densidad para distintos valores del parámetro η. Se observa que la distribución es asimétrica, pero cuando la cantidad de grados de libertad aumenta, esta característica disminuye. Es más, cuando η es lo suficientemente grande, la distribución Chi-cuadrado tiene a una distribución Normal. Esto puede apreciarse en la segunda parte de la figura.



Un aspecto importante de esta distribución es que la suma del cuadrado de n variables normales estandarizadas, se corresponde con la variable Chi-cuadrado con n grados de libertad. Es decir que si tomamos n variables aleatorias $N(0;1)$, las elevamos al cuadrado a cada una de ellas, y las sumamos, la distribución resultante será la Chi-cuadrado. Esto es:

Sean Z_i (i=1,…,n) variables aleatorias Normales iid con media 0 y desvío 1. La variable

$$Y = \sum_{i=1}^{n} Z_i^2$$

tiene distribución Chi-cuadrado con n grados de libertad.

Análogamente, si X_i ($i = 1, 2, ..., n$) constituye una muestra aleatoria de variables Normales iid con media μ y desvío igual a σ, la variable

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

tiene distribución Chi-cuadrado con n grados de libertad.

En base a la proposición anterior, surge la aplicación de la distribución Chi-cuadrado a la distribución de muestreo de la varianza muestral.

Sea $\{X_1; X_2; ...; X_n\}$, una muestra aleatoria de una población Normal con media μ y desvío σ (ambos conocidos), y sea:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$$

la varianza muestral. Entonces, la distribución de muestreo de $(n-1)s^2/\sigma^2$ es Chi-cuadrado con $n-1$ grados de libertad. Es decir,

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

La demostración de las proposiciones anteriores excede el alcance de este libro.

Ejemplo 18

En el mercado financiero, suele ponerse mucho énfasis en el análisis de la variabilidad de los precios y/o rendimientos. Supongamos que deseamos analizar la varianza del rendimiento mensual de un activo financiero, cuya distribución, en base a la experiencia histórica, es Normal con media 0 y desvío estándar 1%. Si obtenemos una muestra de rendimientos mensuales de los últimos 2 años (24 meses), ¿cuál será la probabilidad que el desvío mensual, calculado a partir de la muestra exceda a 1,25%?

Antes de calcular lo deseado, recordemos que los datos son referidos al desvío y la distribución Chi-cuadrado es para las varianzas, por lo que para utilizarla deberemos elevar al cuadrado cada uno de los datos con que contamos. Así, la varianza poblacional será $\sigma^2 = 0,01^2 = 0,0001$, la cantidad de grados de libertad será 23 (tenemos 24 datos muestrales) y el valor cuya probabilidad se desea calcular es $0,0125^2 = 0,00015625$. La probabilidad deseada, entonces, es poco más del 4%:

$$\begin{aligned} P\left(s^2 > 0,00015625\right) &= P\left(\frac{(n-1)s^2}{\sigma^2} > \frac{(24-1) \times 0,00015625}{0,0001} \right) \\ &= P\left(\chi^2_{23} > 35,9375\right) \\ &= 0,0419 \end{aligned}$$

4.5.2 Distribución F (Fischer – Snedecor)

La distribución F es sumamente importante cuando se desean **comparar varianzas** de distintas poblaciones. Su aplicación se realiza especialmente para realizar pruebas de hipótesis

estadísticas, las cuales se verán en el Capítulo 7. A continuación, se expone la función de densidad, su esperanza y su varianza, y luego se explicará su utilización como distribución de muestreo.

La variable x tiene **distribución F** con parámetros η_1 y η_2, si su función de densidad es:

$$f\left(x|\eta_1,\eta_2\right)=\frac{\Gamma\left[\left(\eta_1+\eta_2\right)/2\right]\eta_1^{\eta_1/2}\eta_2^{\eta_2/2}}{\Gamma\left(\eta_1/2\right)\Gamma\left(\eta_2/2\right)}x^{(\eta_1-2)/2}\left(\eta_2+x\eta_1\right)^{-(\eta_1+\eta_2)/2}$$

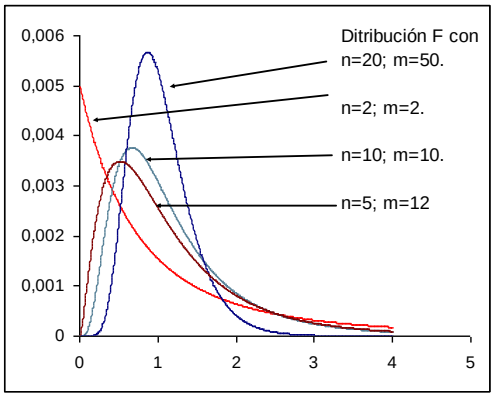
con $x>0$ y $\eta_1;\eta_2>0$.

La **esperanza** y la **varianza** de una variable aleatoria con distribución F:

$$E\left(X|\eta_1,\eta_2\right)=\frac{\eta_2}{\eta_2-2} \qquad \text{para } \eta_2>2$$

$$V\left(X|\eta_1,\eta_2\right)=\frac{2\cdot\eta_2^2\cdot\left(\eta_1+\eta_2-2\right)}{\eta_1\cdot\left(\eta_2-2\right)^2\cdot\left(\eta_2-4\right)} \qquad \text{para } \eta_2>4$$

Las probabilidades de la distribución F pueden calcularse con Microsoft® Excel o SPSS. En la siguiente figura, se observa la función de densidad de la distribución F para distintos valores de grados de libertad.



Tal como en el caso de la distribución Chi-cuadrado, las aplicaciones se derivan de la definición de la variable cuya distribución es F. En este caso, el **cociente** de dos variables Chi-cuadrado, cada una dividida por sus respectivos grados de libertad, tiene una distribución F.

Sean A y B dos variables independientes con distribución Chi-cuadrado, cuyos grados de libertad son, respectivamente, η_1 y η_2. Entonces, la variable:

$$F=\frac{A/\eta_1}{B/\eta_2}$$

tiene **distribución F** con parámetros η_1 y η_2, denominados grados de libertad de numerador y denominador respectivamente (notado $F\sim F_{\eta_1\eta_2}$)

De lo mencionado anteriormente, podemos realizar la siguiente afirmación, que resulta en la aplicación más importante de la distribución F:

Sea $\{X_1; X_2; ...; X_n\}$ una muestra aleatoria de tamaño n de una población Normal con desvío σ_X, y sea $\{Y_1; Y_2; ...; Y_m\}$ una muestra aleatoria de tamaño m de otra población Normal con desvío σ_Y. Entonces, por lo visto en el apartado anterior, $(n-1)s_X^2/\sigma_X^2$ es una Chi-cuadrado con $n-1$ g.l. y $(m-1)s_Y^2/\sigma_Y^2$ es una Chi-cuadrado con $m-1$ g.l.

Si dividimos cada una de estas variables por sus respectivos grados de libertad, obtenemos s_X^2/σ_X^2 y s_Y^2/σ_Y^2. Luego, el cociente de estas dos variables será F, con $n-1$ y $m-1$ grados de libertad:

$$\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} \square F(n-1, m-1)$$

Ejemplo 19

Supongamos que se extraen dos muestras de una misma población Normal, una de tamaño 20 y otra de tamaño 50. ¿Cuál es la probabilidad de que la varianza muestral de la primera muestra sea más del doble que la varianza muestral de la segunda?

Que la varianza muestral de la primera muestra, s_1^2, sea el doble o más que la segunda, s_2^2, es equivalente a decir que $s_1^2 > 2s_2^2$. Teniendo en cuenta que las dos muestras se extrajeron de la misma población, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Por lo tanto, la probabilidad deseada es:

$$\begin{aligned} P(s_1^2/s_2^2 > 2) &= P\left(\frac{s_1^2/\sigma^2}{s_2^2/\sigma^2} > 2\right) \\ &= P(F_{19;49} > 2) \\ &\cong 0,0265 \end{aligned}$$

Por lo tanto, hay aproximadamente un 2,65% de probabilidad de que la varianza muestral con 20 elementos sea más del doble que la varianza con 50 elementos.

Es importante realizar una última aclaración de esta distribución, cuya demostración es inmediata por la definición de la distribución F en términos de dos variables Chi-cuadrado.

Si $F = \frac{A/\eta_1}{B/\eta_2}$ es tiene distribución Fisher-Snedecor con η_1 y η_2 grados de libertad en numerador y denominador respectivamente, entonces $1/F = \frac{B/\eta_2}{A/\eta_1}$ tendrá distribución Fisher-Snedecor con η_2 y η_1 grados de libertad.

4.6 Distribución de $\bar{X}$: varianza poblacional desconocida

En el apartado 3 se analizó la distribución de $\bar{X}$, cuando se tomaban muestras de poblaciones con varianza conocida. Si la población se distribuye como una variable Normal, entonces $\bar{X}$ también será Normal, mientras que para cualquier otras distribución, la distribución de $\bar{X}$ puede aproximarse con una Normal si el tamaño muestral es suficientemente grande.

Cuando **se desconoce la varianza poblacional**, se utiliza otro modelo teórico de probabilidad para la distribución de muestreo de $\bar{X}$: la **distribución t de Student**.

A continuación, expondremos la función de densidad de una variable con distribución t de Student, las fórmulas de cálculo de la esperanza y la varianza, y un teorema que relaciona esta variable con la Normal y la Chi-cuadrado. Luego, aplicando este teorema, utilizaremos la t de Student como distribución de muestreo para $\bar{X}$.

4.6.1 Distribución t de Student

La variable x tiene **distribución t de Student** con η grados de libertad, si su función de densidad es:

$$f\left(x|\eta\right)=\frac{\Gamma\left[\left(\eta+1\right)/2\right]\left(x/2\right)^{\eta/2}}{\sqrt{\eta\pi}\,\Gamma\left(\eta/2\right)}\left[1+x^2/\eta\right]^{-(\eta+1)/2} \qquad -\infty<x<+\infty;\ \eta>0$$

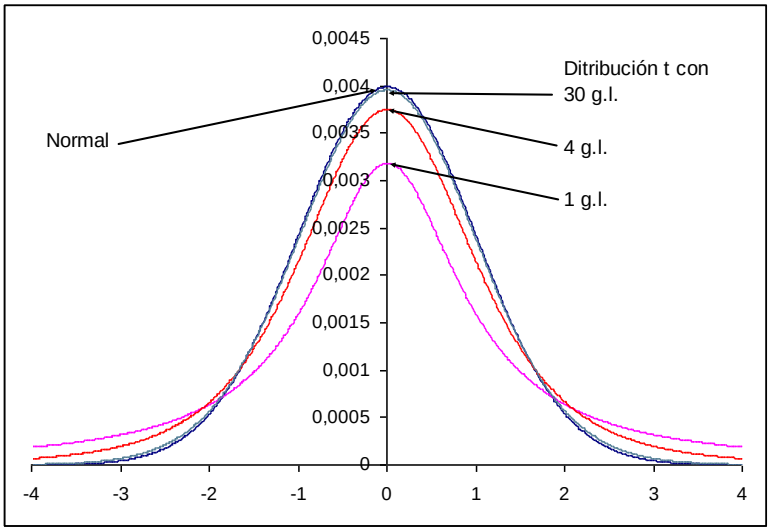
La **esperanza** y la **varianza** de una variable aleatoria con distribución t de Student son:

$$E\left(X|\eta\right)=0$$

$$V\left(X|\eta\right)=\eta/\left(\eta-2\right) \qquad \text{con } \eta>2$$

Al igual que con las distribuciones vistas en el Capítulo 2, la función de distribución de la variable t de Student se encuentra tabulada y sus valores pueden calcularse utilizando Microsoft® Excel o SPSS.

La distribución t de Student es simétrica y tiene forma de campana al igual que la distribución Normal. Cuando la cantidad de grados de libertad es pequeña la distribución *t* posee menos kurtosis y las "colas" más levantadas o más pesadas que la Normal, pero para valores de grados de libertad suficientemente grandes, la distribución t de Student puede aproximarse mediante una Normal estándar. En la siguiente figura, se ilustra una comparación de la distribución *t* para distintos grados de libertad, y la distribución $N(0;1)$. Se puede observar que para 30 grados de libertad, las dos curvas son prácticamente iguales. Por esta razón, la distribución T de Student con grados de libertad superiores a 30, suele aproximarse por la distribución normal estándar.



En este capítulo, hemos visto la aplicación de las distribuciones Normal y Chi-cuadrado a las distribuciones de muestreo. La principal característica de una variable t de Student, y lo que permite su utilización como distribución de muestreo, es que relaciona a las dos distribuciones anteriores. La siguiente propiedad establece la relación mencionada, y la función de densidad expuesta más arriba se desprende del mismo. La demostración excede el alcance de esta obra.

El cociente entre una variable Normal Estándar y la raíz cuadrada de una variable Chi-cuadrado dividida por sus grados de libertad η, es una variable t de Student con η grados de libertda, notada T_η, siempre y cuando las dos variables involucradas sean independientes.

Es decir, que si Z es una variable aleatoria Normal Estándar, Y es una variable Chi-cuadrado con η grados de libertad, y las variables Z e Y son independientes, entonces, la siguiente variable tiene una distribución T de Student con η grados de libertad:

$$X = \frac{Z}{\sqrt{Y/\eta}} \sim T_\eta$$

4.6.2 Poblaciones Normales con muestras pequeñas

En base lo anterior, y a las distribuciones de muestreo de $\bar{X}$ y de s^2 estudiadas en las secciones anteriores, podemos obtener la principal aplicación de la distribución t de Student.

Cuando se muestrea una población Normal con media μ y desvío desconocido, la distribución de:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

es *t de Student* con $n-1$ grados de libertad.

La demostración de esta proposición se basa en la sección anterior, y se encuentra en el Apéndice de este capítulo.

Ejemplo 20

La emisión de dióxido de carbono (CO2) es uno de los principales causantes del efecto invernadero a nivel mundial. Supongamos que el gobierno de un país está interesado en controlar la emisión de CO2 de los automóviles para proteger el medio ambiente. Para ello, ha intimado a las empresas a que la emisión debe ser de como máximo de 140 gramos por kilómetro al finalizar el corriente año.

Al finalizar el año, el gobierno empieza a realizar los controles correspondientes, para lo cual tomará muestras de 20 coches en cada fábrica. Supongamos que en la muestra correspondiente a la fábrica A se observó una media de 143 g/km. y un desvío estándar de 5 g/km. Si la emisión de CO2 de los coches de la fábrica A sigue una distribución Normal ¿Cuál es la probabilidad de que la empresa esté violando el requisito del gobierno? En otras palabras: si su media poblacional (la media de todos sus coches) fuese realmente 140 g/km. ¿cuál sería la probabilidad de haber observado el valor de 143 g/km. o más?

La probabilidad que se desea calcular es $P\left(\bar{X} \geq 143 \middle| \mu = 140\right)$, para lo que utilizaremos la distribución t de student con 19 grados de libertad, ya que la varianza poblacional es desconocida y la muestra es de tamaño 20. La probabilidad deseada es aproximadamente 0,7%, ya que:

$$\begin{aligned} P\left(\bar{X} \geq 143 \middle| \mu = 140\right) &= P\left(\frac{\bar{X} - \mu}{s/\sqrt{n}} \geq \frac{143-140}{5/\sqrt{20}}\right) \\ &= P\left(T_{19} \geq 2,6833\right) \\ &\cong 0,0074 \end{aligned}$$

Por lo tanto, siendo tan baja la probabilidad de haber observado un valor de 143, si la media realmente fuera 140, las autoridades deberían sospechar que la empresa no está cumpliendo con el requisito.

Generalmente, la utilización de la distribución t de Student está asociada a tamaños muestrales pequeños. Esto se debe a que la cantidad grados de libertad de la t de Student que corresponde a la distribución de muestreo de $\left(\bar{X} - \mu\right)/\left(s/\sqrt{n}\right)$ es igual al tamaño muestral menos uno: $n-1$.

Por lo tanto, si el tamaño muestral es grande, la distribución de muestreo de $\left(\bar{X} - \mu\right)/\left(s/\sqrt{n}\right)$, la cual es t de Student, puede aproximarse mediante una distribución Normal (ver sección anterior). En la práctica, si el tamaño muestral es superior a 30, se utiliza directamente la distribución

Normal Estándar, ya que no hay prácticamente diferencia entre los valores de probabilidad de esta última y los correspondientes a una t de Student con 29 g.l.

Por lo tanto, **la distribución t de student se utiliza en la práctica cuando se muestrean poblaciones Normales con varianza desconocida, y cuando el tamaño muestral es pequeño (menor a 30).**

4.7 Estimación: puntual y por intervalo

En el capítulo anterior, hemos visto el concepto de muestra aleatoria y mencionado algunas técnicas de muestreo utilizadas para su obtención. Asimismo, presentamos las distribuciones de muestreo de los principales estadísticos que se utilizan para realizar inferencias, a saber: la media muestral, la proporción muestral y la varianza muestral.

Además, vimos que siempre que se extrae una muestra se corre cierto riesgo de que la misma no sea representativa de la población. A su vez, al tomar diferentes muestras de una misma población, se obtendrán diferentes características muestrales (media, desvío estándar, etc.). Por ello, cuando se realizan inferencias estadísticas no solamente es importante realizar la estimación en sí, sino también medir de alguna manera el riesgo que tiene el resultado obtenido.

Cuando se realizan inferencias estadísticas, generalmente, se supone que un fenómeno determinado (o la población) tiene características similares a algún modelo probabilístico como los estudiados en el Capítulo 2. Por lo tanto, la población tiene una distribución con una función de densidad o probabilidad determinada que depende de ciertos parámetros, cuyos valores numéricos necesitamos conocer para asignar probabilidades a los sucesos. Por ejemplo, si se supone que la población es Normal, entonces, se deberán determinar los valores numéricos de los parámetros μ y σ; si la población es Poisson, se deberá determinar el valor del parámetro λ ; etc. De este modo, en el proceso de estimación, se realizan dos aproximaciones: en primer lugar se supone que la población pertenece a cierta familia de distribuciones (Normal, Poisson, etc.), y luego se estiman los valores numéricos de los parámetros que determinarán el miembro de la familia que caracteriza a la población (Normal estándar ó Normal con $\mu = 5$ y $\sigma = 2$, Poisson con $\lambda = 2$ ó con $\lambda = 5$, etc.).

En términos generales, la función de probabilidad poblacional depende de algún parámetro φ, que puede tomar ciertos valores. El proceso de "Estimación de Parámetros" desarrolla técnicas para construir "estimadores" de dicho parámetro, los cuales permiten asignar valores numéricos al mismo a partir de información muestral. El proceso consiste en especificar criterios para la determinación de los mejores estimadores de φ, a través de la estipulación de propiedades deseables de las estadísticas de muestra utilizadas como estimadores y el desarrollo de técnicas apropiadas para el proceso de estimación en sí.

Un **estimador puntual** (o simplemente estimador) es un estadístico muestral que se utiliza con el fin de inferir el valor de un parámetro poblacional desconocido[43].

Una **estimación** es la realización de un estimador en base a datos provenientes de *una* muestra aleatoria. Es un valor específico observado de un estimador.

Así, con el proceso de estimación puntual se pretende hallar una estimación univaluada del parámetro poblacional de interés, ya que mediante una muestra se estima un único valor del mismo. Por ejemplo, el estadístico $\bar{X}$ es un "estimador" de la media poblacional, y el valor específico $\bar{x}$ que tome cuando es extraída una muestra aleatoria (es decir, la realización de $\bar{X}$) es la "estimación puntual".

Ejemplo 21

Supongamos que la variable "estatura de los alumnos de la facultad" se distribuye Normalmente. El parámetro μ es desconocido, y su valor podría inferirse a través de un **estimador** como la media

[43] La definición de "estadístico muestral" (o simplemente "estadístico") y de "parámetro poblacional" fue expuesta en la Sección 2.1 del Capítulo 4.

muestral $\bar{X}$. Si se extrae una muestra aleatoria de 50 alumnos varones, como la del Ejemplo 6 del Capítulo 4, entonces $\bar{x} = 1,75$ es la **estimación puntual** de μ.

Para que la estimación puntual sea "buena", es deseable que el estimador cumpla ciertas propiedades que serán analizadas en el apartado siguiente.

Ejemplo 22

Si la población de interés es Normal, el parámetro μ es la media, la mediana y la moda poblacional, ya que se trata de una distribución simétrica. Por lo tanto, al extraer una muestra aleatoria, la media muestral, la mediana muestral y la moda muestral son tres posibles estimadores puntuales del parámetro poblacional μ. Teniendo en cuenta que lo más probable es que los valores muestrales de las tres características mencionadas no coincidan ¿Cómo decidimos cuál de ellos utilizar para estimar el parámetro? En la sección siguiente, veremos los criterios que permiten decidir, y concluiremos que el mejor estimador de μ es la media muestral $\bar{X}$.

El proceso de estimación puntual asigna un único valor al parámetro poblacional desconocido. Por lo tanto, la estimación puntual es correcta o equivocada y, generalmente, ocurrirá este segundo caso, ya que es muy poco probable que la estimación obtenida a partir de una única muestra coincida con el parámetro poblacional. Sin embargo, la estimación por sí sola no nos dice nada respecto de cuán equivocada puede ser nuestra estimación. Por ejemplo, si la estatura media de los alumnos de la faculta fuera 1,78m., la estimación del Ejemplo 1 no fue tan mala. Sin embargo, si la media poblacional fuera 1,85m., nuestra estimación estuvo muy alejada del verdadero valor. Por ello, resulta importante asignar alguna medida de riesgo a la estimación, o, mejor aún, indicar un intervalo en el cual puede estar contenido el valor del parámetro.

En el capítulo anterior, hemos visto que los estadísticos (y los estimadores) son variables aleatorias, ya que su valor varía de una muestra a otra. Por lo tanto, por más que usemos el mismo "estimador", la "estimación puntual" resultante variará de una muestra a otra. Sin embargo, al conocer la distribución de muestreo del estimador, podremos hallar un intervalo que con cierta probabilidad contendrá al parámetro. Así, como mediante la estimación puntual se asigna un solo valor numérico a cada parámetro desconocido, y esto generalmente no es del todo satisfactorio, se recurre a la construcción de un intervalo de confianza para el parámetro.

Un **Intervalo de Confianza** (IC) constituye un rango de valores posibles dentro del cual se encuentra el parámetro poblacional a estimar con un nivel de confianza determinado de antemano.

Los límites del intervalo se calculan en base a datos provenientes de una muestra aleatoria, y por lo tanto son aleatorios. Una vez extraída *una* muestra, se especifican los valores numéricos de los límites del intervalo, y decimos que tenemos un **Intervalo de Confianza estimado**.

Así, cuando se lleva a cabo una "estimación por intervalo" se utilizan los datos provenientes de una muestra aleatoria para determinar un intervalo de valores en cual se cree, con cierta probabilidad, que estará el parámetro poblacional de interés.

Reiteramos que toda inferencia estadística debe estar asociada a un nivel de riesgo determinado. En las estimaciones puntuales, el error estándar suele utilizarse como medida del riesgo mientras que, en los intervalos de confianza, esto es medido por el nivel de confianza.

El proceso de estimación que se estudiará en las secciones siguientes busca lograr afirmaciones como la siguiente:

"La estimación de la característica *A* es *xx*.

Con un *zz*% de confianza estará entre *LI* y *LS*".

En la proposición, *A* es la característica poblacional estudiada (estatura media, nivel de ingreso medio, proporción de fumadores, etc.), *xx* es la estimación puntual de la misma en base a los datos de *una* muestra aleatoria, *zz%* es el nivel de confianza del intervalo, cuyos limites inferior y superior son *LI* y *LS*, respectivamente.

4.8 Propiedades deseables de un Estimador

La definición de "estimador" fue expuesta en la sección anterior, y ya había sido introducida en el capítulo previo, cuando se definieron las "estadísticas" y se mencionó que las mismas eran utilizadas para estimar los valores de los parámetros. Pero se pueden definir muchos estadísticos para estimar el valor de un parámetro. Entonces, debemos tener criterios para juzgar qué tan bueno es un estimador, o de modo más general ¿qué es un estimador bueno?

De manera intuitiva, podemos pensar que para que no se produzcan sobre o subestimaciones, la distribución de muestreo del estadístico debería estar centrada en el valor del parámetro. Además, si se cumple lo anterior y el error estándar es pequeño, habrá una mayor probabilidad de que una realización del estadístico esté cerca del valor real del parámetro. Más adelante, formalizaremos estas dos propiedades deseables.

Sea J un estimador del parámetro φ. El error cuadrático medio de J es:

$$ECM(J) = E\left[(J-\varphi)^2\right]$$
$$= Var(J) + \left[\varphi - E(J)\right]^2$$

Si el estimador J cumple con las condiciones mencionadas en el párrafo previo, entonces el ECM será lo más pequeño posible, ya que, al estar su distribución centrada en el parámetro, el segundo término se anula, y como el error estándar es pequeño, también lo será su varianza.

Para hallar al mejor estimador de φ, podría buscarse a aquél que posea el menor ECM. Sin embargo, esto no es tarea fácil y en muchas circunstancias produce resultados ambiguos, ya que para la mayoría de las distribuciones estudiadas no existe ningún estimador que minimice el ECM para todos los posibles valores del parámetro (ver Canavos, 1997).

Considere a $J = g(X_1; X_2; ...; X_n)$ como un estadístico muestral obtenido a partir de una muestra de tamaño n. En los apartados siguientes, veremos qué propiedades debe cumplir J para que sea el "mejor" estimador del parámetro φ, y mencionaremos cuáles de las estadísticas que analizamos en el capítulo anterior cumplen cada una de las propiedades expuestas. Luego, en las secciones posteriores, mencionaremos algunas técnicas que permiten encontrar la función g que define a un estimador.

4.8.1 Insesgamiento

Como se ha mencionado, es deseable que un estimador tenga su distribución de muestreo centrada en el valor del parámetro que pretende estimar.

Un estadístico muestral $J = g(X_1; X_2; ...; X_n)$ es un **estimador insesgado** del parámetro φ, si su esperanza matemática es igual al parámetro:

$$E(J) = \varphi$$

El *sesgo* de un estimador es la diferencia entre su esperanza y el parámetro:

$$\text{Sesgo}(J) = E(J) - \varphi$$

Como puede observarse de la definición anterior, si un estimador es insesgado, entonces su sesgo es nulo. Además, en este caso, el ECM es igual a la varianza.

Algunos autores denominan *imparcialidad* a la propiedad de *insesgamiento* presentada en este apartado, la cual se refiere a que la media de la distribución de muestreo del estimador es igual al parámetro. Es decir que, en promedio, el estimador toma valores por encima del parámetro con la misma frecuencia que valores que están por debajo.

Esta propiedad no garantiza nada respecto de la estimación obtenida a partir de *una única muestra*. Lo que afirma es que, si un estimador es insesgado, entonces, el promedio de todas las estimaciones que se obtendrían con todas las muestras posibles de tamaño n coincidiría con el valor del parámetro. Sin embargo, ya hemos mencionado que es imposible tomar todas las muestras posibles de un determinado tamaño.

Si μ es la esperanza matemática de una distribución de probabilidad cualquiera (es decir, es la media poblacional), entonces,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

es un estimador insesgado de la misma.

Esta proposición ya fue expuesta en el capítulo anterior, donde vimos que $E\left(\bar{X}\right) = \mu$, y la demostración se dio en el apéndice de dicho capítulo.

Si la esperanza matemática de una distribución de probabilidad cualquiera es conocida, entonces:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$$

es un estimador insesgado de la varianza poblacional.

Esta proposición tiene poca aplicación práctica, ya que generalmente la media poblacional no se conoce. En este caso, utilizamos la varianza muestral como estimador.

La varianza muestral es un estimador insesgado de la varianza poblacional de cualquier distribución de probabilidades:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \qquad \text{(Es insesgado)}$$

El siguiente estadístico es un estimador sesgado de la varianza poblacional:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \qquad \text{(Es sesgado)}$$

La demostración excede el alcance de esta obra. El lector interesado puede consultar el Capítulo 9 de Novales (1997).

La proposición previa es la que justifica el uso del divisor $n-1$ en el cálculo de la varianza muestral. Es importante, además, mencionar que la proposición anterior no implica que s sea un estimador insesgado de σ.

La proporción muestral es un estimador insesgado de la proporción poblacional:

$$\bar{p} = \frac{X}{n} \qquad \text{(Es insesgado)}$$

La demostración de la proposición anterior está en el apéndice de este capítulo.

4.8.2 Eficiencia

La eficiencia de un estimador está relacionada con su estabilidad de muestra en muestra: decimos que un estimador es más eficiente que otro si es más estable de una muestra a otra. Por ejemplo, si tomamos muchas muestras, los valores (las realizaciones) de la media muestral estarán más concentrados que las observaciones de la mediana muestral o el modo muestral. Por lo tanto, la media muestral es un estimador más eficiente (estable) que la mediana muestral y el modo muestral. La estabilidad de las observaciones de una variable aleatoria está íntimamente relacionada con su desvío estándar, y por ende la eficiencia de un estimador estará relacionada con su error estándar: **cuanto más chico el error estándar, más eficiente es el estimador**.

En otras palabras, cuando se extrae una única muestra, ésta podría estar sesgada, y por lo tanto es importante tener una idea del riesgo que se podría estar cometiendo al realizar la estimación puntual. El error estándar del estimador es justamente una medida para ello. Cuanto más pequeño sea el error estándar, mejor o más confiables será nuestra estimación puntual.

Si bien existen técnicas para determinar si un estimador tiene la mínima varianza que podría poseer cualquier otro estimador del parámetro, no entraremos en detalles de estas cuestiones aquí. Simplemente, analizaremos la eficiencia en términos relativos, como una herramienta para comparar dos estimadores puntuales.

Sean J_1 y J_2 dos estimadores insesgados de φ. Si $Var(J_1) < Var(J_2)$, entonces J_1 es más eficiente que J_2.

La eficiencia relativa de J_1 respecto de J_2 es $Var(J_2)/Var(J_1)$.

Si los estimadores son sesgados, se suele utilizar el ECM para medir la eficiencia relativa.

Ejemplo 23

Si $\{X_1; X_2; ...; X_n\}$ es una muestra aleatoria de variables iid con media μ, entonces cada una de las X_i individualmente es un es estimador insesgado de μ, porque se verifica que $E(X_i) = \mu$. Sin embargo, el promedio simple de todas las observaciones, $\bar{X}$, es también un estimador insesgado, pero es más eficiente, ya que $Var(X_i) = \sigma^2 > Var(\bar{X}) = \sigma^2 / n$.

Considere todos los estimadores insesgados de φ. Entonces J^* es el **estimador insesgado de varianza mínima**, si $Var(J^*) \leq Var(J)$, siendo J cualquier otro estimador de insesgado de φ.

Si un estimador es insesgado de varianza mínima, entonces, es el más eficiente entre todos los estimadores del parámetro de interés, ya que su varianza es más pequeña que la de cualquier otro estimador. No entraremos en los detalles, pero puede demostrarse que $\bar{X}$ **es el estimador de mínima varianza de la media poblacional** (ver el Capítulo 8 de Canavos, 1997), y por lo tanto el más eficiente.

4.8.3 Consistencia

Es lógico esperar que cuanta más información muestral se posea, mejor sean las estimaciones que se realicen. Por lo tanto, un buen estimador debería mejorar a medida que aumenta el tamaño muestral.

Un estimador J_n del parámetro desconocido φ es **consistente** si "converge en probabilidad" al verdadero valor del parámetro[44]. Es decir, la probabilidad que la diferencia entre el estimador y el parámetro sea muy pequeña tiende al 100% cuando el tamaño *n* de la muestra crece:

$$\lim_{n\to\infty} P(|J_n - \varphi| < \varepsilon) = 1 \quad \text{para todo } \varepsilon > 0.$$

Algunos autores denominan esta propiedad "coherencia", e indica que al aumentar el tamaño de la muestra se tiene *casi la certeza* de que el valor del estimador se aproximará bastante al valor del parámetro.

En términos menos técnicos, la definición anterior quiere decir que, cuando el tamaño muestral aumenta, el valor de cualquier realización del estimador estará muy próximo al valor del parámetro. Por lo tanto, si se esperan que las observaciones estén muy cercanas al valor del parámetro, quiere decir que el sesgo y el error estándar del estimador se hacen cada vez más pequeños conforme crece la cantidad de observaciones de la muestra. A continuación, introduciremos un nuevo concepto relacionado con el sesgo de un estimador, y luego resumiremos las ideas presentadas aquí en un teorema.

[44] El subíndice n se utiliza para indicar que el estimador está calculado con una muestra aleatoria de tamaño n.

Un estimador es **asintóticamente insesgado,** si su sesgo tiende a cero cuando el tamaño de la muestra tiende infinito.

De la definición anterior resulta obvio que cualquier estimador insesgado resultará ser asintóticamente insesgado, ya el sesgo es cero para cualquier tamaño de muestra.

Ejemplo 24

La varianza muestral s^2 y la media muestral $\bar{X}$ son estimadores insesgados, y en consecuencia asintóticamente insesgados, de la varianza poblacional y la media poblacional, respectivamente. Por otro lado, el estimador S^2, expuesto en la Sección 8.1 de este capítulo, es un estimador sesgado. Pero su valor esperado tiende a la varianza poblacional cuando el tamaño de la muestra crece, por lo que su sesgo tiende a cero cuando n tiende a infinito y, por lo tanto, es un estimador asintóticamente insesgado.

Si un estimador J_n del parámetro φ **es asintóticamente insesgado y su varianza tiende a cero** cuando el tamaño muestral aumenta, entonces J_n **es un estimador consistente** de φ.

Puede demostrarse que la media muestral es un estimador consistente de la media poblacional. Este hecho es lo que se conoce como Ley de los Grandes Números, y es lo que permite utilizar el promedio de un número finito de observaciones para estimar la media de la distribución poblacional, teniendo en cuenta que la confiabilidad de este promedio es mayor que la de cualquiera de las observaciones.

Ley de los Grandes Números (LGN): Si $\{X_1; X_2; ...; X_n\}$ es una muestra aleatoria de variables iid con media μ y varianza σ^2 (ambos finitos), entonces, $\bar{X}_n$ es un estimador consistente de μ.

Utilizando la LGN y la desigualdad de Chebyshev, es posible determinar el tamaño muestral necesario para asegurar con una determinada probabilidad que la media muestral no se alejará de la media poblacional más allá de un valor especificado. En la sección final de este capítulo, volveremos sobre este tema.

Antes de proseguir haremos una última aclaración: En una *distribución simétrica* (la mediana poblacional coincide con la media poblacional), la *mediana muestral* es un estimador insesgado y consistente de la *mediana poblacional*, ya que cuando aumenta el tamaño muestral, las realizaciones de este estimador se acercan al valor de la característica poblacional. Sin embargo, el error estándar de la *mediana muestral* es más grande que el de la *media muestral*. Por lo tanto, **en una distribución simétrica, la *media muestral* es un estimador más eficiente de la *mediana poblacional* que la *mediana muestral***.

4.8.4 Suficiencia

En la Sección 8.1 se mencionó que $\bar{X}$ es un estimador insesgado de μ en cualquier población. Asimismo, en el Ejemplo 3 de la Sección 8.2 vimos que cualquier elemento de la muestra aleatoria será un estimador insesgado de la media poblacional: si la muestra está compuesta por las siguientes variables iid $\{X_1; X_2; ...; X_n\}$, entonces, la esperanza de cada una de las variables X_i será μ, y por lo tanto, $J_1 = X_1$ será un estimador insesgado de μ. Es más, cualquier combinación lineal del tipo:

$$J = \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_n X_n \qquad \text{con} \sum_{i=1}^{n} \alpha_i = 1$$

es un estimador insesgado de μ (la demostración de esta afirmación se propone como ejercicio). Sin embargo, obviamente, $J_1 = X_1$ no será el mejor estimador de la media poblacional, ni mucho menos. Por ejemplo, si se toman una muestra de 50 alumnos para estimar la estatura media de los alumnos de la facultad, la primera observación no sería una muy buena estimación, ya que podría haberse observado un alumno muy alto o muy bajo. Esto nos lleva a pensar que un buen estadístico debe utilizar la mayor cantidad de información posible contenida en la muestra.

Cuando se construye un estimador, se busca utilizar toda la información contenida en la muestra. De manera general, podemos decir que cuando un estimador es **suficiente** es porque utiliza una cantidad de información tal que ningún otro estimador podría sacar información adicional sobre el parámetro de la población a partir de la muestra.

Como se mencionó antes, si se toma una muestra aleatoria de tamaño n, el estimador $J = X_1$ es un estimador insesgado de la media poblacional, pero no utiliza toda la información muestral y, por lo tanto, no es suficiente.

La suficiencia de un estimador es quizás la característica más importante, pero la complejidad matemática que contienen los desarrollos asociados con una exposición rigurosa de la misma excede el alcance de este libro.

4.9 Estimación Puntual: métodos

Cuando se extrae una única muestra, y con las observaciones obtenidas con la misma se calcula el valor de un estimador, estamos realizando una **estimación puntual**. En el apartado anterior, hemos visto las propiedades que son deseables para que un estimador sea considerado "bueno". Aquí veremos algunos métodos que suelen emplearse para obtener estimadores, es decir, para obtener la función que transforma los datos de una muestra en un número que corresponde a la estimación del parámetro de interés.

Por ejemplo, la función para estimar la media poblacional hemos visto que es:

$$(X_1 + X_2 + ... + X_n)/n$$

En lo apartados siguientes mencionaremos algunos métodos que permiten obtener esta fórmula, y, de modo más general, cómo obtener la función g que permite definir al estimador $J = g(X_1; X_2; ...; X_n)$. Simplemente, expondremos la lógica de cada método, sin profundizar en su implementación ni en los aspectos prácticos. Puntualmente, se expondrán los métodos de *máxima verosimilitud* y de *momentos*. Otro método importante de estimación es el de *mínimos cuadrados*, pero el mismo será presentado en el Capítulo 6, en el contexto del análisis de *regresión lineal*.

4.9.1 Máxima verosimilitud

Cuando se extrae una muestra de una población que se supone tiene función de densidad $f(x;\varphi)$, el interés recae en obtener un valor para el parámetro desconocido φ. El método de "máxima verosimilitud" consiste en asignar al parámetro aquel valor que maximiza la probabilidad de que la muestra a extraer sea la muestra realmente observada.

Por ejemplo, considere una muestra de tamaño 4: $\{X_1; X_2; X_3; X_4\}$. Suponga que extrae una muestra con las siguientes observaciones: $\{5; 2; 0,75; 3\}$. El método asigna al parámetro poblacional el valor que maximiza la probabilidad de que simultáneamente $X_1 = 5$, $X_2 = 2$, $X_3 = 0,75$ y $X_4 = 3$.

De modo más general, para cualquier muestra de tamaño n, el método maximizará la *probabilidad conjunta* de que $X_1 = x_1$, $X_2 = x_2$,…, y $X_n = x_n$. Es decir, maximiza la probabilidad de que las variables aleatorias que componen la muestra aleatoria tomen los valores que se obtienen en una realización particular de la misma.

Ejemplo 25 [45]

Suponga que en una población se desea estudiar cuántos apoyan a cierto candidato político. Para ello, se extrae una muestra de 20 habitantes, de los cuales 8 están a favor del candidato. ¿Cuál es el valor del estimador de máxima verosimilitud de p?

Para calcularlo, debemos obtener el valor del parámetro p que maximiza la probabilidad (binomial) de que en una muestra de tamaño 20, se obtengan 8 éxitos. Es decir:

$$\max_p l(p) = \max_p \binom{20}{8} p^8 (1-p)^{20-8}$$

[45] Este ejemplo requiere que el lector conozca algunos conceptos de análisis matemático.

Para facilitar el procedimiento, se toman logaritmos:

$$L(p) = \ln\left[l(p)\right]$$
$$= \ln\binom{20}{8} + 8\times\ln(p) + 12\times\ln(1-p)$$

Derivando e igualando a cero obtenemos:

$$\frac{\partial L(p)}{\partial p} = \frac{8}{p} + \frac{12}{1-p}\times(-1) = 0$$

$$\Rightarrow \frac{8}{p} = \frac{12}{1-p} \qquad \Rightarrow 8-8p = 12p$$

$$\Rightarrow \qquad p = \frac{8}{20}$$

El ejemplo anterior puede generalizarse fácilmente: si tomamos una muestra de tamaño n y obtenemos x éxitos, el estimador de máxima verosimilitud de p es x/n. Es decir, la *proporción muestral*.

Es importante mencionar que lo expuesto aquí simplemente ilustra la idea de la técnica sin profundizar demasiado en ella. En Novales (1997) y Canavos (1997) se encuentra una exposición detallada del tema, incluyendo los estimadores máximo-verosímiles de los parámetros de las distribuciones más conocidas y una descripción de las principales propiedades de este tipo de estimadores. Sin más, en la siguiente tabla se ilustran las principales poblaciones utilizadas en la práctica y los estimadores que se obtienen con el método de Máxima Verosimilitud:

Población	Parámetro	Estimador MV
Normal	μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
	σ^2	$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
Binomial		$\bar{p} = \frac{X}{n}$
Poisson	λ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

4.9.2 Momentos

El método de estimación por Momentos se basa en el concepto de que toda distribución puede caracterizarse, al menos parcialmente, por sus momentos. De este modo, **el método iguala algunos momentos muestrales**, calculados de acuerdo a lo visto en el Capítulo 2, **con los momentos teóricos expresados en términos de los parámetros** de una distribución determinada.

Por ejemplo, para el caso de una distribución Normal, la media coincide con el parámetro μ y la varianza con el parámetro σ^2, y por lo tanto la media muestral y la varianza muestral serán los estimadores propuestos para estos parámetros.

Si se supone que la población sigue una distribución Gamma, entonces, tenemos que:

$$E(X) = \frac{\beta}{\alpha}; \qquad \text{y} \qquad Var(X) = \frac{\beta}{\alpha^2}$$

Resolviendo estas dos ecuaciones de manera simultánea, obtenemos que:

$$\alpha = \frac{E(X)}{Var(X)}; \quad \text{y} \qquad \beta = \frac{\left[E(X)\right]^2}{Var(X)}$$

Finalmente, remplazando los momentos teóricos por los muestrales, obtendríamos los estimadores por momentos de los parámetros:

$$\alpha = \frac{\bar{X}}{s^2}; \text{ y} \qquad \beta = \frac{\bar{X}^2}{s^2}$$

Con esta lógica se pueden calcular los momentos de las distintas distribuciones. A continuación, se expone el resumen de los estimadores que se obtendrían para las poblaciones más utilizadas en la práctica.

Población	Parámetro	Estimador Momentos
Normal	μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
	σ^2	$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
Binomial	p	$\frac{X}{n}$
Poisson	λ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Al igual que en el apartado anterior, remitimos al lector a Novales (1997) o Canavos (1997) para profundizar en el tema.

En general, resulta mucho más útil especificar un rango de valores dentro del cual puede estar contenido el parámetro poblacional a analizar. Esto nos lleva directamente a la estimación de Intervalos de Confianza.

4.10 Intervalos de Confianza (IC)

Cuando realizamos estimaciones por intervalos, deseamos conocer un rango de valores dentro del cual es probable que esté el parámetro. En términos generales, si especificamos un nivel de probabilidad (denominado nivel de confianza) con el cual deseamos realizar la estimación, $1-\alpha$, podemos decir que buscamos dos límites de un intervalo, k_1 y k_2, tal que la probabilidad de que la diferencia entre el estimador y el parámetro esté en el rango especificado sea $1-\alpha$: Es decir:

$$P(k_1 < J - \varphi < k_2) = 1 - \alpha$$

Por ejemplo, podemos decir para un fenómeno específico que la probabilidad de que la diferencia entre $\bar{X}$ y μ esté entre -2 y 2 es de 90 %. En este caso, tendríamos que los límites del intervalo deseado son $k_1 = -2$ y $k_2 = +2$, y el nivel de probabilidad con que se realiza la estimación es $1-\alpha = 0{,}9$, por lo que $\alpha = 0{,}1$. Es decir:

$$P\left(-2 < \bar{X} - \mu < 2\right) = 0{,}90$$

Obviamente, los límites del rango expuesto, k_1 y k_2, dependerán tanto del nivel de probabilidad deseado como del error estándar del estadístico. Por un lado, cuanto mayor sea el error estándar, más amplio será el intervalo para una misma probabilidad, y por otro, para un mismo error estándar, mayor será el intervalo para un mayor nivel de probabilidad.

Sin embargo, nos interesa determinar un intervalo en el cual esté el parámetro, y no solamente especificar el intervalo de la diferencia entre éste y su estimador. Para ello, el camino lógico a seguir es remplazar el estimador por una realización del mismo (una estimación) obtenida a través de una muestra aleatoria. Manipulando la expresión anterior, podemos escribir[46]:

$$P\left[J-k_2\left(\sigma_J;\alpha\right)<\varphi<J-k_1\left(\sigma_J;\alpha\right)\right]=1-\alpha$$

Si remplazamos el estimador J, por una realización del mismo, j, no podremos hablar más de probabilidad, ya que no habría más variables aleatorias involucradas en la expresión (recordamos que un parámetro es un número fijo desconocido). Por lo tanto, la expresión entre corchetes sería verdadera o falsa, lo cual podría determinarse si conociéramos el verdadero valor de φ.

Por lo expuesto, cuando se realiza una estimación de los límites del intervalo, hablamos de *confianza* en lugar de probabilidad, y generalmente escribiremos:

$$C\left[j-k_2\left(\sigma_J;\alpha\right)<\varphi<j-k_1\left(\sigma_J;\alpha\right)\right]=1-\alpha$$

Siguiendo con el ejemplo anterior, si extraemos una muestra y obtenemos que $\bar{x}=10$, entonces, podemos decir que con un 90% de confianza la media poblacional está entre 8 y 12:

$$\begin{aligned} C\left(-2<10-\mu<+2\right) &= C\left(10-2<\mu<10+2\right) \\ &= C\left(8<\mu<12\right) \\ &= 90\% \end{aligned}$$

Calcular las realizaciones de cada estimador ya es tarea corriente. Lo que necesitamos determinar es algún método para hallar los valores de k_1 y k_2, los cuales, conjuntamente con la estimación puntual, permitirán establecer el intervalo de confianza del parámetro. Para ello, recurriremos frecuentemente a las distribuciones de muestreo estudiadas en el capítulo anterior.

Antes de estudiar las técnicas de construcción de este tipo de intervalos, remarquemos claramente qué implica un nivel de confianza dado. El IC se construye con *una realización* del estimador, la cual depende de la muestra extraída. Es posible, y muy probable, que si tomáramos otra muestra se obtendría otra realización distinta de J y, por lo tanto, otro intervalo distinto. El nivel de confianza $1-\alpha$ nos dice que si se extraen infinitas muestras de tamaño n y con cada una de ellas se construye el intervalo, el $100\times(1-\alpha)\%$ de los mismos contendrá al verdadero valor de φ.

Por ejemplo, si se extraen 1.000 muestras y con cada una de ellas se calcula el intervalo para φ con un 95% de confianza, esperamos que 950 de estos intervalos contengan al verdadero valor del parámetro. Lo que implica el nivel de confianza, es que "confiamos" en que nuestra muestra particular es una de las 950 que genera un intervalo *correcto* (que contiene a φ), pero "no tenemos la certeza" de ello, ya que la muestra obtenida podría ser una de las 50 que genera un intervalo *equivocado* (que no contienen al verdadero valor de φ). Nunca sabremos si el intervalo es correcto o equivocado, a menos que conozcamos el verdadero valor φ, y en cuyo caso no tendría sentido construir un IC.

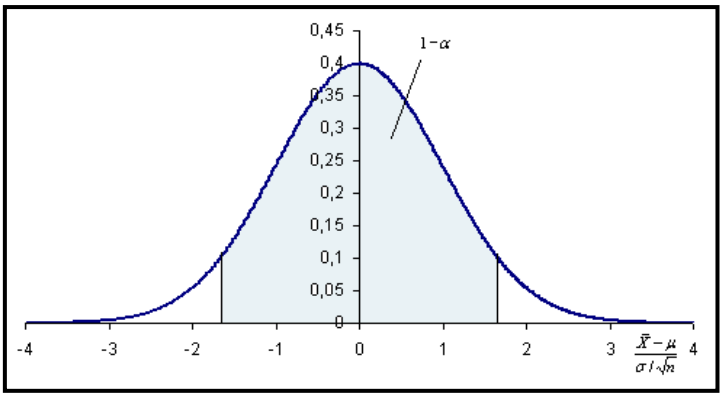
4.10.1 Media en poblaciones normales con varianza conocida

$\bar{X}$ es un "buen" estimador de la media poblacional, en el sentido de que posee las propiedades deseables que se expusieron en la Sección 8. En el capítulo anterior, vimos que $\bar{X}\sim N(\mu;\frac{\sigma^2}{n})$, cuando se muestrean poblaciones $N\left(\mu;\sigma^2\right)$ ó cuando no se conoce la distribución pero n es lo suficientemente grande (siendo σ^2 conocido). Por lo tanto, tenemos que:

[46] Primero restamos J en cada miembro, luego multiplicamos por -1 (recordando invertir el sentido de las desigualdades), y finalmente reordenamos. Además, en la expresión dejamos claro que los límites dependen del valor del error estándar del estadístico y de la probabilidad deseada.

$$P\left(z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2} \right) = 1-\alpha$$

Donde $z_{\alpha/2}$ y $z_{1-\alpha/2}$ son los valores correspondientes a una Normal Estándar que acumulan (a izquierda) una probabilidad de $\alpha/2$ y $1-\alpha/2$, respectivamente, como puede observarse en la figura.



Teniendo en cuenta que la Normal estándar es simétrica respecto del origen, tenemos que $z_{\alpha/2} = -z_{1-\alpha/2}$. Utilizando esta relación, y reagrupando, obtenemos que:

$$\begin{aligned} 1-\alpha &= P\left(-z_{1-\alpha/2} \times \sigma/\sqrt{n} \le \bar{X}-\mu \le z_{1-\alpha/2} \times \sigma/\sqrt{n}\right) \\ &= P\left(-\bar{X} - z_{1-\alpha/2} \times \sigma/\sqrt{n} \le -\mu \le -\bar{X} + z_{1-\alpha/2} \times \sigma/\sqrt{n}\right) \\ \Rightarrow \quad & P\left(\bar{X} - z_{1-\alpha/2} \times \sigma/\sqrt{n} \le \mu \le \bar{X} + z_{1-\alpha/2} \times \sigma/\sqrt{n}\right) = 1-\alpha \end{aligned}$$

El intervalo de confianza de un $100\times(1-\alpha)\%$ para la media poblacional μ cuando se conoce el desvío estándar σ es:

$$C\left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \right) = 1-\alpha$$

Ejemplo 26

Suponga que la estatura de los alumnos varones de la Facultad de Ciencias Económicas se distribuye Normalmente, y se sabe que el desvío poblacional es de 15 cm. ¿Cuál es el intervalo de confianza que contendrá a la verdadera media poblacional con un 99%, si se obtuvo la siguiente muestra de 20 datos?

Varones				
1,83	1,82	1,91	1,80	1,85
1,76	1,70	1,80	1,70	1,83
1,76	1,72	1,91	1,84	1,90
1,76	1,77	1,76	1,74	1,81

Si calculamos el promedio (la realización de $\bar{X}$ para esta muestra particular), tenemos que $\bar{x} = 1,75$. Teniendo en cuenta que $\alpha = 0,01$ (porque $1-\alpha = 99\%$), buscamos en la tabla Normal Estándar el valor que acumula $1-\alpha/2 = 0,995$, y tenemos que $z_{0,995} = 2,5758$. Finalmente, como sabemos que $n = 20$ y $\sigma = 0,15$, podemos construir el IC:

$$C\left(\bar{x} - z_{0,995}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{0,995}\frac{\sigma}{\sqrt{n}} \right) = 0,99$$

$$\begin{aligned} C\left(1,80 - 2,5758\frac{0,15}{\sqrt{20}} \le \mu \le 1,75 + 2,5758\frac{0,15}{\sqrt{20}}\right) &= C\left(1,7120 \le \mu \le 1,8848\right) \\ &= 0,99 \end{aligned}$$

Es decir, que con un 99% la estatura media de los varones estará entre 1,66m. y 1,84m.

El intervalo del ejemplo anterior se obtuvo a partir de *una realización* de $\bar{X}$, y si tomáramos otra muestra, obtendríamos posiblemente otra realización distinta del estimador y, por lo tanto, otro intervalo distinto. Un nivel de confianza de, por ejemplo, un 99%, significa que si se extraen infinitas muestras de tamaño n y con cada una de ellas se construye el intervalo, el 99% de los mismos contendrá a la verdadera media poblacional.

De este modo, si el 99% de los intervalos que se construyen contienen al verdadero valor poblacional, de alguna manera "confiamos" en que nuestra muestra particular genera uno de estos intervalos *correctos*, pero no tenemos la certeza de ello. Es decir, que nuestra muestra particular podría ser una de las que genera el 1% de los intervalos *equivocados* que no contienen al verdadero valor del parámetro. Nunca lo sabremos, a menos que conozcamos el verdadero valor del parámetro y, en cuyo caso, no tendría sentido realizar inferencias.

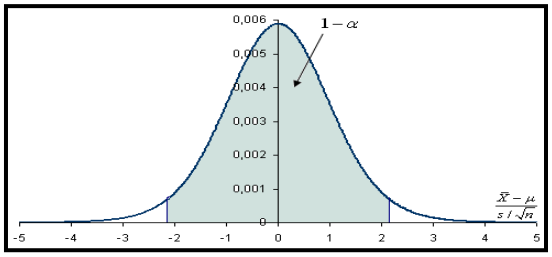
4.10.2 Media en poblaciones normales con varianza desconocida

En la Sección 6, hemos visto que cuando se trabaja con una muestra aleatoria de tamaño n de una población Normal con media y varianza desconocida, la distribución de muestreo de la siguiente variable es *T de Student* con $n-1$ g.l. (grados de libertad)

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

En base a ello, podemos expresar que:

$$P\left(t_{n-1;\alpha/2} \le \frac{\bar{X}-\mu}{s/\sqrt{n}} \le t_{n-1;1-\alpha/2} \right) = 1-\alpha$$



Donde $t_{n-1;\alpha/2}$ y $t_{n-1;1-\alpha/2}$ son los valores correspondientes a una *T de Student* con *n*-1 g.l. que acumulan una probabilidad de $\alpha/2$ y $1-\alpha/2$, respectivamente (ver Figura). Al igual que la Normal Estándar, la *t de Student* es simétrica respecto del origen y, por lo tanto, $t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$. Si utilizamos esta relación entre los valores de la distribución de muestreo y reagrupamos la expresión anterior, podemos obtener el intervalo aleatorio que con una probabilidad de $1-\alpha$ contiene a μ:

$$\begin{aligned} 1-\alpha &= P\left(-t_{n-1;1-\alpha/2} \times s/\sqrt{n} \le \bar{X}-\mu \le t_{n-1;1-\alpha/2} \times s/\sqrt{n} \right) \\ &= P\left(-\bar{X} - t_{n-1;1-\alpha/2} \times s/\sqrt{n} \le -\mu \le -\bar{X} + t_{n-1;1-\alpha/2} \times s/\sqrt{n} \right) \\ \Rightarrow \quad & P\left(\bar{X} - t_{n-1;1-\alpha/2} \times s/\sqrt{n} \le \mu \le \bar{X} + t_{n-1;1-\alpha/2} \times s/\sqrt{n} \right) = 1-\alpha \end{aligned}$$

Cuando obtenemos una muestra y, a partir de ella, una realización de los estadísticos $\bar{X}$ y s^2, obtenemos el Intervalo de Confianza para la media poblacional.

El intervalo de confianza de un $100\times(1-\alpha)\%$ para la media poblacional μ **cuando el desvío estándar poblacional es desconocido** es:

$$C\left(\bar{x}-t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\leq\mu\leq\bar{x}+t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right)=1-\alpha$$

Ejemplo 27

Supongamos que una empresa desea controlar la vida útil de cierto producto. Para ello, se extrae una muestra de 20 artefactos y se los enciende hasta que fallen. La duración de cada artefacto se muestra en la tabla.

774	759	755	724
660	763	742	601
667	707	665	644
696	699	778	780
691	663	765	575

Si la distribución de la duración es Normal, ¿cuál es el intervalo en el cual se encuentra la duración media con un 99% de confianza?

En primer lugar, con los datos calculamos la media muestra y el desvío muestral, los cuales son: $\bar{x}=705,4$ y $s=60$. Luego, buscamos el cuantil de la distribución T de Student con 19 g.l que acumula un 99,5% de probabilidad. Estos valores se justifican en que el tamaño muestral es $n=20$ y $\alpha=1\%$, entonces, $n-1=19$ y $1-\alpha/2=99,5\%$. Buscando en la tabla, obtenemos que $t_{19;0,995}=2,8609$ y, por tanto, el intervalo de confianza deseado es:

$$C\left(705,4-2,8609\frac{60}{\sqrt{20}}\leq\mu\leq705,4+2,8609\frac{60}{\sqrt{20}}\right)=C\left(667,01\leq\mu\leq743,78\right)$$
$$=0,99$$

En palabras: con un 99% de confianza podemos decir que la duración media de las lamparitas está entre 666,02 y 744,78 horas.

Recordemos la interpretación de un intervalo de confianza en el contexto del ejemplo previo: si extraemos infinitas muestras de tamaño 20, el 99% de las mismas generará un intervalo que contendrá a la verdadera duración media, pero no sabemos si nuestra muestra particular es una de la que pertenece a este 99% que genera intervalos "correctos".

4.10.3 Media en poblaciones no Normales

Por aplicación del Teorema Central del Límite, hemos visto que cuando no se conoce la distribución de la población, la distribución de muestreo de $\bar{X}$ puede aproximarse mediante una Normal cuando el tamaño muestral es suficientemente grande. De esta manera, aplicamos lo descrito en el Apartado 4.1, y obtenemos un intervalo de confianza *aproximado*.

Si, además de no saber si la muestra proviene de una población Normal, se desconoce la varianza poblacional, la distribución de muestreo de la variable $\left(\bar{X}-\mu\right)/\left(s/\sqrt{n}\right)$ puede *aproximarse* mediante una distribución *T de Student*. Por lo tanto, en este caso, el intervalo *aproximado* se construirá utilizando la técnica descrita en el apartado anterior.

4.10.4 Varianza en poblaciones Normales con media desconocida

En la Sección 5, expusimos la distribución de muestreo de la varianza muestral s^2 cuando se trabaja con una muestra aleatoria de tamaño n de una población Normal con media desconocida. Más precisamente, hemos expuesto que la distribución de la siguiente variable es *Chi-cuadrado* con $n-1$ g.l.

$$\frac{(n-1)s^2}{\sigma^2} \sim X^2_{n-1}$$

Con esta información, podemos escribir que:

$$P\left(\chi^2_{n-1;\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1;1-\alpha/2} \right) = 1-\alpha$$

La distribución Chi-cuadrado *no es simétrica* por lo que podremos recurrir a la simplificación de los apartados anteriores, y sí o sí deberemos conseguir dos valores de la distribución de muestreo. Manipulando la expresión anterior, podemos construir el intervalo aleatorio que contiene a la varianza poblacional σ^2 con una probabilidad de $1-\alpha$.

$$P\left(\frac{1}{\chi^2_{n-1;1-\alpha/2}} \le \frac{\sigma^2}{(n-1)s^2} \le \frac{1}{\chi^2_{n-1;\alpha/2}} \right) = 1-\alpha$$

Note que en la expresión anterior el valor de la chi-cuadrado que acumula $1-\alpha/2$ quedó a la izquierda, y el que acumula $\alpha/2$ quedó a la derecha[47]. Finalmente,

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} \right) = 1-\alpha$$

El intervalo de confianza de un $100\times(1-\alpha)\%$ para la varianza poblacional σ^2 cuando se muestrean poblaciones Normales es[48]:

$$C\left(\frac{(n-1)s_x^2}{\chi^2_{n-1;1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)s_x^2}{\chi^2_{n-1;\alpha/2}} \right) = 1-\alpha$$

Ejemplo 28

Considere el Ejemplo 27 de la Sección 10.1, donde se obtuvo un desvío muestral $s = 60$ con una muestra de $n = 20$ lamparitas. ¿Cuál es el intervalo que con un 95% de confianza contiene a la varianza poblacional?

Teniendo en cuenta que el intervalo se construye para la varianza, en primer lugar, calculamos $s^2 = 60^2 = 3600$. Luego, teniendo en cuenta el nivel de confianza del 95%, tenemos que: $\alpha = 0,05$, $\alpha/2 = 0,025$ y $1-\alpha/2 = 0,975$. Además, teniendo en cuenta que la muestra es de 20 elementos, utilizaremos una Chi-cuadrado con 19 grados de libertad. Los valores necesarios de la Chi-cuadrado son: $\chi^2_{19;\,0,025} = 8,9065$ y $\chi^2_{19;\,0,975} = 32,8594$. Luego, el intervalo de confianza es:

$$\frac{19\times 3600}{32,8526} \le \sigma^2 \le \frac{19\times 3600}{8,9065} \qquad \Rightarrow \qquad 2082 \le \sigma^2 \le 7681$$

Por lo tanto, podemos decir que con un 95% de confianza el desvío estándar estará entre $\sqrt{2028} \cong 45,63$ y $\sqrt{7681} \cong 87,64$.

4.10.5 Proporción en poblaciones Binomiales

En el Capítulo 2 se vio que cuando el parámetro n es lo suficientemente grande, la distribución binomial podía aproximarse con la Normal. Utilizando esta propiedad, en el Capítulo 4, hemos visto que cuando el tamaño de la muestra es grande, la distribución de muestreo de la siguiente variable puede aproximarse por una Normal Estándar:

[47] Esto se debe a que si $a \le b \Rightarrow 1/a \ge 1/b$.

[48] Utilizamos s_x^2 en lugar de s^2 para indicar que estamos haciendo referencia a una estimación calculada a partir de una muestra, es decir que s_x^2 hace referencia a una *realización* de s^2.

$$\frac{\bar{p}-p}{\sqrt{p(1-p)/n}} \sim N(0;1)$$

Por lo tanto, utilizando la simetría de la distribución Normal, podemos escribir:

$$P\left(-z_{1-\alpha/2} \le \frac{\bar{p}-p}{\sqrt{p(1-p)/n}} \le z_{1-\alpha/2}\right) = 1-\alpha$$

Manipulando el argumento de la probabilidad del miembro izquierdo, obtenemos:

$$\begin{aligned} 1-\alpha &= P\left(-z_{1-\alpha/2} \times \sqrt{p(1-p)/n} \le \bar{p}-p \le z_{1-\alpha/2} \times \sqrt{p(1-p)/n}\right) \\ &= P\left(-\bar{p}-z_{1-\alpha/2} \times \sqrt{p(1-p)/n} \le -p \le -\bar{p}+z_{1-\alpha/2} \times \sqrt{p(1-p)/n}\right) \\ \Rightarrow \quad & P\left(\bar{p}-z_{1-\alpha/2} \times \sqrt{p(1-p)/n} \le p \le \bar{p}+z_{1-\alpha/2} \times \sqrt{p(1-p)/n}\right) = 1-\alpha \end{aligned}$$

En este caso, para poder construir el intervalo de confianza estimado, debemos utilizar la estimación puntual $\bar{p}_x$ para calcular el desvío estándar del estimador $\bar{p}$. Es decir, debemos remplazar $\bar{p}_x$ por $\bar{p}$ en las raíces cuadradas que aparecen en ambos extremos del intervalo[49].

El intervalo de confianza de un $100\times(1-\alpha)\%$ para la proporción poblacional p, cuando el tamaño de la muestra es suficientemente grande, es:

$$C\left(\bar{p}_x - z_{1-\alpha/2} \times \sqrt{\frac{\bar{p}_x(1-\bar{p}_x)}{n}} \le p \le \bar{p}_x + z_{1-\alpha/2} \times \sqrt{\frac{\bar{p}_x(1-\bar{p}_x)}{n}}\right) = 1-\alpha$$

Ejemplo 29

Suponga que, en una encuesta tomada a 500 personas, en una determinada ciudad, 350 respondieron que apoyan al candidato A, lo que significa que la proporción muestral es de un 70%. Entonces, sabiendo que $z_{0,975} = 1,96$, el intervalo de 95% de confianza para la proporción poblacional es

$$\begin{aligned} 0,95 &= C\left(0,60-1,96\times\sqrt{\frac{0,70\times0,30}{500}} \le p \le 0,66+1,96\times\sqrt{\frac{0,70\times0,30}{500}}\right) \\ &= C\left(0,5571 \le p \le 0,6429\right) \end{aligned}$$

Siendo el límite inferior del intervalo mayor a 0,50, el resultado de esta encuesta permite afirmar con un 95% de confianza que el candidato A ganará las elecciones.

4.11IC para comparar poblaciones

Muchas veces se desean comparar poblaciones para saber si sus medias son iguales o no. Por ejemplo, el gobierno Nacional podría estar interesado en comparar el ingreso medio de los habitantes de dos provincias para saber si son iguales, o si su diferencia supera una cantidad determinada. O una empresa podría estar interesada en comprar la proporción de artículos defectuosos que generan dos máquinas para utilizar aquélla que tenga mejor funcionamiento. También, puede darse el caso en que se desee comparar la variabilidad de cierta característica, y el interés recaería en las varianzas.

Estas ideas nos llevan, directamente, a la construcción de intervalos de confianza para la "diferencia de medias", la "diferencia de proporciones" y el "cociente de varianza"

[49] Al igual que con la varianza, utilizamos un subíndice x para indicar que se trata de una realización de $\bar{p}$.

4.11.1 Diferencia de medias

En primer lugar, debemos establecer la distribución de muestreo de la diferencia de medias muestrales.

Sean $\{X_1; X_2; ...; X_n\}$ y $\{Y_1; Y_2; ...; Y_m\}$ dos muestras aleatorias de tamaños n y m, respectivamente, provenientes de dos poblaciones Normales independientes con medias μ_X y μ_Y, y varianzas σ_X^2 y σ_Y^2. Entonces:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{{\sigma_X}^2}{n} + \frac{{\sigma_Y}^2}{m}}} \sim N(0; 1)$$

Sobre la base de la distribución de muestreo expuesta, podemos construir el Intervalo de Confianza correspondiente. Sabemos, entonces, que:

$$P\left[-z_{1-\alpha/2} \le \frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_X - \mu_Y\right)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \le z_{1-\alpha/2} \right] = 1 - \alpha$$

Reagrupando la expresión anterior, tenemos que:

$$P\left[\left(\bar{X} - \bar{Y}\right) - z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \le \left(\mu_X - \mu_Y\right) \le \left(\bar{X} - \bar{Y}\right) + z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right] = 1 - \alpha$$

Finalmente, una vez que se obtienen las dos muestras, se pueden obtener las estimaciones de $\bar{X}$ e $\bar{Y}$, y con ellas estimar el intervalo.

El intervalo de confianza de un $100 \times (1-\alpha)\%$ para la diferencia de medias, $\mu_X - \mu_Y$, cuando se muestrean dos poblaciones Normales con varianzas conocidas σ_X^2 y σ_Y^2 es:

$$C\left[\bar{x} - \bar{y} - z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \le \mu_X - \mu_Y \le \bar{x} - \bar{y} + z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right] = 1 - \alpha$$

Ejemplo 30

Considere el Ejemplo 26 de la Sección 10.1 donde se construyó el IC para l media de estaturas de varones (en metros) con 20 datos (ver tabla) de Varones.

Varones				
1,83	1,82	1,91	1,80	1,85
1,76	1,70	1,80	1,70	1,83
1,76	1,72	1,91	1,84	1,90
1,76	1,77	1,76	1,74	1,81

Mujeres				
1,64	1,71	1,66	1,76	1,75
1,62	1,63	1,63	1,62	1,87
1,66	1,65	1,64	1,61	1,83
1,83	1,63	1,81	1,70	1,63
1,59	1,53	1,81	1,77	1,83
1,66	1,68	1,78	1,58	1,76

Supongamos que se extrae una muestra con 30 datos de mujeres (ver tabla). Si la distribución de estaturas de mujeres es Normal con desvío estándar de 10cm (o 0,10 m). e independiente de las estaturas de varones, ¿cuál es el IC al 99% para la diferencia de las medias?

El intervalo a construir es:

$$C\left[\bar{v}-\bar{m}-z_{0,995}\sqrt{\frac{\sigma_V^2}{n_V}+\frac{\sigma_M^2}{n_M}} \le \mu_V-\mu_M \le \bar{v}-\bar{m}+z_{0,995}\sqrt{\frac{\sigma_V^2}{n_V}+\frac{\sigma_M^2}{n_M}}\right]=0,99$$

Para calcular los valores, en primer lugar, sabemos que en una Normal Estándar $z_{0,995}=2,5758$. Luego, con la información del ejemplo tenemos que:

$$\bar{v}=1,80 \qquad\qquad \bar{m}=1,70$$

$$\sqrt{\frac{\sigma_V^2}{n_V}+\frac{\sigma_M^2}{n_M}}=\sqrt{\frac{0,15^2}{20}+\frac{0,10^2}{30}} \cong 0,038188$$

Remplazando en la expresión anterior, tenemos que:

$$\begin{aligned} 0,99 &= C\left[1,80-1,70-2,5759\times 0,038188 \le \mu_V-\mu_M \le 1,80-1,70+2,5759\times 0,038188\right] \\ &= C\left[0,0042 \le \mu_V-\mu_M \le 0,2009\right] \end{aligned}$$

Por lo tanto, en base a esta muestra, al ser el límite inferior de intervalo superior a cero, podemos decir con un 99% de confianza que la estatura media de los varones es superior a la estatura media de las mujeres.

En el caso en que las varianzas de las dos poblaciones sean desconocidas pero iguales (o pueda suponerse su igualdad), entonces, deberá utilizarse la distribución T de Student con *n+m-2* g.l. Para estimar la varianza común se usa el siguiente estadístico:

$$s_p^2=\frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}$$

Sean $\{X_1;X_2;...;X_n\}$ y $\{Y_1;Y_2;...;Y_m\}$ dos muestras aleatorias de tamaños n y m, respectivamente, provenientes de dos poblaciones Normales independientes con medias μ_X y μ_Y, y **varianzas desconocidas pero iguales**. Entonces:

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$

El IC para la diferencia de medias será:

$$C\left[\bar{x}-\bar{y}-t_{n+m-2\,;\,1-\alpha/2}\times s_p\sqrt{\frac{1}{n}+\frac{1}{m}} \le \mu_X-\mu_Y \le \bar{x}-\bar{y}+t_{n+m-2\,;\,1-\alpha/2}\times s_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right]=1-\alpha$$

Ejemplo 31

Consideremos el ejemplo anterior, y supongamos que se desconocen las varianzas poblacionales, pero pueden suponerse iguales. Para construir el intervalo, primero buscamos el valor $t_{30+20-2;0,995} = 2,6822$. Luego, estimamos la varianza común s_p^2:

$$\left.\begin{aligned} s_p^2 &= \frac{(n_V - 1)s_V^2 + (n_M - 1)s_M^2}{n_V + n_M - 2} \\ &= \frac{(20-1)\times 0,064^2 + (30-1)\times 0,09^2}{20+30-2} \\ &= 0,006464 \end{aligned}\right\} \Rightarrow \quad s_p = 0,0804$$

Remplazando en la expresión del intervalo, tenemos que:

$$\begin{aligned} 0,99 &= C\left[1,80-1,70-2,6822\times 0,0804 \le \mu_V - \mu_M \le 1,80-1,70+2,6822\times 0,0804\right] \\ &= C\left[-0,1131 \le \mu_V - \mu_M \le 0,3182\right] \end{aligned}$$

Vemos que, en este caso, al incluir valores negativos el intervalo, no se puede asegurar que la estatura media de los varones sea mayor a la estatura media de las mujeres. Por lo tanto, la afirmación del ejemplo anterior estaba basada exclusivamente en los supuestos respecto de las varianzas poblacionales.

4.11.2 Diferencia de proporciones

Al igual que cuando se trabaja con una única población, para estimar la diferencia de proporciones se utiliza la aproximación Normal.

Sean $\{A_1; A_2; ...; A_n\}$ y $\{B_1; B_2; ...; B_m\}$ dos muestras aleatorias de tamaños n y m, respectivamente, provenientes de dos poblaciones Bernoulli con parámetros p_A y p_B. Sean X e Y la cantidad de éxitos en cada una de ellas (es decir, $X = \sum_{i=1}^{n} A_i$ y $Y = \sum_{i=1}^{m} B_i$). Entonces, aproximadamente, tenemos que:

$$\frac{\left(\frac{X}{n} - \frac{Y}{m}\right) - (p_A - p_B)}{\sqrt{\frac{p_A(1-p_A)}{n} + \frac{p_B(1-p_B)}{m}}} \sim N(0;1)$$

Con esta distribución de muestreo, es inmediata la construcción del intervalo de confianza.

El intervalo de confianza para la diferencia de proporciones es:

$$C\left[x/n - y/m - z_{1-\alpha/2}\times s_{prop} \le p_A - p_B \le x/n - y/m + z_{1-\alpha/2}\times s_{prop}\right] = 1-\alpha$$

donde para el cálculo del desvío estándar se remplaza la proporción de cada población por sus respectivas medidas muestrales ($\bar{p}_A = x/n$ y $\bar{p}_B = y/m$), obteniendo:

$$s_{prop} = \sqrt{\frac{x(n-x)}{n^3} + \frac{y(m-y)}{m^3}}$$

Ejemplo 32

Consideremos el ejemplo de la Sección 10.5. Supongamos que en otra ciudad se encuesta a 300 personas, y resulta que 150 están a favor del candidato A. ¿Cuál es el IC al 95% para la diferencia entre las proporciones de ambas ciudades?

Primero, recordemos que en el ejemplo mencionado se obtuvo que 350 personas estaban a favor de A, sobre una muestra de 500: $x = 350$ y $n = 500$ ($x/n = 0,6$). A su vez, con los datos aquí presentados, tenemos que $y = 150$ y $m = 300$ ($y/m = 0,5$). Con esta información, podemos calcular el desvío estándar:

$$s_{prop} = \sqrt{\frac{x(n-x)}{n^3} + \frac{y(m-y)}{m^3}}$$
$$= \sqrt{\frac{300 \times 200}{500^3} + \frac{150 \times 150}{300^3}} \cong 0,03624$$

Finalmente, sabiendo que $z_{0,975} = 1,96$, podemos construir el intervalo:

$$0,95 = C\left[0,6 - 0,5 - 1,96 \times 0,03624 \leq p_1 - p_2 \leq 0,6 + 0,5 + 1,96 \times 0,03624\right]$$
$$= C\left[0,0290 \leq p_1 - p_2 \leq 0,1710\right]$$

Por lo tanto, en base a las muestras, con un 95% de confianza podemos afirmar que la diferencia entre la proporción de personas favorables al candidato A en las ciudades 1 y 2, está entre el 14,18% y el 25,82%. Esto evidencia el mayor apoyo en la ciudad 1, ya que el intervalo de la diferencia de proporciones no incluye al cero, por lo que, más allá de cuál sea la diferencia, podemos afirmar con un 95% de confianza que $p_1 > p_2$. Esta conclusión podría llevar al candidato a intensificar la campaña electoral en la ciudad 2 para lograr mayor apoyo.

En el capítulo siguiente se verán técnicas para comparar proporciones. Más precisamente, realizar un testeo para saber si una de las proporciones es superior a la otra.

4.11.3 Cociente de varianzas

En la Sección 5.2 del Capítulo 4, hemos visto que cuando se muestrean dos poblaciones Normales con medias desconocidas, se verifica que:

$$\frac{s_x^2/\sigma_x^2}{s_Y^2/\sigma_Y^2} \sim F(n-1; m-1)$$

Al igual que en los apartados anteriores, una vez conocida la distribución de muestreo, la construcción del intervalo es directa. En este caso, por no ser simétrica la distribución tenemos que:

$$P\left(F_{(n-1;m-1);\alpha/2} \leq \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} \leq F_{(n-1;m-1);1-\alpha/2} \right) = 1-\alpha$$

Si se reagrupa la expresión anterior y se realizan las estimaciones correspondientes de los estadísticos s_X^2 y s_Y^2, obtenemos el IC para el cociente de varianzas:

Si se toman dos muestras aleatorias, $\{X_1; X_2; ...; X_n\}$ y $\{Y_1; Y_2; ...; Y_m\}$, de dos poblaciones Normales con medias desconocidas, entonces el Intervalo de Confianza de un $100 \times (1-\alpha)\%$ es

$$C\left(\frac{s_X^2}{s_Y^2 F_{(n-1;m-1);1-\alpha/2}} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{s_X^2}{s_Y^2 F_{(n-1;m-1);\alpha/2}} \right) = 1-\alpha$$

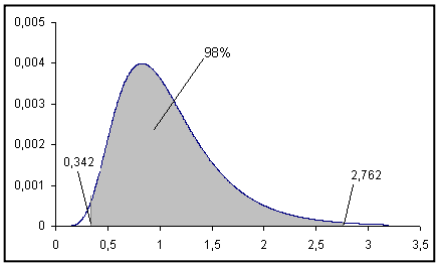
Ejemplo 33

En los Ejemplos 7 y 8 de este capítulo, se analizó la duración de las lamparitas producidas por una determinada empresa. Supongamos que la empresa está analizando la posibilidad de adquirir una nueva máquina, y le interesa especialmente que la duración de los productos sea más estable (es decir, que la varianza de la duración sea menor). Para decidir respecto de la compra, se toma una muestra de la producción de la nueva máquina. Las duraciones de las lamparitas se observan en la tabla, siendo la varianza muestral calculada con estas observaciones: $s_Y^2 = 1871$.

681	688	655	771	670
674	635	657	694	619
662	782	751	650	684
700	720	757	614	678
679	706	705	682	663

A su vez, recordamos de los ejemplos anteriores que con la muestra de la producción actual se obtuvo una varianza de $s_X^2 = 3600$.

Para realizar la comparación de la variabilidad de los dos métodos de producción se decide construir el intervalo para el cociente de varianzas con un 98% de confianza. Para determinar los límites, ya tenemos calculadas las varianzas muestrales, y nos resta determinar los valores de la variable F con 19 y 24 g.l. (ya que la primera muestra era de tamaño 20 y la segunda de tamaño 25).



Entonces, buscamos los cuantiles que acumulan 99% y 1%, de manera que en el centro quede el 98% deseado (ver figura): $F_{(19;24);0,99} = 2,762$ y $F_{(19;24);0,01} = 0,342$. Finalmente, utilizando la fórmula expuesta calculamos el intervalo:

$$0,98 = C\left(\frac{3600}{1871 \times 2,762} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{3600}{1871 \times 0,342}\right)$$
$$= C\left(0,697 \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq 5,629\right)$$

Para asegurar que la nueva máquina es menos variable que la actual, el intervalo debería encontrarse totalmente a la derecha del número uno (el límite inferior debería ser mayor a 1), ya que ello implicaría que la varianza actual es mayor que la nueva:

$$1 < LI \leq \frac{\sigma_X^2}{\sigma_Y^2} \qquad \Rightarrow \qquad 1 < \frac{\sigma_X^2}{\sigma_Y^2} \quad \Rightarrow \qquad \sigma_Y^2 < \sigma_X^2$$

Como el intervalo calculado con un 98% incluye al uno, no podemos asegurar que la varianza del procedimiento actual sea mayor a la varianza de la nueva máquina.

4.12 Tamaño Muestral y poblaciones Finitas

En este capítulo, hemos visto que el error estándar de la media muestral decrece a medida que el tamaño muestral aumenta, lo cual puede apreciarse claramente en su fórmula de cálculo: $\sigma_{\bar{X}} = \sigma / \sqrt{n}$. Por lo tanto, al aumentar el tamaño muestral, es más probable que una realización de la media muestral se encuentre más próxima a la media poblacional. Se podría decir que aumenta la "precisión" de $\bar{X}$ cuando aumenta n. Además, también hemos mencionado que la reducción del error estándar es decreciente, ya que el mismo varía inversamente con la raíz cuadrada de n. Por ejemplo, si tenemos un desvío de 5000 y un tamaño muestral de 40 (como en el Ejemplo 14). Si aumentamos 10 veces n, de 40 a 400, el error estándar de $\bar{X}$ se reduce a poco menos de un tercio del valor original, ya que:

$$\sigma_{\bar{X};40} = 5000/\sqrt{40} \cong 790,57; \qquad \text{y} \qquad \sigma_{\bar{X};400} = 5000/\sqrt{400} = 250$$

Por lo tanto, hay que ver si vale la pena aumentar el tamaño muestral, ya que se debe evaluar tanto el beneficio que produce la muestra mayor, como el costo que implicaría recolectarla.

La fórmula de cálculo utilizada hasta aquí para el error estándar de la media muestral, $\sigma_{\bar{X}} = \sigma / \sqrt{n}$, se refiere a los casos en que se muestrean poblaciones infinitas, o cuando se toman muestras

con reposición. Cuando se toman muestras sin reposición de poblaciones finitas se debe corregir el error estándar[50].

Cuando se trabaja con una **población finita**, y se realiza un **muestreo sin reposición**, entonces el **error estándar de $\bar{X}$ para poblaciones finitas**, $\sigma_{\bar{X};PF}$, es:

$$\sigma_{\bar{X};PF} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \sigma_{\bar{X}}\sqrt{\frac{N-n}{N-1}}$$

donde n y N, son el tamaño muestral y poblacional, respectivamente.

El factor que se multiplica a $\sigma_{\bar{X}}$ se denomina **multiplicador de población finita**:

$$\sqrt{\frac{N-n}{N-1}}$$

Recordemos que cuando se trabajaba con la distribución del estadístico $\bar{p}$, el desvío muestral era $\sqrt{p(p-1)}/\sqrt{n}$. Por lo tanto, el **error estándar de $\bar{p}$ para poblaciones finitas** es:

$$\sigma_{\bar{p};PF} = \sqrt{\frac{p(p-1)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Ejemplo 34

Considerando el Ejemplo 15 de la Sección 4, supongamos que la producción diaria es de 1000 artículos. Si se toma una muestra de 100 elementos, entonces:

$$\begin{aligned}\sigma_{\bar{p};PF} &= \sqrt{\frac{0,10\times 0,90}{100}}\sqrt{\frac{1000-100}{1000-1}} \\ &\cong 0,03\times 0,9492 \\ &\cong 0,0285\end{aligned}$$

Vemos que este desvío es ligeramente menor al del ejemplo mencionado, el cual era de $\sigma_{\bar{p}} = 0,03$, correspondiente al primer factor del cálculo previo. De este modo, la probabilidad de que haya menos de un 6% de artículos defectuosos es:

$$\begin{aligned}P(\bar{p} < 0,06) &= P\left(\frac{\bar{p}-p}{\sigma_{\bar{p};PF}} < \frac{0,06-0,10}{0,0285}\right) \\ &\cong P(Z < -1,4047) \\ &\cong 0,0801\end{aligned}$$

Vemos que la probabilidad del ejemplo anterior (9,12%) se reduce en más de un 1% debido al efecto del multiplicador, alcanzando solamente un 8%.

Notemos que si el tamaño poblacional es muy grande en relación al tamaño muestral, el multiplicador se aproxima a uno, por lo que no resulta necesario realizar la corrección. En general, si la muestra es menor al 5% de la población (es decir, si $n/N < 0,05$)[51], entonces no resulta necesario utilizar el multiplicador. Además, en el cálculo de $\sigma_{\bar{X};PF}$ aparece $\sqrt{n}$ dividiendo, por lo que el tamaño muestral en términos absolutos es el que disminuye el error estándar, sin importar qué porcentaje de la población represente.

Ejemplo 35

En el ejemplo anterior supusimos que la producción diaria (la "población") era de 1000 artículos. Si la producción fuera de 5000 artículos, el multiplicador sería muy cercano a 1:

[50] Para una justificación de esta corrección, véase el Capítulo 11 de Novales Cinca (1997).
[51] Algunos autores denominan "fracción de muestreo" al cociente n/N.

$$\sqrt{\frac{5000-100}{5000-1}} \cong 0,99$$

El desvío corregido sería $\sigma_{\bar{p};PF} = 0,03 \times 0,99 = 0,0297$, muy cercano al del desvío para poblaciones infinitas, el cual era de 0,3. En este caso, $n/N = 100/5000 = 0,02$, por lo que podría no utilizarse el factor de corrección.

Cuando se toma una muestra de un tamaño determinado y se estima el intervalo de confianza, la precisión de la estimación está relacionada con la amplitud del intervalo, y es incierta *a priori*. En ocasiones, se desea obtener cierto grado de precisión, y se calcula cuál es el tamaño muestral necesario para ello. Es decir, primero se fija el nivel de precisión deseado, y luego, en base el mismo, se calcula el tamaño muestral necesario para lograrlo.

Para determinar el tamaño muestral necesario para que el error sea inferior a ε, se utiliza el teorema de Chebyshev[52].

Sea X una variable aleatoria con $E(X) = \mu$ y $Var(X) = \sigma^2$, ambos finitos. Entonces, para cualquier $k \geq 1$ se verifica que:

$$P(|X - \mu| \leq k\sigma) \geq 1 - 1/k^2$$

Este teorema puede aplicarse de manera general a cualquier distribución de probabilidades para determinar el tamaño muestral necesario para que la estimación tenga una precisión determinada.

Ejemplo 36

Supongamos que deseamos estimar el ingreso medio de una población determinada con una precisión de 200, y sabemos que la varianza es $\sigma^2 = 90000$ (el desvío correspondiente es $\sigma = 300$). Además, queremos que nuestra estimación tenga por lo menos un 99% de confianza. ¿Cuál es el tamaño muestra necesario cumplir con los requisitos?

De acuerdo con el teorema anterior, tenemos que para cualquier variable:

$$P(|X - \mu| \leq k\sigma) \geq 1 - 1/k^2$$

E este caso, deseamos estimar la media poblacional, por lo que nuestra variable de interés será su mejor estimador: $\bar{X}_n$. El desvío estándar de la media muestral es $\sigma/\sqrt{n}$. Por lo tanto, en este caso particular tendremos que:

$$P\left(|\bar{X}_n - \mu| \leq k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - 1/k^2 \qquad (1)$$

Como deseamos que el nivel de confianza supere el 99%, igualamos este valor al miembro derecho de la desigualdad anterior, obteniendo:

$$0,99 = 1 - 1/k^2 \quad \Rightarrow k = 10$$

Luego, como queremos que la precisión sea de 200, debemos lograr que la distancia entre el estimador y el parámetro sea menor a dicho valor. Utilizando la expresión (1), sabemos que dicha distancia está dada por $k\sigma/\sqrt{n}$, por lo tanto:

$$k\sigma/\sqrt{n} = 200$$

Por último, como conocemos que $k = 10$ y $\sigma = 300$, podemos despejar n:

$$10 \times \frac{300}{\sqrt{n}} = 200 \quad \Rightarrow \sqrt{n} = 10 \times \frac{300}{200} = 15 \quad \Rightarrow \quad n = 225$$

[52] En el Capítulo 3, Sección 2.5, se expuso la regla de Bienaymé-Chebyshev, referida al análisis de datos. Aquí veremos un teorema relacionado con variables aleatorias. El lector podrá observar la analogía entre ambos.

Finalmente, podemos afirmar que si sabemos que el desvío estándar es 300 y que deseamos un nivel de confianza de por lo menos el 99%, debemos extraer una muestra de tamaño 225.

El ejemplo anterior puede generalizarse para cualquier distribución.

Si deseamos una confianza del $100\times(1-\alpha)\%$, entonces $\alpha = 1/k^2$ y, por lo tanto, $k = 1/\sqrt{\alpha}$. Además, si queremos que la diferencia entre $\bar{X}_n$ y μ sea menor a ε, tenemos que $\varepsilon = k\times\sigma/\sqrt{n}$. Reemplazando en esta última expresión el valor de k recién hallado, se obtiene que $\varepsilon = \sigma/\sqrt{n\alpha}$. Finalmente, despejando n de la ecuación anterior obtenemos que el tamaño muestral necesario.

Cuando la varianza poblacional, σ^2, es conocida y se desea estimar la media poblacional con una precisión de ε y una confianza de *al menos* $100\times(1-\alpha)\%$, el tamaño muestral necesario es:

$$n_{\text{Pobl. desc.}} = \frac{\sigma^2}{\varepsilon^2\times\alpha}$$

El teorema de Chebyshev establece una desigualdad para el caso en que se desconozca la distribución de probabilidades de la variable aleatoria y, a través del mismo, se establece el tamaño necesario para obtener un nivel de confianza igual *o superior* al deseado. En caso de que la distribución sea conocida, no será necesario recurrir al teorema, y se podrá establecer de manera exacta el tamaño muestral.

Ejemplo 37

Consideremos el ejemplo anterior, pero supongamos que la distribución del ingreso es Normal. En este caso, sabemos que el valor de una Normal Estándar que acumula $1-\alpha/2 = 0,995$ es $z_{0,995} = 2,5758$ y, por lo tanto:

$$P\left(-2,5758 \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le 2,5758\right) = 0,99$$

Re-expresando la ecuación anterior se obtiene:

$$P\left(\left|\bar{X}_n - \mu\right| \le 2,5758\times\sigma/\sqrt{n}\right) = 0,99$$

Finalmente, sabiendo que $\sigma = 300$ y que se desea una precisión de 200, tenemos que:

$$\begin{aligned} 2,5758\times 300/\sqrt{n} = 200 \quad &\Rightarrow \sqrt{n} = 2,5758\times 300/200 \cong 3,8637 \\ &\Rightarrow n \cong 14,9285 \end{aligned}$$

Redondeando hacia arriba el valor hallado (para garantizar el nivel de confianza deseado), obtenemos que debiera observarse el ingreso de 15 individuos para que, con un 99% de confianza, la media muestral no difiera de la media poblacional en más de 200 (sabiendo que la varianza poblacional es 300).

El ejemplo puede generalizarse fácilmente. Si se posee una población Normal, se sabe que:

$$P\left(-z_{1-\alpha/2} \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}\right) = 1-\alpha$$

Lo cual puede expresarse como:

$$P\left(\left|\bar{X}_n - \mu\right| \le z_{1-\alpha/2}\times\sigma/\sqrt{n}\right) = 1-\alpha$$

Luego, igualando el nivel de error máximo permitido a la diferencia que hay entre la media poblacional y la muestral, podemos obtener el tamaño muestral necesario. Es decir, que debe despejarse n de la siguiente ecuación:

$$\varepsilon = z_{1-\alpha/2}\times\sigma/\sqrt{n}$$

Cuando se desea estimar la media de una población Normal con varianza σ^2 (conocida) con una precisión de ε y una confianza de $100\times(1-\alpha)\%$, el tamaño muestral necesario es:

$$n_{\text{Normal}} = \left(\frac{z_{1-\alpha/2} \times \sigma}{\varepsilon} \right)^2$$

En el ejemplo anterior, teníamos que $z_{0,995} = 2,5758$, $\varepsilon = 200$ y $\sigma = 300$. Por lo tanto, utilizando directamente la fórmula expuesta, podemos obtener que el tamaño muestral necesario es: $n = (2,5758 \times 300/200)^2 \cong 14,9285$.

Cuando se estima la proporción poblacional, tal como se vio en la Sección 4.5, suele utilizarse la aproximación Normal para la población Binomial. En este caso, el estimador a utilizar para realizar inferencias es la proporción muestral $\bar{p} = X/n$, cuyo desvío estándar es $\sqrt{p(1-p)/n}$. En la mencionada sección, se vio que:

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{p} - p}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2} \right) = 1-\alpha$$

Esta expresión puede presentarse como:

$$P\left(|\bar{p} - p| \leq z_{1-\alpha/2} \times \sqrt{p(1-p)/n} \right) = 1-\alpha$$

Finalmente, procediendo de manera análoga al caso anterior en que estimábamos la media poblacional, tenemos que despejar n de la siguiente ecuación:

$$\varepsilon = z_{1-\alpha/2} \times \sqrt{p(1-p)/n}$$

Sin embargo, en este caso, generalmente desconocemos la varianza poblacional, ya que se desconoce el valor de p. Por ello, para asegurarnos el nivel de confianza deseado, suele utilizarse el valor del parámetro que maximiza el desvío estándar, es decir, que hacemos $p = 1/2$ en la expresión anterior, obteniendo: $\varepsilon = z_{1-\alpha/2} \times \sqrt{1/4/n}$.

Cuando se desea estimar la proporción poblacional con una precisión de ε y una confianza de $100\times(1-\alpha)\%$, el tamaño muestral necesario es:

$$n_{\text{Proporcion}} = \left(\frac{z_{1-\alpha/2}}{\varepsilon} \right)^2 \times p(1-p)$$

Si no se cuenta con ninguna información relacionada con el posible valor de p, éste se remplaza por $p = 1/2$ en la expresión anterior.

Ejemplo 38

Supongamos que un fabricante necesita controlar la calidad de sus productos y desea estimar la proporción de artículos defectuosos con una precisión del 95% y una precisión del 5%. Si desconoce la proporción de artículos fallados ¿cuál sería el tamaño muestral necesario? ¿Cuál sería el tamaño muestral si *a priori* el fabricante estima que hay un 10% de artículos defectuosos?

La precisión es $\varepsilon = 0,05$ y el nivel de confianza del 95% implica un valor de $z_{0,975} = 1,9600$.

Para la primera pregunta, al desconocer la proporción poblacional, utilizamos $p = 1/2$, y obtenemos que:

$$n = \left(\frac{1,96}{0,05} \right)^2 \times 0,5^2 \cong 384,15$$

Por lo tanto, para cumplir con los requisitos se deberían analizar 385 artículos.

En el segundo caso, si se estima *a priori* que la proporción de artículos defectuosos es del 10%, se obtiene que:

$$n = \left(\frac{1,96}{0,05}\right)^2 \times 0,1 \times 0,9 \cong 138,29$$

Es decir, que en este caso la muestra necesaria sería de solamente 139 artículos.

Lógicamente, en el último caso del ejemplo anterior, después del proceso de muestreo y estimación habrá que testear si la creencia *a priori* es justificada y coherente con lo observado en la muestra. Este tema será estudiado en el capítulo siguiente.

4.13 Apéndice: Demostraciones

Esperanza y varianza de la media muestral

La media muestral es:

$$\overline{X} = \frac{1}{n}\left(X_1 + X_2 + ... + X_n\right)$$

Donde las X_i son variables aleatorias iid, con media μ y varianza σ^2 cada una.

Si calculamos la **esperanza**, tenemos que:

$$\begin{aligned} E\left(\overline{X}\right) &= E\left[\frac{1}{n}\left(X_1 + X_2 + ... + X_n\right)\right] \\ &= \frac{1}{n} E\left(X_1 + X_2 + ... + X_n\right) \\ &= \frac{1}{n}\left[E\left(X_1\right) + E\left(X_2\right) + ... + E\left(X_n\right)\right] \\ &= \frac{1}{n}\left[\underbrace{\mu + \mu + ... + \mu}_{n \text{ veces}}\right] \\ &= \frac{n\mu}{n} \\ &= \mu \end{aligned}$$

En el primer paso, hemos utilizado la propiedad "la esperanza de una constante por una variable aleatoria, es la constante por la esperanza de la variable". En el segundo paso, se utilizó "la esperanza de una suma de variables aleatorias es la suma de las esperanzas". Y en el tercer paso se utilizó que las variables tienen idéntica distribución, cada una con esperanza μ.

Si calculamos la **varianza**, tenemos que:

$$
\begin{aligned}
Var\left(\bar{X}\right) &= Var\left[\frac{1}{n}\left(X_1 + X_2 + ... + X_n\right)\right] \\
&= \frac{1}{n^2} Var\left(X_1 + X_2 + ... + X_n\right) \\
&= \frac{1}{n^2}\left[Var\left(X_1\right) + Var\left(X_2\right) + ... + Var\left(X_n\right)\right] \\
&= \frac{1}{n^2}\left[\underbrace{\sigma^2 + \sigma^2 + ... + \sigma^2}_{n \text{ veces}}\right] \\
&= \frac{n\sigma^2}{n^2} \\
&= \sigma^2 / n
\end{aligned}
$$

En el primer paso hemos utilizado la propiedad "la varianza de una constante por una variable aleatoria, es la constante elevada al cuadrado por la varianza de la variable". En el segundo paso, se utilizó "la varianza de una suma de variables aleatorias *independientes* es la suma de las varianzas". Y en el tercer paso se utilizó que las variables tienen idéntica distribución, cada una con varianza σ^2. Nótese la importancia que tiene la independencia de las variables en este caso.

Distribución t de Student: $\frac{\bar{X} - \mu}{s/\sqrt{n}}$

(1) En la sección 3.2, se vio que cuando se muestrea una población Normal, la distribución $Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ es Normal Estándar.

(2) En la sección 5.1 vimos que $W = (n-1)s^2/\sigma^2$ tiene una distribución Chi-cuadrado con $n-1$ g.l.

(3) Si dividimos a la variable W por sus grados de libertad y luego le tomamos raíz cuadrada, obtenemos una variable que es una "la raíz cuadrada de una Chi-cuadrado dividida por sus grados de libertad":

$$\sqrt{\frac{W}{g.l.(W)}} = \sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{n-1}} = \sqrt{\frac{s^2}{\sigma^2}} = \frac{s}{\sigma}$$

(4) Si dividimos la variable Y de (1), por la variable obtenida en (3), habremos dividido una Normal Estándar por la raíz de una Chi-cuadrado dividida por sus grados de libertad, la cual, según el teorema del final de la sección 6.1, sigue una distribución t de student, cuyos grados de libertad se corresponden con los de la Chi-cuadrado, que en este caso son $n-1$:

$$T = \frac{Y}{\sqrt{\frac{W}{g.l.(W)}}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{s/\sigma} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Al final de la Sección 2.1 se afirmó que:

La proporción muestral es un estimador insesgado de la proporción poblacional:

$$\bar{p} = \frac{X}{n}$$

El hecho de que sea insesgado quiere decir que $E(\bar{p}) = p$ y es lo que se demostrará a continuación.

Sea $\{B_1; B_2; ...; B_n\}$ una muestra aleatoria de variables Bernoulli iid. Es decir, que cada una de ellas puede tomar el valor 1 con probabilidad p y el valor 0 con probabilidad $1-p$. Por lo tanto, la esperanza de cada una de ellas es:

$$E(B_i) = 1 \times p + 0 \times (1-p) = p$$

Si X es la cantidad de "éxitos" de la muestra, tenemos que $X = B_1 + B_2 + ... + B_n$. Al tomar esperanza a esta variable obtenemos:

$$\begin{aligned} E(X) &= E(B_1 + B_2 + ... + B_n) \\ &= E(B_1) + E(B_2) + ... + E(B_n) \\ &= \underbrace{p + p + ... + p}_{n \text{ veces}} \\ &= n \times p \end{aligned}$$

Finalmente, utilizando esta expresión podemos calcular la esperanza de $\bar{p}$, y observar que efectivamente coincide con el parámetro p:

$$\begin{aligned} E(\bar{p}) &= E\left(\frac{X}{n}\right) \\ &= \frac{E(X)}{n} \\ &= \frac{n \times p}{n} \\ &= p \end{aligned}$$

'La estimación es una de las aplicaciones más importantes de la estadística moderna. Si se detiene a pensar un poco, todo el mundo realiza, permanentemente, estimaciones. Por ejemplo, alguien que vive en Olivos y tiene una reunión en el centro, saldrá una hora antes porque "estima" que tardará ese tiempo en llegar; si está por ir al supermercado a realizar una compra y toma $100, estima que gastará esa suma o menos; un gerente que contrata un nuevo empleado estima que los beneficios que aportará a la empresa serán mayores que el costo salarial del mismo; cuando el jefe de un departamento realiza un presupuesto estima que su sector gastará ese dinero en el año próximo; etc. Como se dará cuenta, podemos escribir páginas completas de las estimaciones que uno realiza diariamente casi sin darse cuenta. Además, decimos que todas estas son *estimaciones* porque "a seguro lo llevaron preso"[53]: si hay problemas de tránsito demorará más de una hora en ir desde Olivos hasta el centro; si encuentra una oferta muy buena quizás desee gastar más de $100 en el supermercado; si la selección no fue buena el empleado podría traer más dolores de cabeza que beneficios; por diversos motivos podría gastarse más que lo presupuestado (inflación, crecimiento de la empresa o del departamento, etc.); etc. Podemos ver, en estos casos, la importancia de que las estimaciones sean buenas, dependiendo del contexto: si demora más de una hora podría llegar tarde a la función de teatro; si llevó sólo $100 podría perderse la oferta; si seleccionó mal al empleado perjudicaría a toda la empresa; si el jefe presupuestó mal y no hay flexibilidad, podría dejar a su sector sin recursos; etc.

[53] Dicho popular que indica que no hay nada seguro y, por lo tanto, siempre hay que tomar ciertos recaudos.

En algunos casos de los mencionados anteriormente, las estimaciones se realizan intuitivamente (“a ojo” o mediante el método de los “cinco dedos oscilantes”). Sin embargo, en otros casos, se requiere de un mayor análisis y rigurosidad científica. Es en estos casos donde la “Inferencia Estadística” juega su rol fundamental. Si bien ya hemos definido a esta rama de la estadística, repasaremos brevemente sus características. La Estadística Inferencial es la rama que utiliza conceptos relacionados con la Teoría de la Probabilidad para la toma de decisiones en situaciones de incertidumbre. Las dos principales aristas de esta rama son la Estimación, que se tratará en este capítulo, y las Pruebas de Hipótesis, que serán analizadas en el capítulo siguiente.

En otras palabras, podemos decir que la estadística aplicada pretende realizar afirmaciones o sacar conclusiones en relación a ciertas características de una población determinada sobre la base de observaciones realizadas en una muestra de la misma y que este proceso se determina “Inferencia Estadística”. Por ejemplo, en los siguientes, casos se requiere el análisis de muestras para inferir comportamientos relacionados a las poblaciones:

- En algunos procesos de control de calidad, las pruebas que se realizan implican la utilización del producto hasta que deje de funcionar. En estos casos, no se analiza toda la producción ¡porque no quedaría nada para vender!
- En los análisis de mercado no se encuesta a toda la población, porque ello requeriría un gasto demasiado elevado no justificado.

En este capítulo, analizaremos la manera de estimar el valor de alguna característica poblacional mediante el análisis de una muestra. Por ejemplo,

- Mediante el estudio de 100 lamparitas (una muestra), se tratará de estimar la duración media de toda la producción.
- A través de la consulta a 200 personas, se tratará de determinar las preferencias de todos los consumidores de una región.

Como se trabaja con muestras, habrá que determinar el nivel de riesgo que poseen las conclusiones que se obtengan, ya que si no se analiza la población completa no se puede tener certeza en relación a sus características. Los métodos de estimación y la determinación del nivel de riesgo son analizados a continuación.

5 Pruebas de Hipótesis

Dario Bacchini
Lara Vazquez
Andrea Lepera

En general, en casi todas las disciplinas, los investigadores formulan hipótesis referidas al comportamiento de los problemas que están estudiando y luego, a través de experimentos, tratan de comprobar si sus hipótesis son correctas. Sobre esta idea se apoyan las Pruebas de Hipótesis Estadísticas.

Cuando se desea hacer una prueba para validar una Hipótesis Estadística se debe trabajar con muestras aleatorias (ver Capítulo 4). La información suministrada por la muestra puede rechazar la hipótesis formulada o no rechazarla[54]. La decisión en relación a si los datos apoyan o no la hipótesis formulada, se toma con bases probabilísticas (Canavos, 1997).

El testeo que se realiza, en general, tiene que ver con parámetros relacionados con la población bajo estudio o con características numéricas descriptivas de la misma.

Consideremos algunos ejemplos para clarificar el concepto:

- El responsable de producción de una empresa puede estar interesado en saber si el promedio de tiempo que demora un producto en la línea de montaje es de 15 minutos. En este caso, la hipótesis que se desea probar o testear es que "el tiempo promedio es 15 minutos".

- A una fundación de salud que estudia las causas del cáncer pulmonar puede interesarle testear si la proporción de fumadores es de del 50%. En este caso la hipótesis sería "la proporción de fumadores es 50%".

- El responsable de control de calidad de una fábrica podría testear si la cantidad de artículos defectuosos en un día de producción es como máximo 50 ó si la proporción de artículos defectuosos es inferior al 10%. Las hipótesis serían "la cantidad promedio de artículos defectuosos es inferior o igual a 50" ó "la proporción de artículos defectuosos es menor o igual a 0,10".

En estos ejemplos se han enunciado hipótesis "en palabras". En la sección siguiente, se expondrá la manera de formularla en términos estadísticos.

A continuación, se exponen algunos conceptos generales relacionados con el testeo de hipótesis, luego se desarrolla la metodología para poner en práctica las pruebas en distintos casos.

5.1 Conceptos Generales del Testeo de Hipótesis

Las pruebas de hipótesis se realizan para testear si una creencia que se tiene respecto de algún parámetro poblacional tiene sustento en base a datos muestrales. A continuación, se presenta la manera en que se formulan las hipótesis en términos estadísticos y luego, se presentan los conceptos básicos sobre los cuales se basan las técnicas de testeo que se describen en las secciones posteriores.

5.1.1 Formulación de las Hipótesis Nula y Alternativa

La hipótesis que se formula y que se cree verdadera *a priori* es la que se desea testear. Ésta se denomina **Hipótesis Nula** y se denota por H_0. Las hipótesis nulas de los ejemplos mencionados al inicio del capítulo son:

- Encargado de producción: $H_0 : \mu = 15\min.$

- Fundación que estudia el cáncer pulmonar: $H_0 : p = 0,50$

- Responsable de Control de Calidad: $H_0 : \mu \leq 50$ ó $H_0 : p \leq 0,10$

Para rechazar o no la hipótesis nula, se deberán utilizar estimaciones de la característica a testear obtenidas de una muestra. Cabe destacar aquí que el interés no recae directamente en la estimación del parámetro (la media o la proporción en los ejemplos mencionados), sino que habiendo realizado una hipótesis sobre el valor del parámetro se desea comprobar la validez de la misma.

[54] Hemos expresado intencionalmente "no rechazar" en lugar de "aceptar". El motivo quedará claro más adelante.

La decisión respecto del rechazo o no de la Hipótesis Nula se realiza sobre la base del valor de la estimación del parámetro de interés obtenida con la muestra aleatoria.

El estimador de la media poblacional es $\bar{X}$ y el de la proporción es $\bar{p}$ (ver Capítulo 6). Siguiendo con los ejemplos mencionados:

- Si el valor particular $\bar{x}$ obtenido con una muestra es "muy lejano" a 15min., entonces, se deberá rechazar la hipótesis realizada y considerar que el tiempo promedio de armado no es igual al valor supuesto.

- Si la proporción de fumadores de la muestra es "muy distinto" del 50%, deberá rechazarse la hipótesis nula.

- Si el promedio muestral de artículos defectuosos es "muy superior" a 50 se rechazará la hipótesis nula. La segunda hipótesis se rechazaría si la proporción muestral de artículos defectuosos es "muy superior" a 0,10.

Como habrá notado, en los párrafos previos se hace hincapié en la evidencia muestral para **rechazar** la hipótesis. En este aspecto resulta interesante la analogía con el principio judicial de que una persona acusada de un delito "es inocente hasta que se pruebe lo contrario". En este caso, la evidencia en el juicio debe ser contundente para estar seguros de que no se está culpando a alguien inocente, ya que se considera más grave condenar a alguien inocente que dejar libre a alguien culpable. Podemos decir que la justicia se basa en la hipótesis nula de que el acusado es inocente, y testea dicha hipótesis para comprobar la culpabilidad. De esta manera, por más que haya indicios de culpabilidad, las pruebas tienen que ser determinantes para condenar al acusado.

En el caso de las hipótesis estadísticas, se deben hallar pruebas suficientes para asegurar que la Hipótesis Nula es falsa (culpable). En caso contrario, no podremos "condenarla" por ser falsa, por más que haya indicios de su falsedad. Por ello, al inicio del capítulo y en lo que sigue, utilizamos la expresión "no se rechaza" la hipótesis nula, en lugar de "se acepta". No podemos decir que es culpable en base a la evidencia, pero tampoco podemos afirmar que es inocente.

En base a lo mencionado, queda claro el porqué se expresó que la característica muestral observada debe estar "muy lejos" de valor hipotético para rechazar la hipótesis nula: las pruebas deben ser contundentes. Si no podemos rechazar las hipótesis nulas, tampoco podemos afirmar con certeza que la misma sea cierta, simplemente no podemos afirmar que es falsa.

Para rechazar o no la Hipótesis Nula, debe crearse una **regla de decisión**. Por ejemplo, si el jefe de producción en base a su muestra obtiene que $\bar{X} < 10$ ó $\bar{X} > 20$, rechazará la hipótesis $H_0 : \mu = 15\text{min}$. Si los miembros de la fundación hallan que $\bar{p} < 0{,}30$ ó $\bar{p} > 0{,}80$ se rechazará $H_0 : p = 0{,}50$. Finalmente, si en control de calidad encuentran que en la muestra $\bar{X} > 65$ se rechazará $H_0 : \mu \leq 50$, mientras que si $\bar{p} > 0{,}14$ se rechazará $H_0 : p \leq 0{,}10$.

Los valores de los estimadores que limitan la región de rechazo se denominan **Valores Críticos**. En el caso de producción, los valores críticos serían de 10 y 20 minutos, en el caso de la proporción de fumadores serían 30% y 80% y, en el caso de control de calidad, sería 65 (para testear la media) y 14% (para testear la proporción). La determinación de estos valores está basada en la distribución de muestreo de los estimadores (Capítulo 4) y está íntimamente relacionada con los intervalos de confianza (Capítulo 6). La metodología para calcularlos será analizada en detalle más adelante. Todos los valores para los cuales se rechaza la Hipótesis Nula constituyen la **Región (o Zona) Crítica (o de Rechazo)** de la prueba o test.

Para establecer de manera clara la regla de decisión, se formula una **Hipótesis Alternativa**, que implica la negación de la nula y se denota por H_1. Para el caso de la línea de producción, la hipótesis alternativa sería $H_1 : \mu \neq 15$, para el caso de la proporción de fumadores es: $H_1 : p \neq 0{,}50$, mientras que para el control de calidad tendríamos $H_1 : \mu > 50$ ó $H_1 : p > 0{,}10$.

La formulación de la Hipótesis Alternativa puede realizarse de distintas maneras en relación a lo que se quiere testear. Supongamos que deseamos testear el valor del parámetro φ suponiendo en la hipótesis nula un valor de φ_0, entonces tenemos tres posibles formulaciones para la hipótesis alternativa:

Prueba Bilateral: la hipótesis alternativa tendrá la forma $H_1 : \varphi \neq \varphi_0$. En este caso, cualquier desviación respecto del valor hipotético es relevante, ya sea por exceso o por defecto (en los ejemplos de la línea de producción o la fundación, teníamos este caso).

Prueba Unilateral Superior (o derecha): Si estamos preocupados porque el valor real del parámetro φ sea superior al (esté a la derecha del) valor hipotético φ_0, entonces, la formulación de la hipótesis alternativa sería: $H_1 : \varphi > \varphi_0$ (en el ejemplo de control de calidad, utilizamos esta formulación).

Prueba Unilateral Inferior (o izquierda): Si la preocupación radica en cuanto a si el parámetro es menor al (está a la izquierda del) valor hipotético, entonces, la hipótesis alternativa será $H_1 : \varphi < \varphi_0$.

De acuerdo con la formulación de la hipótesis alternativa, la región crítica estará a la derecha, izquierda o ambos lados del valor φ_0.

En base a lo expuesto en este apartado, podemos definir formalmente una prueba de hipótesis de la siguiente manera.

Una **Prueba de Hipótesis Estadística** es una regla que permite decidir, en base a una muestra aleatoria, si se rechaza una Hipótesis (Nula) relacionada con una característica de la población.

5.1.2 Errores de Tipo I y II

Obviamente, cuando se toma una decisión se pueden cometer errores. Continuando con la analogía judicial, se pueden cometer básicamente dos errores: condenar a un inocente o dejar libre a un culpable.

En las pruebas estadísticas también se pueden cometer dos errores: rechazar una Hipótesis Nula cuando es verdadera o no rechazarla cuando es falsa. Estos errores se llaman Errores de Tipo I y II, respectivamente. La probabilidad de cometer el error de tipo I se denota con la letra α, mientras que la probabilidad de cometer el error de tipo II es β.

El **Error de Tipo I** se comete cuando se rechaza una Hipótesis Nula verdadera. La probabilidad de cometerlo es:

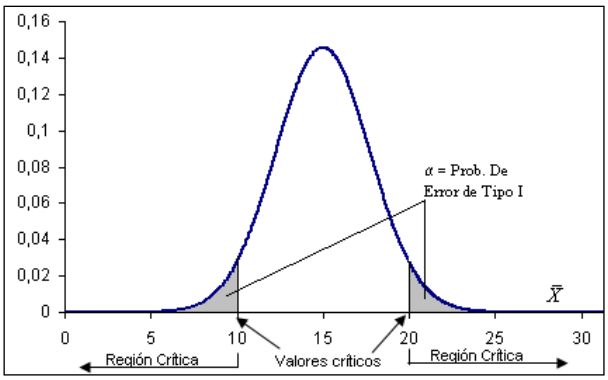
$$P(\text{Error de Tipo I}) = P(\text{rechazar } H_0 \,|\, H_0 \text{ es verdadera}) = \alpha$$

Ejemplo 1

Considere el caso del testeo de la hipótesis de que el tiempo medio en la línea de montaje de cada producto es 15 minutos, y la regla de decisión es que se rechazará la hipótesis si la media muestral es menor a 10 ó mayor a 20. Las hipótesis son:

$$H_0 : \mu = 15$$
$$H_1 : \mu \neq 15$$

Bajo ciertos supuestos, el estimador $\bar{X}$ tiene distribución Normal (ver Capítulo 4), y si asumimos que H_0 es verdadera su media será 15. En la siguiente figura, se muestra el gráfico de la distribución de $\bar{X}$ bajo H_0 (es decir, asumiendo que la media es 15), los valores críticos de 10 y 20, la región crítica y la probabilidad de cometer el error de tipo 1.

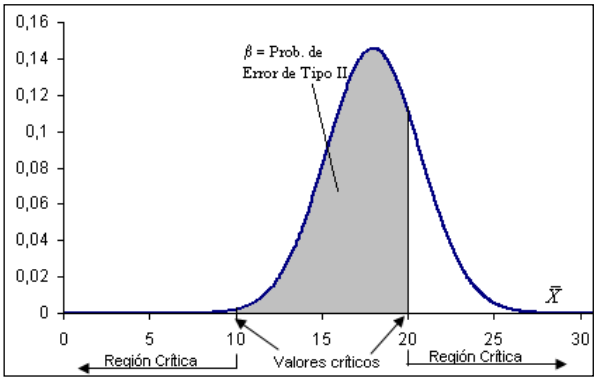


El **Error de Tipo II** se comete cuando no se rechaza una Hipótesis Nula falsa. La probabilidad de cometerlo es:

$$P(\text{Error de Tipo II}) = P(\text{No rechazar } H_0 \,|\, H_0 \text{ es falsa})$$
$$= \beta$$

Ejemplo 2

Considere el ejemplo anterior, y asuma que la verdadera media es $\mu = 18$, con lo que la hipótesis alternativa es verdadera ($\mu \neq 15$). Asumiendo esto, en la siguiente figura, se grafica la distribución de $\bar{X}$ bajo H_1 (es decir, asumiendo que la media es 18 –distinta a 15-), los valores críticos de 10 y 20, la región crítica y la probabilidad de cometer el error de tipo 2 (se recuerda que no se rechaza si $\bar{X} > 10$ ó $\bar{X} < 20$).



Cabe señalar que la probabilidad β de cometer error de tipo II depende, en este caso del valor 18 que toma la verdadera media poblacional. Si la media poblacional verdadera fuera, por ejemplo 30, $\beta = \beta(30)$ sería menor (observar que la zona entre 10 y 20 quedaría más en "la cola" de la distribución. Por lo tanto β depende del valor de la media poblacional verdadera, por lo cual es una función $\beta(\mu)$. En el siguiente ejemplo se analizará esto más en detalle.

Lógicamente, uno espera realizar una prueba de hipótesis con la menor probabilidad posible de cometer un error. Es decir, reducir al mínimo los valores de las probabilidades de Error de Tipo I y II. Sin embargo, esto no es posible ya que los dos errores interactúan entre sí. Si se reduce α mucho, entonces, necesariamente aumentará el valor de β . En el siguiente ejemplo, se ilustra la interacción entre los dos tipos de errores.

Ejemplo 3

Consideremos una prueba de hipótesis unilateral donde las hipótesis son:

$$H_0 : \mu = 30$$
$$H_1 : \mu < 30$$

Supongamos que el estadístico de prueba es la media muestral $\bar{X}$, que la muestra aleatoria será de tamaño 15 y que la varianza poblacional es 2.

Antes de analizar la muestra se debe determinar el valor crítico, para lo cual se proponen tres valores: $VC_1 = 28,50$, $VC_2 = 29,00$ y $VC_3 = 29,30$.

Además, se desea que el testeo tenga una probabilidad de error de tipo I no mayor al 5% ($\alpha \leq 0,05$).

En primer lugar, calculemos los errores de tipo I de cada uno de los valores críticos propuestos:

$$\alpha = P\left(\text{Rechazar } H_0 \mid H_0 \text{ es verdadera}\right)$$

Si el valor crítico es 28,5, entonces[55]:

$$\begin{aligned} P\left(\bar{X} < 28,5 \mid \mu = 30\right) &= P\left(\frac{\bar{X} - 30}{2/\sqrt{15}} < \frac{28,5 - 30}{2/\sqrt{15}}\right) \\ &= P\left(z < -2,905\right) \\ &= 0,002 \end{aligned}$$

En los otros dos casos tenemos:

$$\begin{aligned} P\left(\bar{X} < 29 \mid \mu = 30\right) &= P\left(z < -1,936\right) \\ &= 0,026 \end{aligned}$$

$$\begin{aligned} P\left(\bar{X} < 29,3 \mid \mu = 30\right) &= P\left(z < -1,356\right) \\ &= 0,088 \end{aligned}$$

Dado que con el tercer valor crítico propuesto se obtiene un valor de probabilidad de error de tipo I mayor al tolerado ($\alpha_3 = 0,088 > 0,05$), se descarta este valor crítico.

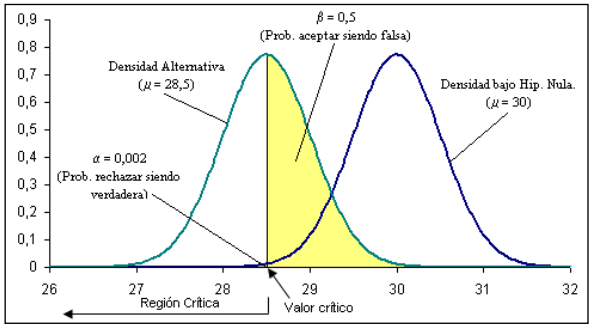
Podemos observar que $\alpha_1 < \alpha_2$, pero ambos cumplen con el máximo tolerado.

Analicemos ahora el error de tipo II. Para poder calcula β, se debe especificar un valor particular de $\mu < 30$. Por ejemplo, consideremos $\mu_1 = 28,5 < 30$, y calculemos:

$$\beta = P\left(\text{No Rechazar } H_0 \mid H_0 \text{ es falsa}\right)$$

Para el $VC_1 = 28,5$ tenemos que (ver figura):

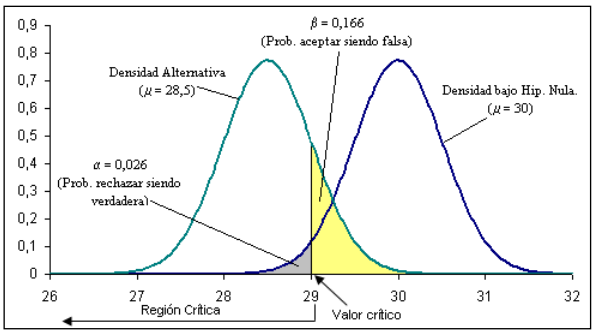
$$\begin{aligned} P\left(\bar{X} \geq 28,5 \mid \mu = 28,5\right) &= P\left(z \geq 0\right) \\ &= 0,500 \end{aligned}$$



Para el $VC_2 = 29$, el error de tipo II es (ver figura):

$$\begin{aligned} P\left(\bar{X} \geq 29 \mid \mu = 28,5\right) &= P\left(z \geq 0,9682\right) \\ &= 0,166 \end{aligned}$$

[55] Recordamos que $\bar{X}$ se distribuye normalmente con media μ y desvío $\sigma/\sqrt{n}$.



Entonces, para $\mu = 28{,}5$, $\beta_2 < \beta_1$. Obviamente, el valor de β dependerá del valor que se le dé a μ_1 (el valor de la media bajo la hipótesis alternativa). En la siguiente tabla, se ilustra el valor de β para los dos valores críticos considerados y para distintos valores de μ_1, lógicamente menores a $\mu_0 = 30$.

VC	β (28)	β (28.25)	β (28.5)	β (28.75)	β (29)	β (29.25)	β (29.5)
28.500	0.166	0.314	0.500	0.686	0.834	0.927	0.974
29.000	0.026	0.073	0.166	0.314	0.500	0.686	0.834

De esta manera, al disminuir el valor crítico baja la Probabilidad de Error de Tipo I, α, pero aumenta la Probabilidad de Error de Tipo II, β.

El ejemplo anterior conduce a preguntarnos cómo determinar la mejor región crítica en cuanto a la interacción de las probabilidades de error. Para ello, resulta importante analizar la función de potencia de una Prueba de Hipótesis.

5.1.3 Potencia de la Prueba

La Potencia de una contrastación de hipótesis indica la probabilidad de tomar una decisión correcta al rechazar la hipótesis nula.

La **Potencia** de una Prueba de Hipótesis, ψ, es la probabilidad de rechazar la hipótesis nula cuando es falsa, es decir, que es el complemento de la probabilidad de error de Tipo II:

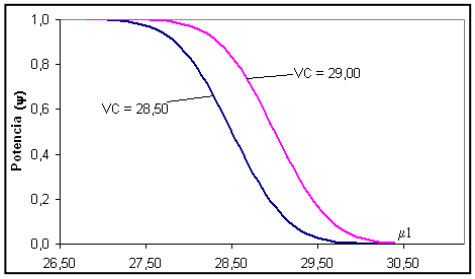
$$\psi = P(\text{Rechazar } H_0 \mid H_0 \text{ es falsa})$$
$$= 1 - \beta$$

Cuando se plantean Hipótesis alternativas compuestas, el valor de la probabilidad de Error de Tipo 2 depende del valor real del parámetro φ cuyo valor se está testeando. Por lo tanto, la potencia, al igual que la probabilidad de error de tipo II, dependerá del valor verdadero del parámetro: $\psi(\varphi)$. Los distintos valores que tome la potencia de acuerdo al valor del parámetro φ se denomina **Función de Potencia**.

Ejemplo 4

Consideremos el ejemplo anterior, donde se ilustró una Tabla con las probabilidades de Error de Tipo II, de acuerdo a los distintos valores del parámetro μ. La Potencia, es simplemente la probabilidad del complemento de β, por lo cual:

VC	ψ (28)	ψ (28.25)	ψ (28.5)	ψ (28.75)	ψ (29)	ψ (29.25)	ψ (29.5)
28.500	0.834	0.686	0.500	0.314	0.166	0.073	0.026
29.000	0.974	0.927	0.834	0.686	0.500	0.314	0.166



Cuando un Valor Crítico tiene una Potencia mayor, tendrá también una mayor Probabilidad de Error de Tipo I. En el Ejemplo 3 se calculó que $\alpha_1 = 0,002$ y $\alpha_2 = 0,026$ corresponden a los valores críticos $VC_1 = 28,5$ y $VC_2 = 29,00$. En la figura se ilustra el gráfico de la Función de Potencia de las Pruebas con dichos VC. Se observa que la Potencia del Test con $VC_2 = 29$ es siempre superior a la potencia considerando $VC_1 = 28,5$. De este modo, podemos ver que $\psi_2 \geq \psi_1$ mientras que $\alpha_2 > \alpha_1$.

Como ya hemos mencionado, **cuanto menor sea la Probabilidad de Error de Tipo I, α , menor será la Potencia, ψ , y viceversa**. En base a esta relación, si se establece el valor máximo permitido de Probabilidad de Error de Tipo I, α^*, el valor crítico VC^* que tenga exactamente esa probabilidad de rechazar H_0 dado que es verdadera (Error de Tipo I) será el que maximice la potencia del test. Veamos esta relación en el ejemplo anterior. El valor α^* mencionado se denomina **Nivel de Significación del Test.**

Ejemplo 5

En base a los supuestos del Ejemplo 3, el test se realizará con un valor máximo de Probabilidad de Error de Tipo I del 5%, es decir que $\alpha^* = 0,05$. Entonces, VC^* se calcula de la siguiente relación:

$$P\left(\bar{X} < VC^* \mid \mu = 30\right) = 0,05$$

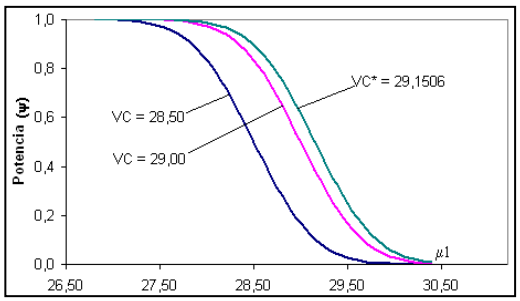
$$\Rightarrow \qquad P\left(z < \frac{VC^* - 30}{2/\sqrt{15}}\right) = 0,05$$

El valor que acumula un 5% en una Normal Estándar es -1,645, por lo tanto

$$-1,645 = \frac{VC^* - 30}{2/\sqrt{15}}$$

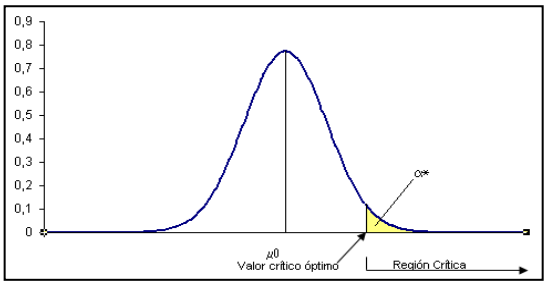
$$\Rightarrow \qquad VC^* = 29,1506$$

En la figura se observa la Función Potencia para este valor crítico. Podemos notar que la potencia, en este caso, es siempre mayor que para los otros valores propuestos.



El **Valor Crítico Óptimo**, VC^*, que maximiza la Potencia de una **Prueba de Hipótesis unilateral inferior** habiendo fijado el nivel máximo permitido para la Probabilidad de Error de Tipo I, α^*, es el valor del estimador que, bajo la Hipótesis Nula, acumula exactamente la probabilidad mencionada {ver figura):

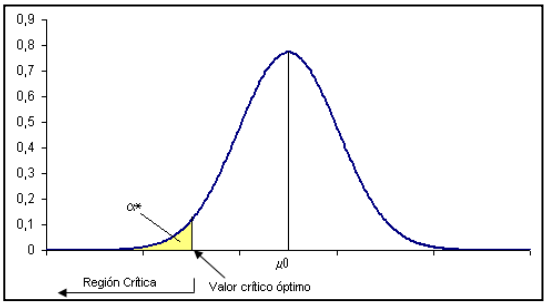
$$P\left(J \le VC^* \middle| \varphi = \varphi_0\right) = \alpha^*$$



La expresión anterior se estableció para Pruebas unilaterales a la izquierda de φ_0. Sin embargo, una expresión análoga es obtenida para pruebas unilaterales a la derecha de φ_0.

El **Valor Crítico Óptimo**, VC^*, que maximiza la Potencia de una **Prueba de Hipótesis unilateral superior,** habiendo fijado el máximo permitido para el error de Tipo I, α^*, será aquel que satisfaga la siguiente relación (ver figura):

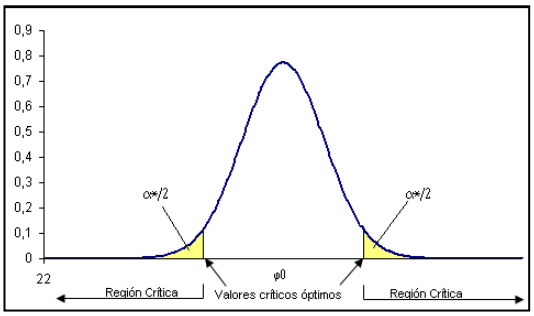
$$P\left(J \le VC^* \middle| \varphi = \varphi_0\right) = 1 - \alpha^* \qquad \Rightarrow P\left(J \ge VC^* \middle| \varphi = \varphi_0\right) = \alpha^*$$



Finalmente, si la Prueba es bilateral, siempre es conveniente elegir la región crítica de manera que la probabilidad de Error de Tipo I quede equi-distribuida en los extremos de la distribución de estimador J. Es decir, que el valor crítico a la izquierda de φ_0, VC_I^*, será aquel que acumula $\alpha^*/2$, mientras que el valor crítico a la derecha, VC_D^*, acumulará $1-\alpha^*/2$ (es decir que dejará una probabilidad de $\alpha^*/2$ a su derecha).

Los **Valores Críticos Óptimos a la izquierda y a la derecha** del valor supuesto para el parámetro bajo la hipótesis nula (VC_I^* y VC_D^*, respectivamente) que maximizan la Potencia de una **Prueba de Hipótesis bilateral** habiendo fijado α^*, serán los que satisfagan (ver figura):

$$P\left(J \le VC_I^* \middle| \varphi = \varphi_0\right) = \alpha^*/2 \qquad \text{y} \qquad P\left(J \le VC_D^* \middle| \varphi = \varphi_0\right) = 1 - \alpha^*/2$$



En base a lo expuesto en esta Sección, en lo que resta del capítulo se utilizará aquel valor crítico que maximice la potencia del test para un nivel de Probabilidad de Error de Tipo I determinado. Eligiendo este valor crítico, obtenemos lo que se denomina la **Prueba uniformemente más potente**.

5.2 Testeo para Medias

En esta sección, mediante la utilización de los conceptos generales expuestos previamente, determinaremos la metodología para testear hipótesis estadísticas relacionadas con la media de una población. Luego, analizaremos la técnica para realizar pruebas cuando el interés recae en la comparación de la media de dos poblaciones.

5.2.1 Prueba para la media con una muestra

En esta sección, expondremos los lineamientos para realizar pruebas donde la hipótesis nula es de la forma $H_0 : \mu = \mu_0$. Analizaremos tanto testeos unilaterales como bilaterales, por lo que la hipótesis alternativa podrá ser cualquiera de las siguientes:

Bilateral	Unilateral Inferior	Unilateral Superior
$H_1 : \mu \neq \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$

El estadístico que se utilizará para realizar la prueba será naturalmente la media muestral $\bar{X}$.

Si la población es Normal (denominado habitualmente “normalidad poblacional”) con media μ y desvío σ , entonces $\bar{X}$ tendrá distribución Normal con media μ y desvío $\sigma/\sqrt{n}$ (Ver Capítulo 4, Sección 3.2. Si, en cambio, la distribución poblacional no es conocida o se sabe que no es normal, pero el tamaño de la muestra es suficientemente grande y se conoce la varianza poblacional, podemos recurrir al Teorema Central del Límite. En ambos casos, se tiene que:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0;1)$$

Siendo α el nivel de significación o sea la probabilidad de rechazar H_0 cuando ésta es verdadera, tenemos que:

$$\alpha = P\left(\text{Rechazar } \mathrm{H}_0 \,|\, \mu = \mu_0\right)$$

Analicemos cómo expresar el evento "Rechazar H_0" en términos estadísticos de acuerdo con cada una de las posibles formulaciones de la Hipótesis Alternativa, y en base a esta formulación cómo determinar el/los Valor/es Crítico/s.

Considerando la prueba bilateral $H_1: \mu \neq \mu_0$, rechazaremos H_0 cuando $\bar{X}$ sea mayor al valor crítico óptimo de la derecha, VC_D^*, o inferior al valor crítico óptimo de la izquierda, VC_I^* (ver Sección 1.3). Es decir:

$$\alpha = P\left(\bar{X} < VC_I^* \ \vee \ \bar{X} > VC_D^* \,|\, \mu = \mu_0\right)$$

Al distribuir la probabilidad de error en la cola superior e inferior de igual manera, tenemos que:

$$\alpha/2 = P\left(\bar{X} < VC_I^* \,|\, \mu = \mu_0\right) \qquad \alpha/2 = P\left(\bar{X} > VC_D^* \,|\, \mu = \mu_0\right) = 1 - P\left(\bar{X} \leq VC_D^* \,|\, \mu = \mu_0\right)$$

Como sabemos que $\bar{X}$ es Normal, podemos escribir que:

$$\alpha/2 = P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < \frac{VC_I^*-\mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z < \frac{VC_I^*-\mu_0}{\sigma/\sqrt{n}}\right) \Rightarrow z_{\alpha/2} = \frac{VC_{\text{Izquierda}}^*-\mu_0}{\sigma/\sqrt{n}}$$

$$1-\alpha/2 = P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \leq \frac{VC_D^*-\mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z \leq \frac{VC_D^*-\mu_0}{\sigma/\sqrt{n}}\right) \Rightarrow z_{1-\alpha/2} = \frac{VC_{\text{Derecha}}^*-\mu_0}{\sigma/\sqrt{n}}$$

Finalmente, despejando los valores críticos de las expresiones anteriores, y utilizando la relación de simetría de la variable Normal Estándar ($z_{\alpha/2} = -z_{1-\alpha/2}$), obtenemos que:

En una prueba bilateral sobre la media, los valores críticos son:

$$VC_{\text{Izquierda}}^* = \mu_0 - z_{1-\alpha/2}\sigma/\sqrt{n} \qquad VC_{\text{Derecha}}^* = \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n}$$

La regla de decisión sería:

Rechazar H_0 **si** $\bar{X} < VC_{\text{Izquierda}}^*$ **ó** $\bar{X} > VC_{\text{Derecha}}^*$.

Ejemplo 6

Considere el Ejemplo 26 del Capítulo 4, donde se supuso que la estatura de los alumnos varones de la Facultad de Ciencias Económicas se distribuye Normalmente con desvío estándar poblacional de 15cm. ¿Cuáles son los valores críticos del siguiente test de hipótesis si se desea un error de tipo I máximo del 5% y se toma una muestra de 20 datos?

$$H_0: \mu = 1{,}80m.$$

$$H_1: \mu \neq 1{,}80m.$$

Utilizando las fórmulas recién expuestas, podemos calcular:

$$VC_{\text{Izquierda}}^* = \mu_0 - z_{0{,}975} \times 0{,}15/\sqrt{20} = 1{,}80 - 1{,}96 \times 0{,}15/\sqrt{20} \cong 1{,}7343$$

$$VC_{\text{Derecha}}^* = \mu_0 + z_{0{,}975} \times 0{,}15/\sqrt{20} = 1{,}80 + 1{,}96 \times 0{,}15/\sqrt{20} \cong 1{,}8657$$

Por lo tanto, si la media muestral se encuentra entre estos dos valores, no podrá rechazarse la hipótesis nula. Por el contrario, si excede al VC_{Derecha}^* o es inferior al $VC_{\text{Izquierda}}^*$, no se aceptará H_0.

Con los datos del Ejemplo 26 del Capítulo 4 se obtuvo que $\bar{x} = 1{,}75$ y, por lo tanto, no puede rechazarse la hipótesis de que la media poblacional sea de 1,80 m.

Para el caso de una prueba unilateral inferior, se rechazará la Hipótesis Nula si el estadístico es inferior al valor crítico. Es decir que:

$$\alpha = P\left(\bar{X} < VC^* \,|\, \mu = \mu_0\right)$$

Sabiendo que $\bar{X}$ tiene distribución Normal, podemos escribir que:

$$\alpha = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{VC^* - \mu_0}{\sigma/\sqrt{n}}\right)$$
$$= P\left(Z < \frac{VC^* - \mu_0}{\sigma/\sqrt{n}}\right)$$
$$\Rightarrow z_\alpha = \frac{VC^* - \mu_0}{\sigma/\sqrt{n}}$$

Utilizando la relación de simetría ($z_\alpha = -z_{1-\alpha}$), y despejando de la última expresión obtenemos el Valor Crítico.

En una prueba unilateral inferior sobre la media, el valor crítico es:

$$VC^* = \mu_0 - z_{1-\alpha}\sigma/\sqrt{n}$$

La regla de decisión sería:

Rechazar H_0 **si** $\bar{X} < VC^*$

Por último, si se trata de una prueba de hipótesis unilateral superior, se rechazará H_0 si el estadístico muestral supera al valor crítico. Es decir que:

$$\alpha = P\left(\bar{X} > VC^* \,|\, \mu = \mu_0\right) \quad \Rightarrow \qquad 1-\alpha = P\left(\bar{X} \le VC^* \,|\, \mu = \mu_0\right)$$

Utilizando la distribución de $\bar{X}$, obtenemos que:

$$1-\alpha = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \le \frac{VC^* - \mu_0}{\sigma/\sqrt{n}}\right)$$
$$= P\left(Z \le \frac{VC^* - \mu_0}{\sigma/\sqrt{n}}\right)$$
$$\Rightarrow z_{1-\alpha} = \frac{VC^* - \mu_0}{\sigma/\sqrt{n}}$$

No queda más que despejar el Valor Crítico de la última expresión.

En una prueba unilateral superior sobre la media de una población con varianza conocida, el valor crítico es:

$$VC^* = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$$

La regla de decisión sería:

Rechazar H_0 **si** $\bar{X} > VC^*$

Ejemplo 7

Como en el Ejemplo 13 del Capítulo 4, supongamos que se está analizando el contenido medio de los envases de dulce de leche que produce la empresa A. Los envases dicen en su etiqueta que el contenido es de 500 gr., y si el contenido real fuera inferior se estaría estafando al consumidor. De este modo, el interés de los agentes de control recae testear las hipótesis siguientes:

$$H_0 : \mu = 500$$
$$H_1 : \mu < 500$$

El control es muy estricto, por lo que se tolera un error de Tipo I del 1%, y se tomará una muestra de 40 potes. Se supone que la distribución es Normal con desvío estándar de 20gr.

Con esa información, se plantea el VC para test unilateral inferior:

$$\begin{aligned} VC^* &= \mu_0 - z_{0,99}\sigma/\sqrt{n} \\ &= 500 - 2,3263 \times 20/\sqrt{40} \\ &\cong 492,64 \end{aligned}$$

Por lo tanto, si en la muestra se obtiene que $\bar{X} < 492,64$ se rechazará la hipótesis nula de que la media es de 500 gr., y correspondería alguna sanción a la empresa.

Ya hemos visto en capítulos anteriores que, generalmente, no se conoce la varianza poblacional, por lo que se suele recurrir a una estimación de la misma. En esos casos, hemos visto en el Capítulo 4 que la distribución de muestreo apropiada para $\bar{X}$ es la *t-Student*:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \square t_{n-1}$$

Utilizando los mismos razonamientos que en el caso de varianza poblacional conocida, pueden obtenerse los siguientes valores críticos:

En una prueba de hipótesis sobre la media de una población con varianza desconocida, los valores críticos y las reglas de decisión son:

	Bilateral	Unilateral Inferior	Unilateral Superior
Valor/es Crítico/s	$VC^*_{Izq.} = \mu_0 - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}$ $VC^*_{Der.} = \mu_0 + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}$	$VC^* = \mu_0 - t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}$	$VC^* = \mu_0 + t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}$
Regla de Decisión	**Rechazar** H_0 **si** $\bar{X} < VC^*_{Izq.}$ ó $\bar{X} > VC^*_{Der.}$	**Rechazar** H_0 **si** $\bar{X} < VC^*$	**Rechazar** H_0 **si** $\bar{X} > VC^*$

Ejemplo 8

En el Ejemplo 20 del Capítulo 4, se planteó el caso de una empresa productora de automóviles que debía reducir la emisión de dióxido de carbono (CO2) de los coches que produce, debido a una nueva legislación por parte del gobierno: la emisión debe ser de cómo máximo de 140 gramos por kilómetro al finalizar el corriente año. Al finalizar el año, el gobierno tomó una muestra de 20 coches de la fábrica y observó una media de 143 g/km. y un desvío estándar de 5 g/km. Se supone que la emisión de CO2 sigue una distribución Normal. En el ejemplo mencionado, se calculó la probabilidad de que la media muestral sea 143, siendo la media real 140.

Ese procedimiento es un tanto rudimentario, y se realizó únicamente para presentar la distribución de muestreo. Lo más natural sería formular una prueba de hipótesis que testee si la media realmente es 140 g/km. Por lo tanto, las hipótesis nula y alternativa serían:

$$H_0 : \mu = 140$$
$$H_1 : \mu > 140$$

H_1 fue formulada de esa manera porque el interés del gobierno radica en un exceso de emisión.

Utilizando lo expuesto en esta sección, y suponiendo que el test se realiza al 95%, tenemos que:

$$VC = \mu_0 + t_{19;0,95} \times s/\sqrt{n}$$
$$= 140 + 1,729 \times 5/\sqrt{20}$$
$$= 141,93$$

Si $\bar{X}$ supera ese valor, se rechazará la hipótesis nula. Como el valor observado es de 143, se rechaza la hipótesis de que la emisión es de 140 g/km. Y se concluye que la empresa no cumplió con el requisito.

5.2.2 Prueba para comparar medias de dos poblaciones

Siguiendo los mismos razonamientos que los expuestos en la sección anterior, podemos formular pruebas de hipótesis referidas a la diferencia entre la media de dos poblaciones. En este caso, generalmente se desea testear si existe diferencia entre la media de las dos poblaciones, por lo que las hipótesis nula y alternativa suelen ser:

$$H_0 : \mu_A = \mu_B \quad \Rightarrow \quad H_0 : \mu_A - \mu_B = 0$$
$$H_1 : \mu_A \neq \mu_B \quad \Rightarrow \quad H_1 : \mu_A - \mu_B \neq 0$$

Si bien pueden plantearse pruebas de hipótesis unilaterales, expondremos aquí solamente el desarrollo para testear hipótesis con dos colas como la recién expuesta, siendo directa la extensión para pruebas de una cola.

De modo más general, las hipótesis nula y alternativa del test bilateral pueden expresarse como:

$$H_0 : \mu_A - \mu_B = c_0$$

$$H_1 : \mu_A - \mu_B \neq c_0$$

siendo c_0 una constante que indica en cuánto excede μ_A a μ_B bajo la hipótesis nula. Como hemos mencionado, en las aplicaciones prácticas, generalmente, se utiliza $c_0 = 0$, y se testea si las dos medias son iguales o no.

En la Sección 5.1 del Capítulo 6 se vio que si las variables X_A y X_B provienen de poblaciones normales (supuesto de "normalidad poblacional") independientes, con varianzas conocidas, entonces:

$$\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{{\sigma_A}^2}{n} + \frac{{\sigma_B}^2}{m}}} \sim N(0;1)$$

Al conocer la distribución de este estadístico que será utilizado para realizar el test, la determinación de los valores críticos para la variable $\bar{X}_A - \bar{X}_B$ es directa. Si se asigna la mitad de la probabilidad de error de Tipo I a cada lado del valor de la diferencia entre las medias bajo la hipótesis nula (a cada lado de c_0), entonces, el valor crítico de la derecha acumulará una probabilidad de $1-\alpha/2$. Por lo tanto, se tiene que:

$$1-\alpha/2 = P\left(\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}} \leq \frac{VC_D^* - c_0}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}} \right)$$
$$= P\left(Z \leq \frac{VC_D^* - c_0}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}} \right)$$
$$\Rightarrow z_{1-\alpha/2} = \frac{VC_D^* - c_0}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}}$$

Despejando obtenemos el VC correspondiente a la cola superior:

$$VC_D^* = c_0 + z_{1-\alpha/2}\sqrt{\sigma_A^2/n + \sigma_B^2/m}$$

Por otra parte, el valor crítico inferior será aquel que acumule la mitad de la probabilidad de error de Tipo I, por lo que:

$$\begin{aligned}\alpha/2 &= P\left(\bar{X}_A - \bar{X}_B < VC_I^*\right)\\ &= P\left(\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}} < \frac{VC_I^* - c_0}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}}\right)\\ &= P\left(Z < \frac{VC_I^* - c_0}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}}\right)\\ \Rightarrow z_{\alpha/2} &= \frac{VC_I^* - c_0}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}}\end{aligned}$$

Despejando, y utilizando que $z_{\alpha/2} = -z_{1-\alpha/2}$, obtenemos que:

$$VC_{\mathrm{I}}^* = c_0 - z_{1-\alpha/2}\sqrt{\sigma_A^2/n + \sigma_B^2/m}$$

Cuando se realiza la siguiente prueba de hipótesis:

$$H_0 : \mu_A - \mu_B = c_0$$

$$H_1 : \mu_A - \mu_B \neq c_0$$

La regla de decisión es:

Rechazar si $\bar{X}_A - \bar{X}_B < VC_{\mathrm{I}}^*$ ó $\bar{X}_A - \bar{X}_B > VC_{\mathrm{D}}^*$

Siendo

$$VC_{\mathrm{I}}^* = c_0 - z_{1-\alpha/2}\sqrt{\sigma_A^2/n + \sigma_B^2/m} \qquad VC_{\mathrm{D}}^* = c_0 + z_{1-\alpha/2}\sqrt{\sigma_A^2/n + \sigma_B^2/m}$$

Ejemplo 9

Una ONG desea comparar si el ingreso medio de dos ciudades es igual. Se supone que la distribución de los ingresos es Normal, y que los desvíos son $\sigma_X = 600$ y $\sigma_Y = 750$. Para realizar el estudio se realizará una prueba de hipótesis con $\alpha = 0,05$. Si los datos obtenidos del proceso de muestreo son los siguientes, ¿a qué conclusión se llega?

Ciudad X	Ciudad Y
$n_X = 60$	$n_Y = 70$
$\bar{x} = 2100$	$\bar{y} = 1850$

Expresando en términos estadísticos lo enunciado, tenemos que:

$$H_0 : \mu_X - \mu_Y = 0$$

$$H_1 : \mu_X - \mu_Y \neq 0$$

Los valores críticos para la prueba son:

$$\begin{aligned}VC_{\mathrm{I}}^* &= c_0 - z_{0,975}\sqrt{\sigma_A^2/n + \sigma_B^2/m} & VC_{\mathrm{D}}^* &= c_0 + z_{0,975}\sqrt{\sigma_A^2/n + \sigma_B^2/m}\\ &= 0 - 1,96 \times \sqrt{600^2/60 + 750^2/70} & &= 0 + 1,96 \times \sqrt{600^2/60 + 750^2/70}\\ &= -232,20 & &= 232,50\end{aligned}$$

Si calculamos $\bar{x} - \bar{y} = 2100 - 1850 = 250$, obtenemos que la diferencia entre las medias muestrales excede el valor crítico superior y, por lo tanto, no se acepta la hipótesis nula.

Cuando bajo el supuesto de "normalidad poblacional" las varianzas poblacionales son desconocidas, pero pueden suponerse iguales, como se expuso en el Capítulo 6, se debe utilizar la distribución *t-Student*. Una vez conocida la distribución de muestreo, la deducción de los valores críticos es inmediata. Si consideramos la siguiente prueba de hipótesis:

$$H_0 : \mu_A - \mu_B = c_0$$

$$H_1 : \mu_A - \mu_B \neq c_0$$

Los valores críticos serían:

$$VC^*_{\text{Izquierda}} = c_0 - t_{n+m-2;1-\alpha/2} s_P \sqrt{1/n + 1/m}$$

$$VC^*_{\text{Derecha}} = c_0 + t_{n+m-2;1-\alpha/2} s_P \sqrt{1/n + 1/m}$$

Donde $s_P^2 = \dfrac{(n-1)s_A^2 + (m-1)s_B^2}{n+m-2}$

Note que a diferencia de los casos anteriores, aquí los VC dependen de los datos muestrales, ya que incluye la varianza muestral común. En estas situaciones, suele realizarse el testeo con la variable estandarizada directamente, comparándola con la distribución teórica.

Considere la siguiente prueba de hipótesis en dos poblaciones Normales cuyas varianzas son desconocidas, pero pueden suponerse iguales:

$$H_0 : \mu_A - \mu_B = c_0$$

$$H_1 : \mu_A - \mu_B \neq c_0$$

La regla de decisión es:

Rechazar si $\dfrac{\bar{X}_A - \bar{X}_B - c_0}{s_P \sqrt{1/n + 1/m}} < -t_{n+m-2;1-\alpha/2}$ ó $\dfrac{\bar{X}_A - \bar{X}_B - c_0}{s_P \sqrt{1/n + 1/m}} > t_{n+m-2;1-\alpha/2}$

Ejemplo 10

Considere el ejemplo anterior, y suponga que no se conocen los desvíos poblacionales y, que en base a la muestra, se obtiene lo siguiente:

Ciudad X	Ciudad Y
$n_X = 60$	$n_Y = 70$
$\bar{x} = 2100$	$\bar{y} = 1850$
$s_X = 620$	$s_Y = 670$

Calculando la varianza común, tenemos que:

$$s_P = \sqrt{\frac{(n-1)s_A^2 + (m-1)s_B^2}{n+m-2}}$$

$$= \sqrt{\frac{(60-1)\times 620^2 + (70-1)\times 670^2}{60+70-2}}$$

$$\cong 647,43$$

Los valores críticos de la distribución *t-Student* son:

$$t_{n+m-2;1-\alpha/2} = t_{128;0,975} = 1,6568 \quad \text{y} \quad -t_{128;0,975} = -1,6568$$

Si calculamos el estadístico para realizar la prueba, tenemos que:

$$\frac{\bar{x}-\bar{y}-c_0}{s_P\sqrt{1/n+1/m}}=\frac{2100-1850-0}{647,43\times\sqrt{1/60+1/70}}\cong 2,19$$

Siendo este valor superior al VC superior de la variable t, no se acepta la hipótesis nula.

5.3 Testeo para proporciones

En esta sección expondremos los lineamientos para la realización de pruebas de hipótesis relacionadas con la proporción poblacional cuando se realizan estudios de índole cualitativa. Para la realización de los testeos, se utiliza la aproximación Normal para la distribución Binomial, por lo que los desarrollos serán muy similares a los expuestos en la sección anterior.

5.3.1 Prueba para la proporción con una muestra

Hemos visto en la Sección 4 del Capítulo 4, que cuando el tamaño de la muestra es lo suficientemente grande, la distribución de la proporción muestral puede aproximarse mediante una distribución Normal. Más precisamente, hemos visto que:

$$\frac{\bar{p}-p}{\sqrt{p(1-p)/n}}\sim N(0;1)$$

Utilizando esta distribución de muestreo podemos, obtener los valores críticos necesarios para probar la hipótesis nula:

$$H_0: p = p_0$$

contra cualquiera de las siguientes hipótesis alternativas:

Bilateral	Unilateral Inferior	Unilateral Superior
$H_1: p \neq p_0$	$H_1: p < p_0$	$H_1: p > p_0$

La deducción de los valores críticos es análoga a la expuesta en la sección anterior, y se propone como ejercicio. En este punto, simplemente mencionamos que para hallar los valores críticos se parte de la base que la hipótesis nula es verdadera, por lo que para el cálculo de la varianza se utilizará el valor p_0

En una prueba de hipótesis sobre la proporción de una población de la forma $H_0: p = p_0$, los valores críticos y las reglas de decisión para cada posible hipótesis alternativa son:

	Bilateral	Unilateral Inferior	Unilateral Superior
Valor/es Crítico/s	$VC^*_{Izq.} = p_0 - z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$ $VC^*_{Der.} = p_0 + z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$	$VC^* = p_0 - z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$	$VC^* = p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$
Regla de Decisión	**Rechazar** H_0 **si** $\bar{p} < VC^*_{Izq.}$ ó $\bar{p} > VC^*_{Der.}$	**Rechazar** H_0 **si** $\bar{p} < VC^*$	**Rechazar** H_0 **si** $\bar{p} > VC^*$

Ejemplo 11

Si un candidato cree que ganará las próximas elecciones obteniendo aproximadamente el 50% de los votos (con el 45% se asegura no tener que ir a ballotage). Por lo tanto, ya que el interés radica en encontrar evidencia suficiente para negar su creencia de que ganará, sus hipótesis nula y alternativa serán:

$$H_0: p = 0,50$$

$$H_1: p < 0,50$$

Suponga que se realizará la prueba de hipótesis con nivel de significación del 5% (probabilidad de error de Tipo I) y que para realizar la prueba se tomará una muestra de 500 personas. Entonces, utilizando lo expuesto en esta sección, podemos hallar el VC y construir la regla de decisión: El valor crítico, al tratarse de una prueba unilateral inferior es:

$$VC = p_0 - z_{1-\alpha/2}\sqrt{p_0(1-p_0)/n}$$
$$= 0,5 - 1,645\sqrt{0,5\times(1-0,5)/500}$$
$$\cong 0,463$$

Por lo tanto, si más del 46,3% de los encuestados está a favor del candidato, no se podrá rechazar la hipótesis nula, mientras que en caso contrario se rechazará la misma concluyendo que la evidencia muestral rechaza la hipótesis de que el candidato ganará con el 50%.

Si deseamos expresar ese porcentaje crítico en cantidad de personas, nada más deberemos multiplicarlo por el tamaño muestral: $0,463\times 500 = 231,61$. Por lo tanto, si 232 personas o más favorecen al candidato, no se podrá rechazar la hipótesis nula.

5.3.2 Prueba para comparar proporciones de dos poblaciones

En ocasiones, resulta de interés comparar las proporciones de alguna propiedad cualitativa de cierto fenómeno. Por ejemplo, podría existir interés en saber si una técnica de producción produce la misma cantidad de fallas que otra, o si la intención de voto para cierto candidato es la misma en dos provincias. En estos casos, se plantea la Hipótesis Nula siguiente:

$$H_0 : p_A - p_B = q_0$$

Esta formulación es general, y cuando el interés recaiga en testear la igualdad de las proporciones, se remplazará $q_0 = 0$.

Las hipótesis alternativas pueden formularse de cualquiera de las formas que se exponen a continuación.

Bilateral	Unilateral Inferior	Unilateral Superior
$H_1 : p_A - p_B \neq q_0$	$H_1 : p_A - p_B < q_0$	$H_1 : p_A - p_B > q_0$

En este apartado, nos focalizaremos únicamente en el desarrollo del caso bilateral para el valor particular $q_0 = 0$. Es decir, que se analizará la siguiente prueba de hipótesis:

$$H_0 : p_A = p_B$$

$$H_1 : p_A \neq p_B$$

En el Capítulo 4, Sección 11.2, hemos visto que:

$$\frac{(\bar{p}_A - \bar{p}_B) - (p_A - p_B)}{\sqrt{\frac{p_A(1-p_A)}{n} + \frac{p_B(1-p_B)}{m}}} \sim N(0;1)$$

Donde $\bar{p}_A = X/n$ y $\bar{p}_B = Y/m$ son las proporciones muestrales de la población A y B, respectivamente. Bajo la hipótesis nula tenemos que $p_A = p_B$, por lo que remplazando en la expresión anterior se obtiene que:

$$\frac{(\bar{p}_A - \bar{p}_B)}{\sqrt{p_A(1-p_A)(\frac{1}{n} + \frac{1}{m})}} \sim N(0;1)$$

Al fijar el nivel de la probabilidad de error de tipo I, α, se construye una región crítica que deje $\alpha/2$ en la cola derecha y $\alpha/2$ en la cola izquierda de la distribución, siendo la media de la

misma igual a cero (ya que bajo H_0 tenemos que $p_A - p_B = 0$). Sin embargo, para calcular el denominador de la expresión anterior, debe conocerse el valor común $p = p_A = p_B$ supuesto para la proporción poblacional. Como el mismo es desconocido, se utiliza el estimador de máxima verosimilitud bajo H_0 (ver Novales, 1997, Capítulo 10):

$$\hat{p} = \frac{X+Y}{n+m}$$

De este modo, para la cola derecha tenemos que:

$$\alpha/2 = P\left(\frac{\bar{p}_A - \bar{p}_B}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} > \frac{VC_{\mathrm{D}}^*}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} \right)$$

$$= P\left(\frac{\bar{p}_A - \bar{p}_B}{\sqrt{(X+Y)(n+m-X-Y)/\left[nm(n+m)\right]}} > \frac{VC_{\mathrm{D}}^*}{\sqrt{(X+Y)(n+m-X-Y)/\left[nm(n+m)\right]}} \right)$$

$$= P\left(Z > \frac{VC_{\mathrm{D}}^*}{\sqrt{(X+Y)(n+m-X-Y)/\left[nm(n+m)\right]}} \right)$$

Utilizando el complemento de probabilidades, podemos escribir:

$$1-\alpha/2 = P\left(Z \le \frac{VC_{\mathrm{D}}^*}{\sqrt{(X+Y)(n+m-X-Y)/\left[nm(n+m)\right]}} \right)$$

Luego, buscando el valor correspondiente de la distribución Normal Estándar, tenemos que:

$$z_{1-\alpha/2} = \frac{VC_{\mathrm{D}}^*}{\sqrt{(X+Y)(n+m-X-Y)/\left[nm(n+m)\right]}}$$

Siguiendo un razonamiento similar obtenemos que:

$$-z_{1-\alpha/2} = \frac{VC_{\mathrm{I}}^*}{\sqrt{(X+Y)(n+m-X-Y)/\left[nm(n+m)\right]}}$$

Nótese que, en este caso, no es posible despejar directamente los valores críticos, ya que en los denominadores de las expresiones de los miembros derechos se encuentran también datos muestrales. Sin embargo, de las expresiones anteriores podemos ver que se rechazará la hipótesis nula siempre que:

$$\frac{\bar{p}_A - \bar{p}_B}{\sqrt{(X+Y)(n+m-X-Y)/\left[nm(n+m)\right]}} < -z_{1-\alpha/2} \text{ ó}$$

$$\frac{\bar{p}_A - \bar{p}_B}{\sqrt{(X+Y)(n+m-X-Y)/\left[nm(n+m)\right]}} > z_{1-\alpha/2}$$

Ejemplo 12

Suponga que una empresa fabrica un determinado artículo a través del Proceso A, y se planea modificarlo para utilizar el Proceso B, siempre y cuando haya evidencia suficiente de que este último produce una cantidad de artículos de defectuosos significativamente menor al primero. Para tomar la decisión, se realizará una prueba de hipótesis tomando 100 artículos producidos con cada proceso ($n = m = 100$). Como el cambio de proceso implica ciertos costos, el test se realizará con un 1% de significación; para realizar el cambio solamente existe "mucha" evidencia de que las proporciones de

artículos defectuosos no son iguales. El valor crítico correspondiente a la Normal Estándar es $z_{0,995} = 2,5758$.

Para ilustrar la metodología, supongamos que en la muestra del proceso A se obtuvieron 10 artículos defectuosos, mientras que con el proceso B solamente 5. Bajo el test planteado, ¿es este resultado suficiente para concluir que la proporción de artículos defectuosos de los dos procesos son diferentes?

Con estos datos, podemos calcular el valor del estadístico siguiente y compararlo con el valor correspondiente a la Normal Estándar:

$$VC = \frac{\bar{p}_A - \bar{p}_B}{\sqrt{(X+Y)(n+m-X-Y)/\left[nm(n+m)\right]}}$$
$$= \frac{10/100 - 5/100}{\sqrt{(10+5)(100+100-10-5)/\left[100\times 100(100+100)\right]}}$$
$$= 1,3423$$

Como este valor es menor a 2,5758, resulta que la evidencia muestral no es suficiente para rechazar la hipótesis nula bajo la cual las proporciones de artículos defectuosos de los dos procesos son iguales.

5.4 Testeo para varianzas

Muchas veces, el interés radica en conocer la varianza de cierta población, o bien, en saber si un valor estimado *a priori* sin información muestral es consistente con lo observado.

Por ejemplo, cuando se construyen IC o se realizan pruebas de hipótesis referidas a la media de una población Normal, y se supone conocida la varianza poblacional, resultaría de interés un test que nos permitiera determinar si el valor que se supone conocido de la varianza es consistente con los datos observados en la muestra.

Además, cuando se construyen IC o se realizan pruebas de hipótesis para la diferencia de medias, siendo las varianzas poblacionales desconocidas, se supone que las dos poblaciones tienen la misma varianza para utilizar la distribución *t-Student*. En este caso, resulta fundamental realizar un test de hipótesis para saber si el supuesto de igualdad de varianzas tiene razón de ser, o si los datos refutan dicha afirmación. Debido a este especial interés, en la Sección 4.2, nos remitiremos al testeo de igualdad de varianzas únicamente.

Para la obtención de los valores críticos utilizaremos (bajo el supuesto de "normalidad poblacional") las distribuciones Chi-Cuadrado y F de Fisher-Snedecor. La primera, cuando se realizan pruebas sobre una población (ver Capítulo 4, Sección 5.1), y la segunda, cuando el interés radica en la comparación de las varianzas de dos poblaciones (ver Capítulo 4, Sección 5.2)

5.4.1 Prueba para la varianza con una muestra

En la Sección 5.1 del Capítulo 4, se ha expuesto la distribución de muestreo de la varianza muestral s^2 cuando se trabaja con poblaciones Normales con media desconocida. Esa distribución nos permitirá realizar pruebas de hipótesis estadísticas del tipo:

Hipótesis Nula	$H_0 : \sigma^2 = \sigma_0^2$		
Hipótesis Alternativa	Bilateral	Unilateral Inferior	Unilateral Superior
	$H_1 : \sigma^2 \neq \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$

Hemos visto en el Capítulo 4 que:

$$\frac{(n-1)s^2}{\sigma^2} \sim X^2_{n-1}$$

Por lo tanto, una vez que se ha especificado el máximo tolerado de probabilidad de Error de Tipo I, α , el valor crítico del estadístico s^2 que limita las regiones de rechazo y no rechazo para una **prueba unilateral superior,** puede obtenerse utilizando la distribución mencionada:

$$\begin{aligned}\alpha &= P\left(\text{Rechazar H}_0 \,\middle|\, \sigma^2 = \sigma_0^2\right) \\ &= P\left(s^2 > VC^* \,\middle|\, \sigma^2 = \sigma_0^2\right) \\ &= P\left(\frac{(n-1)s^2}{\sigma_0^2} > \frac{(n-1)VC^*}{\sigma_0^2}\right) \\ &= P\left(\chi^2_{n-1} > \frac{(n-1)VC^*}{\sigma_0^2}\right)\end{aligned}$$

Luego, utilizando las propiedades de la probabilidad de eventos complementarios tenemos que:

$$1-\alpha = P\left(\chi^2_{n-1} \le \frac{(n-1)VC^*}{\sigma^2}\right)$$

Con un Software (o utilizando las tablas tradicionales), puede obtenerse el valor de la variable Chi-cuadrado con $n-1$ grados de libertad que acumula una probabilidad de $1-\alpha$: $\chi^2_{n-1;1-\alpha}$. Por lo tanto, igualando este valor al miembro derecho del argumento de la probabilidad anterior, tenemos la siguiente ecuación de la cual se puede despejar directamente el valor crítico:

$$\chi^2_{n-1;1-\alpha} = \frac{(n-1)VC^*}{\sigma_0^2}$$

En una prueba de hipótesis sobre la varianza, una población Normal con media desconocida formulada como, $H_0 : \sigma^2 = \sigma_0^2$, los valores críticos del estadístico s^2 y las reglas de decisión para cada posible hipótesis alternativa son:

	Bilateral	Unilateral Inferior	Unilateral Superior
Valor/es Crítico/s	$VC^*_{\text{Izq.}} = \frac{\chi^2_{n-1;\alpha/2}\sigma_0^2}{(n-1)}$ $VC^*_{\text{Der.}} = \frac{\chi^2_{n-1;1-\alpha/2}\sigma_0^2}{(n-1)}$	$VC^* = \frac{\chi^2_{n-1;\alpha}\sigma_0^2}{(n-1)}$	$VC^* = \frac{\chi^2_{n-1;1-\alpha}\sigma_0^2}{(n-1)}$
Regla de Decisión	**Rechazar** H_0 **si** $s^2 < VC^*_{\text{Izq.}}$ ó $s^2 > VC^*_{\text{Der.}}$	**Rechazar** H_0 **si** $s^2 < VC^*$	**Rechazar** H_0 **si** $s^2 > VC^*$

La determinación del valor crítico para el caso de la prueba unilateral superior se expuso anteriormente. Los desarrollos para el caso bilateral y prueba unilateral inferior se proponen como ejercicio.

Ejemplo 13

Considere los Ejemplos 7 y 8 del capítulo anterior, donde se analizó la duración de la producción de lamparitas. Suponiendo que la duración está distribuida de forma Normal, se desea testear si el desvío estándar es superior a las 50 horas, con un error de Tipo I máximo de 0,01 y una muestra de 20 artículos.

En base a lo mencionado, el test a realizar es:

$$H_0 : \sigma_0^2 = 2500$$
$$H_1 : \sigma_0^2 > 2500$$

Note que la prueba se plantea en términos de la varianza.

El valor crítico es el expuesto en la tabla anterior:

$$VC^* = \frac{\chi^2_{n-1;1-\alpha}\sigma_0^2}{(n-1)}$$

Siendo $n = 20$ el tamaño muestral y $\alpha = 0,01$ el error de Tipo I, podemos calcular que $\chi^2_{19;0,99} = 36,1909$, y por lo tanto:

$$VC^* = \frac{36,1909 \times 2500}{19} \cong 4761,956$$

Por lo tanto, si la varianza muestral supera este valor, deberá rechazarse la hipótesis nula.

Con los datos del Capítulo 6 se obtuvo que $s^2 = 3600$, y por lo tanto la evidencia muestral no es suficiente para rechazar H_0.

5.4.2 Prueba para comparar varianzas de dos poblaciones

Cuando resulta de interés comparar la variabilidad de dos poblaciones, toman importancia las pruebas de hipótesis para la igualdad de varianzas. En este caso, se plantea como hipótesis nula la igual de las varianzas de dos poblaciones y el valor crítico se obtiene utilizando la distribución F de Fisher-Snedecor que se utiliza para la distribución de muestreo del cociente de varianzas bajo el supuesto de "normalidad poblacional".

De manera general, la hipótesis nula puede formularse como:

$$H_0 : \frac{\sigma_A^2}{\sigma_B^2} = c_0$$

Sin embrago, como nuestro interés recae en la igualdad de las varianzas, nos abocaremos únicamente al caso en que $c_0 = 1$ y, en consecuencia, la hipótesis quedaría:

$$H_0 : \sigma_A^2 = \sigma_B^2$$

La formulación de la hipótesis alternativa puede plantearse de manera de construir una prueba bilateral o una prueba unilateral superior o inferior. Aquí, nos ocuparemos únicamente del caso de la prueba de dos colas, por lo que la hipótesis alternativa será:

$$H_1 : \sigma_A^2 \neq \sigma_B^2$$

De acuerdo a lo visto en la Sección 5.2 del Capítulo 4, tenemos que:

$$\frac{{s_X}^2/{\sigma_X}^2}{{s_Y}^2/{\sigma_Y}^2} \sim F(n-1; m-1)$$

Como siempre en las pruebas bilaterales, distribuimos la probabilidad de error de Tipo I en las dos colas de la distribución de muestreo. Por lo tanto, para la determinación del valor crítico de la cola inferior tenemos que:

$$\begin{aligned} \alpha/2 &= P\left(s_A^2/s_B^2 > VC_{\mathrm{D}}^* \middle| \sigma_A^2 = \sigma_B^2\right) \\ &= P\left(\frac{s_A^2/\sigma_A^2}{s_B^2/\sigma_B^2} > \frac{\sigma_B^2}{\sigma_A^2} VC_{\mathrm{D}}^*\right) \\ &= P\left(F_{n-1;m-1} > \frac{\sigma_B^2}{\sigma_A^2} VC_{\mathrm{D}}^*\right) \end{aligned}$$

Teniendo en cuenta que bajo H_0 las varianzas son iguales, $\sigma_A^2 = \sigma_B^2$, el cociente de las mismas puede simplificarse en el argumento de la probabilidad anterior. Además, utilizando las propiedades de las probabilidades de eventos complementarios tenemos que:

$$1-\alpha/2 = P\left(F_{n-1;m-1} \leq VC_{\mathrm{D}}^{*}\right)$$

Por lo tanto, el valor crítico inferior es directamente el percentil de una variable F que acumula una probabilidad de $1-\alpha/2$ con $n-1$ g.l en el numerador y $m-1$ g.l en el denominador, el cual puede buscarse directamente en una tabla o calcularse con un Software:

$$F_{(n-1;m-1);1-\alpha/2} = VC_{\mathrm{D}}^{*}$$

Siguiendo un razonamiento similar, obtenemos que:

$$F_{(n-1;m-1);\alpha/2} = VC_{\mathrm{I}}^{*}$$

Note que si bien utilizamos los subíndices "Izq." y "Der.", estos valores no se corresponden a la derecha e izquierda de la distribución, ya que el valor "Izq." acumula $\alpha/2$ mientras que el valor "Der." acumula $1-\alpha/2$, y siendo generalmente $\alpha < 0,5$ el valor "Izq." estará a la derecha de la distribución mientras que el valor "Der." estará a la izquierda. Los subíndices se utilizaron simplemente para mantener la notación de las secciones anteriores y lo que en definitiva importa es la regla de decisión que se expone a continuación:

Rechazar H_0 **si** $\frac{s_A^2}{s_B^2} < F_{(n-1;m-1);\alpha/2}$ **ó** $\frac{s_A^2}{s_B^2} > F_{(n-1;m-1);1-\alpha/2}$

Ejemplo 14

En el Ejemplo 13 del capítulo anterior, se analizó la duración de las lamparitas producidas por una determinada empresa, y se comparó la variabilidad de la producción actual con aquélla correspondiente a la producción obtenida con una nueva máquina. La empresa desea testear si la variabilidad en la duración de ambas máquinas es igual o distinta con un error de Tipo I de 0,05:

$$H_0 : \sigma_A^2 = \sigma_B^2$$

$$H_1 : \sigma_A^2 \neq \sigma_B^2$$

$$\alpha = 0,05$$

En el ejemplo mencionado, los tamaños muestrales fueron $n = 20$ y $m = 25$ y, por lo tanto, los valores críticos son:

$$F_{(n-1;m-1);\alpha/2} = F_{(19;24);0,025} = 0,4078 \qquad F_{(n-1;m-1);1-\alpha/2} = F_{(19;24);0,975} = 2,3452$$

En consecuencia, si el cociente entre la varianza muestral del proceso actual y la del nuevo proceso está comprendido entre estos dos valores, no podrá rechazarse H_0 .

En el capítulo anterior habíamos calculado $s_{\text{Nueva}}^2 = 1871$ y $s_{\text{Actual}}^2 = 3600$, y por lo tanto, no puede rechazarse la hipótesis nula de que la varianzas son iguales, ya que:

$$s_A^2 / s_B^2 = 3600/1871 \cong 1,9241$$

5.5 Relación de las pruebas de hipótesis con los Intervalos de confianza

Habrá notado que existen muchas similitudes en los procedimientos para la realización de Pruebas de Hipótesis y la construcción de Intervalos de Confianza IC. De hecho, siendo los conceptos subyacentes los mismos, se puede establecer una relación de manera tal que construyendo un IC se pude concluir sobre el rechazo o no, de una Hipótesis Nula.

En términos generales, si se desea realizar un test **bilateral** para la siguiente hipótesis:

$$H_0 : \varphi = \varphi_0$$

puede construirse el IC para el parámetro φ con el nivel de confianza $1-\alpha$ siendo α el nivel de significación del test o sea la probabilidad de cometer error Tipo I.

Luego, si el IC estimado con los datos de la muestra contiene al valor del parámetro supuesto en la hipótesis nula, φ_0, ésta no podrá rechazarse, mientras que si dicho valor no está contenido entre los límites del intervalo, se rechazará la hipótesis formulada.

Es importante destacar aquí que sólo puede utilizarse el intervalo de confianza para testear hipótesis bilaterales (no unilaterales) para el parámetro φ en estudio, con el nivel de confianza $1-\alpha$ del intervalo correspondiente al nivel de significación α del test.

Ejemplo 15

Considere el ejemplo de la Sección 4.2 de este capítulo. Del capítulo anterior, se recuerda que:

$$C\left(\frac{s_X^2}{s_Y^2 F_{(n-1;m-1)\,;\,1-\alpha/2}} \le \frac{\sigma_X^2}{\sigma_Y^2} \le \frac{s_X^2}{s_Y^2 F_{(n-1;m-1)\,;\,\alpha/2}}\right) = 1-\alpha$$

En este caso, X correspondería a la producción actual e Y a la producción con la nueva máquina. Por lo tanto, con los datos del ejemplo tenemos que:

$$0,95 = C\left(\frac{3600}{1871\times 2,3452} \le \frac{\sigma_X^2}{\sigma_Y^2} \le \frac{3600}{1871\times 0,4087}\right)$$

Luego, el IC estimado es $\left[0,8205\,;\,4,7185\right]$.

En la hipótesis el ejemplo anterior se formuló que $\sigma^2_{\text{Nueva}} = \sigma^2_{\text{Actual}}$, o lo que es lo mismo, $1 = \sigma^2_{\text{Actual}} / \sigma^2_{\text{Nueva}}$. Como el número uno que se supone para el cociente en la hipótesis nula está incluido en el IC estimado, no puede rechazarse esta última.

6 Regresión Lineal

Dario Bacchini
Lara Vazquez
Andrea Lepera

En el presente capítulo, se trabaja con dos variables que presentan entre sí una relación en el comportamiento, de manera tal que una de ellas podría ser, en parte, explicada por la otra. Esto resulta útil, por ejemplo, si se estudia la relación entre los precios de dos bienes o la relación que se presenta entre el ingreso que percibe un individuo y su consumo.

Nos limitaremos, en el análisis, a los casos en donde esta relación sea lineal, lo cual implica que la relación entre las variables es constante en todo el rango de posibles valores.

Las regresiones lineales son frecuentemente utilizadas en la estimación de valores dado que permiten estudiar las variables no en forma aislada sino en su interacción con otras.

Asimismo, dado que su desarrollo es simple y puede hacerse utilizando, por ejemplo, una planilla de Microsoft® Excel, su uso es también usual en la determinación de tendencias.

En el presente capítulo, introducimos los fundamentos de la Regresión Lineal y los supuestos que subyacen en la estimación. Consideramos que, más allá de la sencillez en términos de cálculo, conocer el significado de una recta de regresión será útil al lector al momento de determinar su aplicabilidad.

6.1 Regresión Lineal

En los capítulos anteriores, se ha trabajado con variables aleatorias y los distintos tipos de estimaciones al buscar conocer ciertas características de una población a partir del análisis de una muestra. En muchos casos, el análisis requiere de analizar conjuntamente más de una variable dado que, por sus características, existe una asociación entre las mismas. Un caso que podemos considerar es, por ejemplo, el comportamiento del incremento de las ventas de un producto y el dinero destinado a campañas publicitarias. Podríamos pensar que parte del incremento de las ventas de determinado producto se explica por los fondos destinados a publicidad: el incremento de ventas sería la *variable explicada* y el gasto en publicidad la *variable explicativa* o *independiente*. Mediante un análisis de regresión lineal, se propone una función lineal que permita estimar el valor promedio de la demanda a partir del conocimiento del gasto publicitario.

Es importante, en este punto, que conozcamos qué se entiende por función lineal, de manera que podamos distinguir si esta relación resulta o no adecuada para las variables que están siendo objeto de análisis. La existencia de una relación lineal implica que existe un *cambio proporcional constante*: el cambio en la variable explicada es proporcional al cambio en la variable independiente y esta proporción es siempre la misma para todo el rango de posibles valores. ¿Qué queremos decir con esto? Si la variable independiente se modifica en un determinado valor, la variación en la variable explicada será un porcentaje fijo de esa variación, porcentaje que se mantiene constante para todo el rango de posibles valores.

Lo anterior quiere decir que: si cuando $x = x_1$ se da que $y = y_1$, y se modifica el valor de x en Δx, entonces tendremos que $x_2 = x_1 + \Delta x$, resultando $y_2 = y_1 + p \times \Delta x$ será el valor de y correspondiente, donde p es la proporción mencionada. Por ejemplo, $y = 0,5x + 1$ es una función lineal en la cual para cada cambio en x, el cambio en y será igual a la mitad del cambio en la variable independiente ya que la *pendiente* es 0,5: para $x_1 = 1$ se obtiene que $y_1 = 1,5$. Si x crece 0,2 el valor de y debería crecer 0,1 (la mitad), es decir que $x_2 = 1,2$ tendremos que $y_2 = y_1 + 0,1 = 1,5 + 0,1 = 1,6$. Comprobemos que esto es efectivamente así, remplazando $x = 1,2$ en la ecuación original: $y_2 = 0,5 \times 1,2 + 1 = 1,6$.

En general, para $y = a \cdot x + b$, ante un cambio en el valor de x igual a Δx, la variable *y* se modificará en $\Delta y = a \times \Delta x$, siendo a la *pendiente* de la función.

Esto se puede ver fácilmente reemplazando:

$$\begin{aligned} \Delta y &= y_1 - y_0 \\ &= a \cdot x_1 + b - \left(a \cdot x_0 + b\right) \\ &= a \cdot \left(x_1 - x_0\right) \\ &= a \times \Delta x \end{aligned}$$

Notemos aquí que si la pendiente a es positiva, la modificación $\Delta y = a \times \Delta x$ tendrá el mismo signo que Δx, pero si la pendiente a es negativa, dicha modificación tendrá sentido contrario al de Δx.

Luego, ante un aumento del valor de x, si la pendiente a es positiva, el valor de y se incrementaría, mientras que si la pendiente a fuera negativa, el valor de y se reduciría.

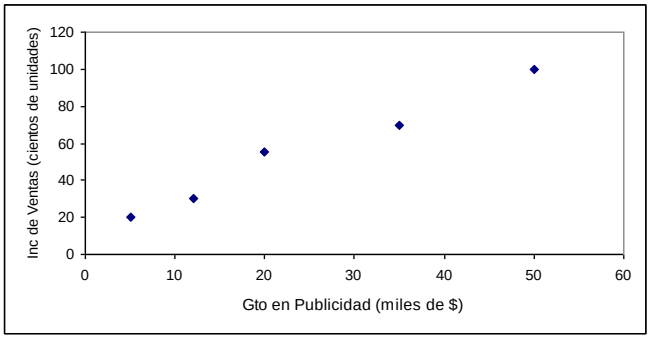
Una vez comprendida la idea de una relación lineal, retomemos el concepto de regresión planteado. La calidad de la función que se adopte como representativa de la relación entre las variables dependerá de las características de las variables y de la muestra con la que se cuente. Claramente, en nuestro ejemplo inicial, la demanda del producto es una *variable aleatoria* y, si lo que queremos es conocerla a partir de una estimación de los gastos para determinado momento, esta última variable deberá considerarse *determinística.*

6.1.1 Concepto

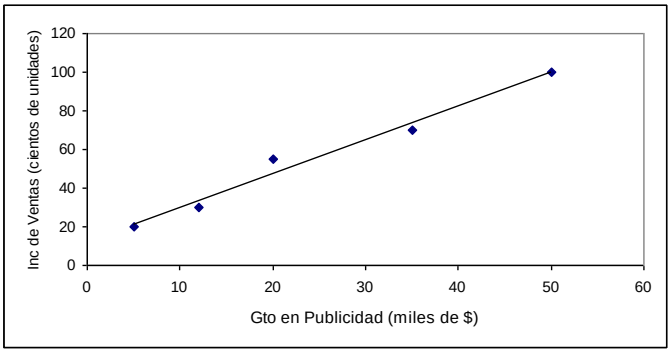
Al trabajar con un análisis de regresión se vinculan dos variables *cuantitativas*. No podemos relacionar variables que no tienen nada que ver entre sí sólo porque los datos con los que contamos pueden hacer suponer una relación lineal. Esa característica no resulta suficiente porque, probablemente, se deba simplemente a la muestra tomada y no a una relación directa entre ellas. La cantidad de vasos de agua que usted toma en el día no parece relacionarse con la cantidad de ventas de una empresa de golosinas. No existe conexión entre esas variables. Sin embargo, si cuento con una muestra determinada podría inclinarme a "inventar" una relación lineal si los datos así me lo permiten. Esto sería erróneo porque no existe ningún sustento que nos permita creer que esa relación va a mantenerse.

Consideremos el ejemplo planteado en la introducción de la relación entre el incremento de las ventas y el gasto en publicidad. Podemos tabular y graficar el resultado si contamos con datos empíricos respecto de esta relación.

Gto Publicidad	Inc. Ventas (u)
5	20
12	30
20	55
35	70
50	100



Una primera impresión de la observación anterior refleja que existe similitud en el comportamiento: al aumentar el gasto en publicidad también se percibe un incremento mayor en la demanda del producto. Podríamos decir que la recta que se muestra a continuación refleja correctamente, *en promedio*, el comportamiento conjunto de las dos variables



En la sección siguiente se tratará uno de los métodos para el cálculo de la recta que ajusta de manera adecuada los datos.

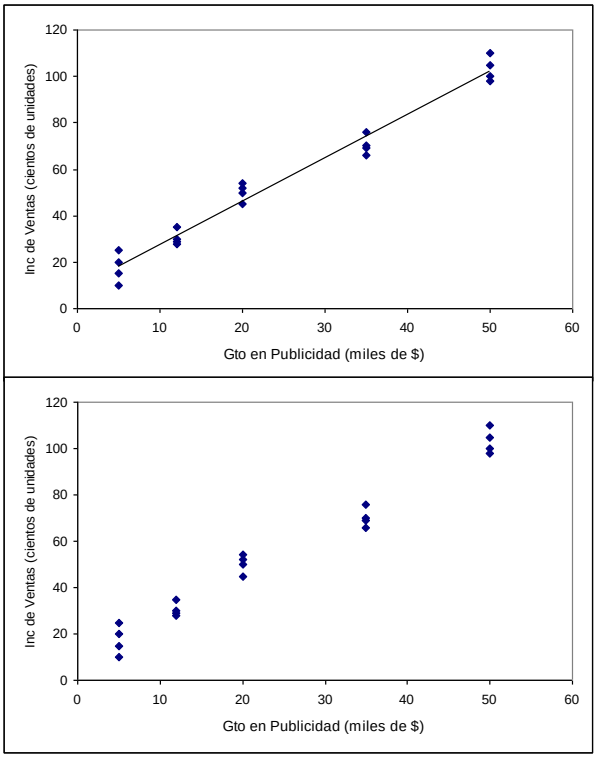
En principio, lo que debe tenerse presente es a qué nos referimos cuando se habla de regresión lineal simple, que es la que trataremos en este capítulo:

En un análisis de regresión se busca estimar la **media** de una variable aleatoria en base a valores predeterminados de otra variable con la cual presenta **dependencia estadística**.

Al hablar de dependencia estadística se pone de manifiesto el carácter no determinístico de esta relación: claramente, si bien podemos observar una relación entre el incremento de las ventas y el gasto en publicidad la misma no se presenta siempre de igual forma dado que existen otros factores que influirán sobre la demanda y que serán ajenos al modelo. Por ejemplo: el tiempo que transcurrió desde la campaña anterior, los precios que maneja la competencia respecto de los propios o un cambio en la preferencia de los individuos, entre otros. Asimismo, la demanda proyectada para un momento dado es una variable aleatoria en sí misma, por lo que la estimación tendrá asociada una probabilidad dada.

El ejemplo anterior lo podemos completar considerando que los gastos en publicidad se repitieron en determinados momentos del tiempo y que, en consecuencia, para cada valor de gasto contamos con distintos resultados en término de incremento de ventas, según lo refleja el cuadro siguiente:

Gto Publicidad	Inc. Ventas A	Inc. Ventas B	Inc. Ventas C	Inc. Ventas D
5	20	15	25	8
12	30	30	30	30
20	55	55	55	55
35	70	70	70	70
50	100	100	100	100



Una visión más clara puede obtenerse de la observación del **diagrama de dispersión** y de la **recta de regresión** asociada a esos datos:

¿Qué estaríamos haciendo al calcular la recta de regresión? Estaríamos considerando que el valor esperado del incremento en las ventas es una *función lineal* del gasto destinado a publicidad:

$$E\left(Inc.Ventas \middle| Gtos\,en\,Publicidad\right) = a_1 \cdot Gtos + a_0$$

Si bien la acepción de linealidad varía según los casos, para nuestro tratamiento, al hablar de regresión lineal simple, estaremos considerando el caso en donde una función lineal se ajuste de manera tal de explicar los valores de una variable en base a otra única variable.

Otros modelos lineales pueden intentar vincular el comportamiento de una variable con los valores que tomen **otras**, es decir, vincularlo con el comportamiento de más de una variable explicativa. En este caso los modelos reciben el nombre de **modelos de regresión múltiple**.

Para el caso del ejemplo que está siendo planteado, el incremento en las ventas puede relacionarse con el monto destinado a publicidad y aquel invertido en el sueldo de los vendedores (considerando que, a mejores condiciones laborales mejor el desempeño de los mismos). Es decir, puede pensarse en:

$$E\left(Inc.Ventas \middle| Gtos\,en\,Publicidad; Salarios\right) = a_2 \times Salarios + a_1 \times Gtos + a_0$$

Más allá de que los factores que influyen sobre el incremento en las ventas no se limitan a estos dos, la descripción parecería ser más completa que la anterior. De cualquier modo, debe tenerse en cuenta el costo que representa la inclusión de cada una de las variables explicativas. Independientemente de lo expuesto, en este libro, nos limitaremos a analizar conjuntamente dos variables y no, los casos en que la regresión incluya más de un factor generador de otros.

Sean X e Y dos variables, de manera tal que Y (variable dependiente) puede explicarse en parte por las realizaciones de X, un análisis de regresión lineal simple implicará encontrar los valores β_1 y β_0 tal que el valor de la variable Y pueda **estimarse a partir de la realización de la variable independiente a través de una función lineal**:

$$y = \beta_1 \cdot x + \beta_0 + \varepsilon$$
$$= E(Y|X = x) + \varepsilon$$

siendo ε una variable aleatoria denominada término de *error estocástico*[56].

El término de error estocástico al cual se hace referencia en la definición tiene razón de ser en el marco de la dependencia estadística de la cual hablábamos. Implica que la estimación no será perfecta, sino que dependerá de la aleatoriedad propia de Y y también de todos los factores que la determinan, más allá de X.

Para que esta regresión sea válida en cuanto a que $E(Y|X = x) = \beta_1 \cdot x + \beta_0$, se deben enunciar ciertas **hipótesis de comportamiento**.

Estas hipótesis, bajo las cuales es válido realizar el método de mínimos cuadrados que se plantea en una sección posterior, son conocidas como las **condiciones de Gauss Markov**:

(1) La esperanza de los errores es igual a cero.

(2) Los errores, para cualquier valor de X, tienen la misma varianza.

(3) Los errores son independientes entre sí.

(4) Los errores tienen distribución normal.

Las hipótesis (1), (2) y (4) pueden resumirse con la siguiente notación:

$$\varepsilon_i \sim N(0; \sigma_\varepsilon)$$

Asimismo, la hipótesis (3) puede expresarse como $E(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i, j \,/\, i \neq j$

Veremos con mayor detenimiento la importancia de estas hipótesis al momento de estimar los parámetros y analizar si el modelo resulta ser o no adecuado para el caso planteado.

Cuando utilizamos un subíndice para las variables nos referimos a que para cada valor de X tenemos asociado un valor distinto de Y:

$$y_i = \beta_1 \cdot x_i + \beta_0 + \varepsilon_i$$

Así, para la primera observación de X (x_1), tendremos asociada la primera observación de la variable Y (y_1) y la primera del error (ε_1). Y así, sucesivamente, para cada observación de X. El hecho de que se observen errores se debe a que en general la recta no es "perfecta". Volveremos sobre este punto más adelante.

Como ya se mencionó, lo que se buscará determinar en este capítulo son los valores de β_0 y β_1 que mejor se "ajusten" a los datos. En el siguiente ejemplo suponemos que se conocen las estimaciones de estos parámetros, y más adelante se verá la técnica para estimarlos.

[56] Entendemos por estocástico a una variable cuyo comportamiento se explica a partir de probabilidades, es decir, está sujeto a aleatoriedad. En otras palabras, estocástico podría pensarse como sinónimo de aleatorio.

Ejemplo 1

Si consideramos el caso de la relación entre el incremento de ventas y el gasto en publicidad, debe entenderse que, para un valor dado del gasto, x, el valor del incremento en las ventas, Y, será una variable aleatoria cuyo valor esperado estará dado por la recta de regresión que estimamos. Si, por ejemplo, $x_1 = 5.000$, y hemos estimado que $\beta_0 = 50$ y $\beta_1 = 0,7$, entonces, el valor esperado del incremento en las ventas será 3.550:

$$\begin{aligned} E\left(Y \middle| X = 5.000\right) &= 5.000 \times \beta_1 + \beta_0 \\ &= 5.000 \times 0,7 + 50 \\ &= 3.550 \end{aligned}$$

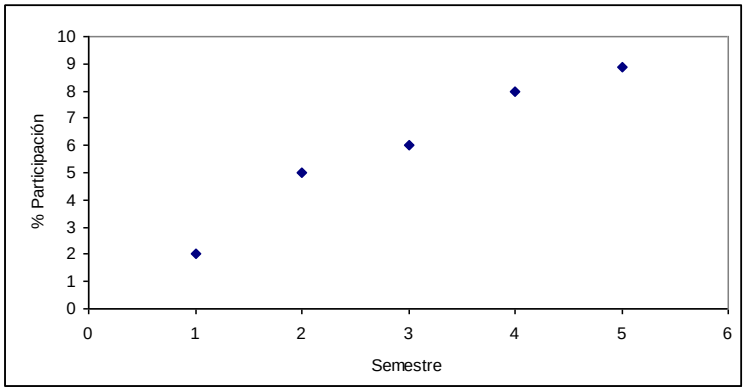
En este ejemplo, la estimación de la pendiente de la función es igual a 0,7. ¿Cómo interpretamos este valor obtenido? Implica que, dada una situación inicial en la relación $(Gasto, Ventas)$, ante un incremento del volumen del gasto en publicidad, las unidades vendidas aumentarán 70% respecto del mencionado incremento del gasto: cada $1000 de incremento (lo cual implica $\Delta x = 1$ al representarse la variable en miles de pesos), las ventas aumentarán en 700 ($\Delta y = 0,7 \times 1$).

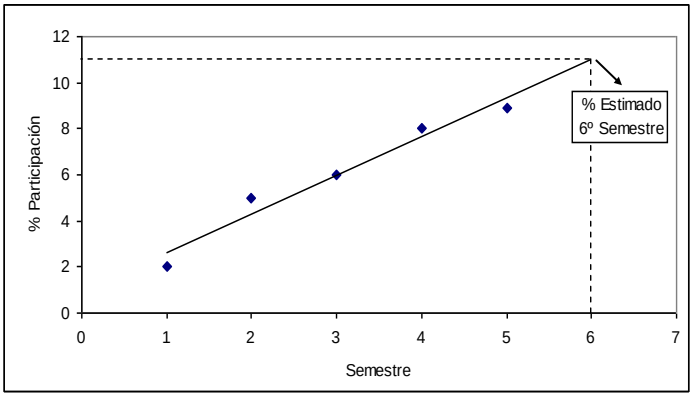
Como se mencionaba párrafos atrás, al momento de relacionar las variables no debe tenerse en cuenta únicamente que los datos muestren un comportamiento similar, sino que, además, sea lógico suponer una relación. Esto es, a pesar de que, podamos ver que hay un movimiento similar entre las ventas de una empresa de llaves con las notas que obtienen los alumnos de un colegio, no puede pensarse en otra cosa más que "*casualidad*". Uno de los casos más frecuentes en donde la asociación tiene sentido es el caso en el cual se relaciona una variable cuantitativa dada (que es aquella que quiere explicarse) con la variable tiempo. De esta manera, se estudia la evolución de la variable a lo largo del tiempo, conformando su trayectoria un *proceso estocástico*. Los procesos estocásticos están fuera del alcance de este análisis, por lo cual sólo consideraremos como ejemplo el tipo de proceso en donde la trayectoria de la variable es función lineal del tiempo, lo cual constituye un *proceso estocástico con tendencia lineal*: $\left|E\left(X_t\right) = a_1 \cdot t + a_0\right|$.

Ejemplo 2

Asumamos la evolución del porcentaje de participación en el mercado de una compañía dada. Consideremos, asimismo, que la misma mide semestralmente dicha participación y quiere estimar su participación para el semestre siguiente, contando con los siguientes datos:

Semestre	% Part.
1	2
2	5
3	6
4	8
5	8,9





Puede observarse una **tendencia al crecimiento** a medida que transcurren los meses. Consideraríamos la recta de regresión, de forma tal que, a partir de ella, pudiéramos estimar la participación para el sexto semestre.

Esta estimación será de la forma: $E\left(Part \middle| X=6\right)=\beta_1 \cdot 6+\beta_0$, siendo los valores estimados $\hat{\beta}_0=0,94$ y $\hat{\beta}_1=1,68$ [57].

Asumiendo una relación lineal, entonces, la participación esperada en el mercado para el siguiente semestre es del 11,02%, según surge del siguiente cálculo y se observa en el gráfico:

$$E\left(Part \middle| X=6\right)=1,68\times 6+0,94$$
$$=11,02$$

Ahora, ¿es coherente suponer esta relación? Es importante tener en cuenta que una regresión lineal, posiblemente, no resultará adecuada para todos los semestres, dado que ella supone que la tasa de crecimiento de la empresa se mantiene constante a lo largo del tiempo y, probablemente, ese supuesto no pueda mantenerse por un período prolongado. En este caso, dependerá también del analista realizar las consideraciones pertinentes.

Así como en el ejemplo anterior, la aplicabilidad de un modelo lineal podía discutirse al implicar que el crecimiento en el mercado se da a tasa constante, en todos los casos en los cuales se estime a través de este tipo de función debe considerarse el impacto de asumir la tasa de variación constante (a cada valor Δx le corresponde la misma variación en Y, es decir, le corresponde un único valor Δy).

Las relaciones lineales serán válidas si puede suponerse que un cambio en la variable independiente de Δx genera el mismo cambio en la variable explicada, cualquiera sea el valor de x inicial y el momento del tiempo en el cual nos encontremos. Esto no siempre es así. Si llamamos Y al consumo en alimentos de una persona y X al nivel de ingreso, podemos fácilmente suponer que a medida que aumenta X, también aumenta Y (no sólo por cantidad, sino también por la calidad de los alimentos consumidos). Ahora, para determinados valores del ingreso, los datos muestrales pueden llevarnos a suponer una relación lineal. Sin embargo, si expandimos el rango de ingresos, probablemente el impacto que genere un incremento de \$100 ($\Delta x=\$100$) cuando la persona recibe \$900 no es el mismo que cuando recibe \$3000. Esperaríamos que el cambio en consumo de alimentos sea mayor si se parte de un salario de \$900 ($x=\$900$), por ejemplo, un incremento de consumo de \$70 ($\Delta y=\$70$). Sin embargo, una persona con salario de \$3000 ($x=\$3000$) tendrá satisfechas sus necesidades de consumo y probablemente destinará el incremento de salario a otros bienes, aumentando su consumo en alimentos en sólo \$20 ($\Delta y=\$20$). La tasa de variación no es constante porque al mismo valor de Δx le corresponden distintos valores de Δy, según sea el valor de x inicial.

[57] Tal como anticipamos, el modo de cálculo se considerará en la sección siguiente.

¿Qué queremos mostrar con el ejemplo anterior? Que no toda relación directa (como en este caso, sube X baja Y) puede mostrarse a partir de una relación lineal, porque la tasa de variación puede ser distinta según cual sea el valor de X considerado.

6.1.2 Covarianza y Correlación entre las Variables

De la forma en la que definimos la regresión lineal (en el acápite anterior), es claro que para poder hablar de un análisis de regresión es necesario que se observe una relación entre las variables. Ahora, lo que podríamos plantearnos es: ¿de qué manera se evidencia esta relación? ¿cómo puede medirse? Los párrafos siguientes nos acercarán un poco a la respuesta.

Al hablar de la probabilidad conjunta de dos eventos (ver Capítulo 2), hemos considerado el caso en donde existe independencia entre dos variables aleatorias. Cuando nos referimos a independencia entre dos variables (a las cuales podemos llamar X e Y), consideramos que la probabilidad de que la variable Y tome un valor particular no se ve afectada por el valor que presente la otra variable en consideración y, al mismo tiempo, también implica que la probabilidad de que X tome un valor particular se mantendrá constante independientemente de la realización de la variable Y. En este capítulo, por el contrario, buscamos detectar cierta *dependencia* entre las variables; en particular, buscamos encontrar evidencia de **dependencia lineal** entre las mismas de manera de que sea válido efectuar un análisis de regresión.

Covarianza

Las medidas más importantes que nos proporcionan evidencia de **dependencia lineal** son la **Covarianza** y el **Coeficiente de Correlación**. En ambos casos, dados los valores de las variables, lo que nos interesa es ver el comportamiento en simultáneo de éstas, con lo cual tendremos que utilizar la **distribución de probabilidad conjunta**.

A fin de interpretar la definición de las medidas antes mencionadas, tendremos en cuenta la siguiente notación: la probabilidad conjunta de dos variables se escribe como $P_{X,Y}(x,y)$ siendo $P_{X,Y}(x,y) = P(X = x, Y = y) = P(X = x \cap Y = y)$. En este caso, cada uno de los eventos que se considera es el par ordenado (x,y), por lo cual el espacio *muestral está dado por todos los posibles pares* (x,y). Lo que nos interesa conocer, entonces, es la probabilidad de que se den de manera simultánea los dos eventos, de manera tal que pueda verificarse si hay una relación entre el valor que toma X y el valor que toma Y.

En el caso de variables continuas, al igual que se utiliza el concepto de función de densidad de probabilidades en una variable, se emplea para la distribución conjunta el concepto de función de densidad de probabilidad continua $f_{X,Y}(x,y)$, pero esto no será desarrollado en profundidad aquí, pues requiere conceptos de análisis matemático con dos variables.

Ejemplo 3

Sea X la variable aleatoria que representa el resultado del lanzamiento de un dado y sea Y aquélla que representa el número de una carta seleccionada al azar de entre dos (con números 1 y 2). Si quiere analizarse de manera conjunta los dos resultados, es decir, la función de probabilidad conjunta, el espacio muestral estará dado por:

$$\Omega = \{(1,1);(1,2);(2,1);(2,2);(3,1);(3,2);(4,1);(4,2);(5,1);(5,2);(6,1);(6,2)\}$$

En este caso, recordando los conceptos de cálculo vistos en el Capítulo 2, cada uno de los eventos tiene una probabilidad igual a $1/12$.

Nos encontramos entonces ya en condiciones de definir a la covarianza entre dos variables aleatorias:

Sean X e Y dos variables aleatorias cuyas medias son, respectivamente, μ_X y μ_Y. La **covarianza** de X e Y se define como:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

También puede escribirse como:

$$Cov(X,Y) = E(XY) - \mu_X \cdot \mu_Y$$

A continuación, indicamos cómo se aplica esta definición al caso de variables discretas y al caso de variables continuas.

En el caso en que X e Y son variables discretas con probabilidad conjunta $P_{X,Y}(x,y)$ la **covarianza** es:

$$Cov(X,Y) = \sum_x \sum_y (x-\mu_X)(y-\mu_Y) P_{XY}(x,y)$$

También puede escribirse como:[58]

$$Cov(X,Y) = \left[\sum_x \sum_y x \cdot y \cdot P_{XY}(x,y) \right] - \mu_X \cdot \mu_Y$$

En el caso en que X e Y son variables aleatorias continuas y $f_{X,Y}(x,y)$ es la función de densidad conjunta, la **covarianza** se calcula como la integral doble:

$$Cov(X,Y) = \int_x \int_y (x-\mu_X)(y-\mu_Y) f_{XY}(x,y)\, dy\, dx$$

Dado que mediante el análisis de la covarianza buscamos encontrar evidencia de correlación entre variables a partir de los datos observados, en general, el análisis que hacemos es el de **covarianza muestral**, remitiéndonos al cálculo de la misma mediante:

$$Cov(X,Y) = \frac{1}{N} \sum_{i=1}^{N} x_i \cdot y_i \cdot n(x_i, y_i) - \bar{x} \cdot \bar{y}$$

siendo N el tamaño de la muestra (número de pares observado), $n(x_i, y_i)$ el número de pares (x_i, y_i) que se observan en la misma y $\bar{x}$ e $\bar{y}$ las respectivas medias muestrales.

¿Cómo puede interpretarse esta medida? Si se observa la primera de las expresiones vemos que, para cada par de posibles valores de X e Y, se realiza una suma ponderada por la probabilidad de ocurrencia del producto de la distancia de cada una de esas realizaciones respecto de su media. ¿Qué nos va a indicar el resultado de esa suma?

Si cuando la desviación de una variable respecto de su media es positiva la de la otra variable también lo es, el producto de estas desviaciones es positivo. Lo mismo sucede en caso de que las dos desviaciones sean menores a cero. Estos signos no se ven afectados con la ponderación dado que la probabilidad, por definición, es siempre de signo positivo. Asimismo, puede darse el caso en el cual el signo de las desviaciones no sea el mismo. En consecuencia, el producto de las mismas es negativo.

El análisis de signos anteriores nos dará una pauta de la relación lineal. Para hablar de que existe una relación lineal, claramente, las variables deben mantener la misma relación en todos sus valores, sea esta relación directa (se mueven en el mismo sentido) o inversa (se comportan en sentido contrario). En consecuencia, si la suma ponderada de los desvíos es **positiva** (y significativa), podemos, en principio, asumir una **relación lineal directa** dado que en la mayor

[58] Ver deducción en el Apéndice.

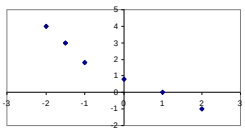
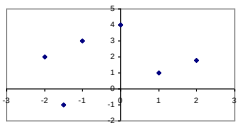
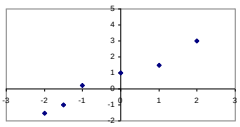
parte de los casos el sentido en el cual las variables se mueven es el mismo. Por el contrario, si la suma de los mismos es **negativa**, puede considerarse una **relación lineal inversa**.

¿Qué significa que la covarianza sea igual a cero? Para que esto suceda, el signo de cada uno de los términos de la sumatoria no debe ser constante, sino que, por el contrario, se alterna dependiendo del par que se considere. Eso significa que, para determinados valores, las variables se mueven en el mismo sentido mientras que para otros lo hacen en sentido contrario. En consecuencia, no podría hablarse de una relación lineal. Se dice, entonces, que:

Si la covarianza de dos variables aleatorias es **igual a cero**, entonces, las variables **no** presentan una **relación lineal** entre ellas. En este caso se dice que las variables son **no correlacionadas** o **incorrelacionadas**.

Los gráficos siguientes ejemplifican lo presentado en los párrafos anteriores. En el primer caso, vemos que, en caso de ajustar la relación que muestran los puntos mediante una recta, la misma tendría pendiente negativa: a medida que aumenta el valor de X disminuye el valor de Y; la relación es inversa y la covarianza negativa. Lo contrario ocurre en el caso del tercer gráfico: vemos que a mayor valor de X, también se manifiesta un mayor valor de Y. Así, la recta que puede ajustarse debería tener pendiente negativa y la covarianza manifiesta esta relación tomando valores mayores a cero. En el gráfico también observamos que, cuando la covarianza es igual a cero, no podemos hablar de la existencia de una relación lineal dado que la variación de Y ante variaciones de X no sólo no es constante, sino que tampoco mantiene el mismo signo para todos los posibles valores de la variable independiente: del primer punto al segundo vemos que aumenta X y disminuye Y ($\Delta x > 0$ y $\Delta y < 0$), mientras que al pasar del segundo al tercero, X aumenta y también lo hace Y ($\Delta x > 0$ y $\Delta y > 0$).

Cabe destacar aquí que dos variables X e Y que son estocásticamente independientes (es decir que $P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$) no tendrán relación líneal, resultando por lo tanto no correlacionadas. Sin embargo el concepto de "incorrelación" no es equivalente al de "independencia". Dos variables incorrelacionadas pueden tener una dependencia a partir de una relación no lineal entre ambas (por ejemplo una relación sinusoidal o exponencial, etc.).

		
Relación lineal inversa $Cov(X,Y) < 0$	Ausencia de relación lineal $Cov(X,Y) = 0$	Relación lineal directa. $Cov(X,Y) > 0$

Ejemplo 4

La tabla que se expone muestra la evolución de los precios de ajuste que se presentó en dos contratos de futuro distintos (Trigo Enero 2007 y Trigo Enero 2008) durante la semana del 6 al 11 de noviembre de 2006[59]. ¿Puede plantearse la existencia de correlación lineal? Calcule la covarianza.

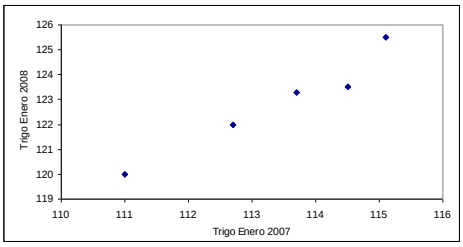
Fecha	Trigo Enero 08	Trigo Enero 07
06/11/2006	111	120
07/11/2006	112,7	122
08/11/2006	113,7	123,3
09/11/2006	115,1	125,5
10/11/2006	114,5	123,5

[59] Fuente: DataCenter del Mercado a Término de Buenos Aires (www.matba.com.ar)

Lo que calcularemos en este ejemplo es la covarianza de esta muestra en particular.

Si se observa el gráfico, esperaríamos que la misma nos dé evidencia de una relación lineal directa: el precio de los contratos se mueve en el mismo sentido y podríamos explicar el comportamiento de uno de ellos en base al del otro.

Dado que trabajamos a partir de una muestra y no conocemos la distribución de probabilidades, debemos recurrir al cálculo a partir de las frecuencias observadas para cada par. En principio, debemos calcular la media muestral de la variable *"TrigoEnero08"* (Y) y la de *"TrigoEnero07"* (X), siendo respectivamente:



$$\overline{x} = \frac{120+122+112,3+125,5+123,5}{5} = 122,86$$

$$\overline{y} = \frac{111+112,7+113,7+115,1+114,5}{5} = 113,4$$

Ahora, debemos sumar el producto de todos los posibles pares por su probabilidad. Según la fórmula de cálculo expuesta, N es la cantidad de pares observados, con lo cual $N = 5$ y, dado que cada par lo observamos una vez, la frecuencia es de 1 para todos los pares.

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} x_i \cdot y_i \cdot n(x_i, y_i) &= \frac{1}{5}\cdot(111\cdot 120+112,7\cdot 122+113,7\cdot 123,3+ \\ &\quad +115,1\cdot 125,5+114,5\cdot 123,5) \\ &= \frac{69674,41}{5} \\ &= 13934,882\end{aligned}$$

Entonces,

$$\begin{aligned}Cov(x,y) &= 13934,882 - 122,86\cdot 113,4 \\ &= \boxed{2,558}\end{aligned}$$

El valor positivo de la covarianza nos presenta evidencia de una relación lineal positiva.

Ejemplo 5

Por otro lado, podemos volver a considerar nuestro ejemplo de los dados y las cartas (Ejemplo 3). Trabajamos, entonces, con variables aleatorias discretas cuya distribución conocemos. En este caso, las esperanzas de cada una de las variables son:

$E(X) = \frac{1+2+3+4+5+6}{6} = 3,5$, al ser $1/6$ la probabilidad de obtener cada uno de los posibles valores del dado

$E(Y) = \frac{1+2}{2} = 1,5$, siendo 1 y 2 los posibles valores de las cartas, con probabilidad igual para cada uno de ellos

Continuando con el cálculo para hallar la covarianza:

$$\sum_x \sum_y x \cdot y \cdot P_{XY}(x,y) = \frac{1}{12} \cdot (1+2+2+4+3+6+4+8+5+10+6+12)$$
$$= \frac{63}{12} = 5,25$$

$$Cov(X,Y) = 5,25 - 3,5 \cdot 1,5$$
$$= 0$$

Con lo cual se refleja la **incorrelación** de las variables.

Es muy importante volver a mencionar que **el hecho de que la covarianza sea cero no implica independencia** entre las variables dado que puede existir entre ellas alguna otra relación distinta de la lineal. Esto se verá con más detalle más adelante, cuando analicemos el coeficiente de correlación de Pearson. Puede resumirse, entonces, que:

Si dos variables son independientes, su covarianza es igual a cero[60] pero que la covarianza sea cero no significa que las mismas sean independientes.

Un nuevo problema aparece cuando queremos comparar, entre pares de valores de variables distintas, cuáles presentan entre sí una mayor relación lineal. En principio, nos veríamos tentados a creer que, cuanto mayor la covarianza, mayor también la relación. Sin embargo, esto no es así dado que, si recordamos la definición, el valor de la covarianza dependerá del valor de los desvíos de cada variable respecto de su valor medio y las escalas que se manejan respecto a estos desvíos son distintas en función del significado de la variable (un desvío de 10 respecto de la media al hablar de salarios mensuales parece ser bajo, no así si consideramos un desvío de 10 respecto de la media en el precio de una remera). En consecuencia, se hace necesario contar con una medida del **grado de relación lineal**, es decir, una medida que nos permita conocer qué variables están más relacionadas que otras. Presentamos, entonces, el **coeficiente de correlación**, el cual nos acercará a este objetivo.

Coeficiente de Correlación

El coeficiente de correlación (o coeficiente de *Pearson*) está relacionado con el cálculo de la covarianza. La ventaja de éste, respecto de la medida anterior, radica en que el coeficiente de correlación puede tomar valores únicamente dentro del rango $[-1\,;1]$.

Esta delimitación del rango nos permite establecer qué tan fuerte es la relación entre las variables: cuando el coeficiente es igual a 1 o a -1 se dice que la correlación entre ambas es perfecta mientras que si es igual a cero no existe correlación de tipo lineal. En general, podemos decir que cuando el coeficiente de correlación sea, en módulo, más cercano a 1, más fuerte será la relación de linealidad entre ambas.

Sean X e Y dos variables aleatorias discretas cuyos desvíos estándar son, respectivamente, σ_X y σ_Y. Su **correlación**, la cual se simboliza como ρ_{XY}, se define como:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

De la definición se observa que: si $Cov(X,Y) = 0 \Rightarrow \rho_{XY} = 0$. Asimismo, dado que, por definición, el desvío estándar es siempre mayor a cero, el signo del coeficiente de correlación está dado por el signo de la covarianza. Como ya hemos mencionado, este signo depende de la forma en la que se manifieste la relación.

[60] Ver Apéndice para la demostración de la afirmación anterior

Cuando se estima a partir de una muestra este coeficiente, se obtiene el **Coeficiente de correlación muestral**

$$\hat{\rho}_{XY} = \frac{\widehat{cov}(X,Y)}{s_X s_Y}$$

Ejemplo 6

Podemos retomar ahora el ejemplo en el cual tratamos la evolución de los precios de ajuste de los contratos de futuro. La covarianza, ya calculada, nos dio un valor igual a 2,558. Para calcular el coeficiente de correlación necesitamos, además, considerar el valor de los desvíos. Realizando el cálculo de la manera vista en el Capitulo 3, se obtiene que:

$$\sigma_{TE07} = 1,8161 \qquad \sigma_{TE08} = 1,445$$

En consecuencia,

$$\rho_{TE07,TE08} = \frac{2,558}{1,8161 \cdot 1,445} = 0,9747$$

Este valor del coeficiente de correlación muestra que el grado de correlación lineal es alto.

Ejemplo 7

X	Y
-2	-1,5
-1,5	-1
-1	-0,12
0	1
1	1,6
2	3

Tomemos en cuenta, a continuación, los valores de la tabla que incluimos. ¿Presentan una menor o una mayor relación que los contratos de futuro?

Comencemos calculando la Covarianza muestral de estas variables, para ello, necesitamos previamente el valor de las medias muestrales:

$$\bar{x} = \frac{-2-1,5-1+0+1+2}{6} = -0,25 \text{ y } \quad \bar{y} = \frac{-1,5-1-0,12+1+1,6+3}{6} = 0,49\hat{6}$$

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} x_i \cdot y_i \cdot n(x_i, y_i) &= \frac{1}{6} \times \left((-2)\cdot(-1,5) + (-1,5)\cdot(-1) + (-1)\cdot(-0,12) + 0,1 + 1,6\cdot 1 + 3\cdot 2\right) \\ &= \frac{12,22}{6} \\ &= 2,03\hat{6}\end{aligned}$$

Por lo tanto,

$$Cov(X,Y) = 2,03\hat{6} - (-0,25) \times 0,49\hat{6} \cong 2,1608$$

La covarianza es menor a la que se presenta para los contratos de futuro, ¿nos indica esto una menor presencia de relación entre las variables?

Como ya dijimos, esto no es necesariamente así y para sacar conclusiones necesitamos conocer el coeficiente de correlación y compararlo con el obtenido para los contratos de futuro.

De acuerdo a la forma de cálculo vista en el Capítulo 2, los desvíos estándares para X y para Y son, respectivamente: $\sigma_X = 1,4068 \quad \sigma_Y = 1,5462$

El Coeficiente de correlación para X e Y es:

$$\rho_{X,Y} = \frac{2,1608}{1,4068 \cdot 1,5462} \cong 0,9934$$

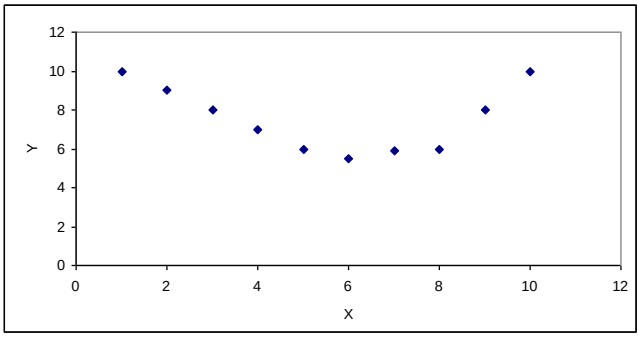
Vemos entonces que, a pesar de que la covarianza es menor que la que se presenta en los contratos de futuro, la correlación entre las variables X e Y es mayor que la que se observa entre Trigo Enero 07 y Trigo Enero 08.

Es importante mostrar que estas medidas (tanto la covarianza como la correlación) son representativas de la dependencia lineal entre variables, por lo cual, que la covarianza sea cero no implica que las variables son independientes, sólo evidencian que no hay una relación *lineal* entre ellas. Lo mismo sucede respecto de la correlación.

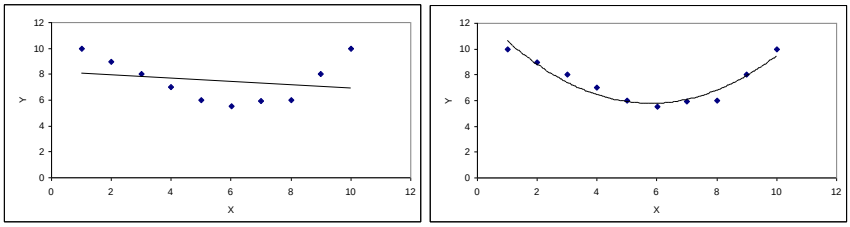
Ejemplo 8

Consideremos los valores de estas dos variables que se muestran en el eje de coordenadas.

X	Y
1	10
2	9
3	8
4	7
5	6
6	5,5
7	5,9
8	6
9	8
10	10



Si calculamos el coeficiente de correlación, tal como lo indica la fórmula anterior, éste toma un valor de -0.22 , el cual no indicaría que la relación entre las variables sea lineal. Esto es evidente si intentamos ajustar una función lineal a estos datos (ver figura). Sin embargo, tampoco parece que exista independencia entre las variables, dado que el comportamiento de la serie parece seguir la forma de una parábola, tal como también se observa en la figura.



En este ejemplo, entonces, puede verse claramente que el hecho de que dos variables no posean dependencia lineal **no implica su independencia**. De cualquier manera, en los términos de nuestro análisis, un grado de correlación poco significativo indicará que el modelo lineal no resulta adecuado.

Significatividad de la Correlación

Al momento de calcular el coeficiente de correlación debe mantenerse presente que estamos trabajando sobre una muestra y que, en consecuencia, a partir de ella estamos intentando inferir las características de toda la población. Por lo tanto, en casos en donde el coeficiente de correlación de dos variables es muy cercano a cero, podríamos dudar de la existencia de una relación lineal entre ellas. Para ello, se construye un test de hipótesis denominado **Test de correlación** en el cual las hipótesis que se manejan son:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Trabajamos, entonces, con una prueba de dos colas en donde, al rechazar la hipótesis nula, mantenemos la posibilidad de que la relación lineal exista. Por el contrario, un no rechazo de la hipótesis nula refleja que la evidencia empírica no es suficiente para afirmar la existencia de una relación lineal entre las variables.

El estadístico de prueba que se utiliza es el siguiente, el cual tiene una distribución t de Student con $n-2$ grados de libertad:

$$t = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t_{n-2}$$

Para una probabilidad α de cometer error de tipo I (rechazar la hipótesis nula cuando ésta es verdadera) los valores críticos de la región de aceptación serán: $t_{n-2;\,1-\alpha/2}$ y $t_{n-2;\,\alpha/2} = -t_{n-2;\,1-\alpha/2}$.

Es importante mencionar que, en caso de que el resultado de este test nos lleve a considerar que $\rho = 0$, entonces, estaría considerando que no existe relación lineal entre las variables. En consecuencia, no tendrá sentido realizar una estimación de ese tipo. Por lo tanto, ante cualquier análisis de variables, antes de iniciar la estimación de los coeficientes a partir del método de mínimos cuadrados debemos verificar que el coeficiente de correlación sea significativo (caso contrario, correremos el riesgo de estar trabajando en la estimación inútilmente, es decir, cuando la misma en realidad no tiene sentido).

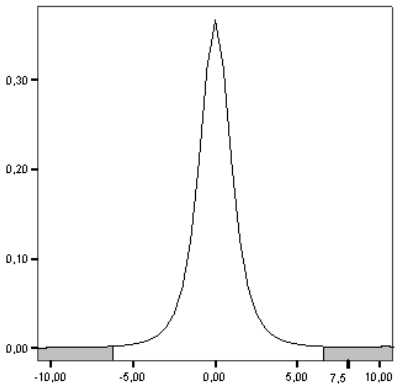
Ejemplo 9

Para el tratamiento de los precios de los contratos de futuro, estimamos en el ejemplo 10.5 un coeficiente de correlación igual a 0,9747 a partir de una muestra de $n = 5$.

Si queremos testar su significatividad con un nivel del 99% de confianza, tendremos que $\alpha = 0,01$ y que el valor crítico $t_{0.995} = +5,841$ ($g.l. = 3$)

El estadístico de prueba t, toma el valor: $t_e = \frac{0,9747 \cdot \sqrt{3}}{\sqrt{1-0,9747^2}} = 7,553$

Como $t_e > +t_{0,995} \Rightarrow$ no se acepta H_0, es decir, podemos seguir sosteniendo la existencia de que existe relación lineal entre las variables.



Ejemplo 10

Consideremos ahora el caso del ejemplo 10.6, en el cual obtuvimos un valor muy pequeño del coeficiente de correlación. Los grados de libertad que manejamos en este caso son 8, dado que la muestra es de 10 observaciones. Con un nivel de confianza igual al del ejercicio anterior: $t_{0,995} = +3,355$

$$t_e = \frac{-0,2222 \cdot \sqrt{8}}{\sqrt{1-0,2222^2}} = -0,6447$$

En este caso, $-3,355 < t_e < 3,355$, con lo cual la evidencia empírica nos hace aceptar la hipótesis nula y no considerar la existencia de una relación lineal entre las variables.

6.2 Estimación de la Recta de Regresión

En la sección anterior, se introdujo el concepto de regresión lineal y de medidas que resultan representativas de la existencia de esta relación. En los puntos siguientes, se expondrá el *Método de Mínimos Cuadrados*, de manera tal de poder estimar cuál es la recta que mejor ajusta a los datos. Asimismo, se analizará el nivel de confianza asociado a estas estimaciones dado que no debe perderse de vista que, al realizar estimaciones, las mismas se construyen sobre la base de una muestra y, en consecuencia, la calidad de la estimación dependerá no sólo de la metodología utilizada sino también de la calidad de la muestra.

6.2.1 Estimación Puntual por Mínimos Cuadrados

Dadas dos variables aleatorias entre las cuales se evidencia algún grado de dependencia lineal, el problema de la regresión lineal consiste en estimar los parámetros de la función lineal que mejor representen esta relación.

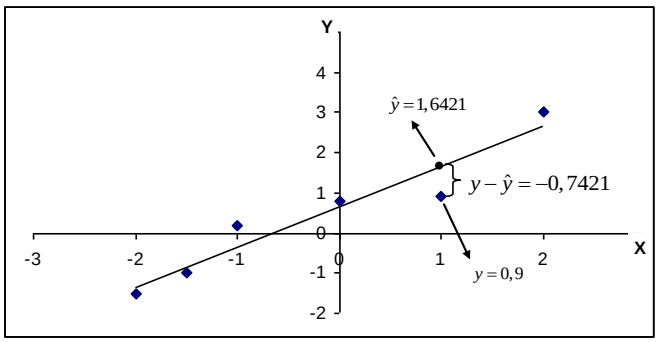
Recordemos que la recta de regresión es de la forma:

$$E(Y_i|x_i) = \beta_1 \cdot x_i + \beta_0$$

Esto implica que, para cada valor que tome la variable independiente (este valor es ya una realización de la variable aleatoria o se encuentra predeterminado), el valor medio de la variable explicada puede conocerse a partir de la función descripta.

Cuando se trabaja con la estimación por mínimos cuadrados, los parámetros β_1 y β_0 se estiman de modo tal que se **minimice la suma de los cuadrados de las desviaciones de las observaciones respecto de la recta**.

Analicemos, entonces, la expresión anterior. Para trazar una recta de regresión necesitamos contar con n observaciones de los pares (x, y) . A partir de ellos, construimos la recta. Ahora, salvo que exista correlación perfecta entre estos datos, la recta no va a pasar exactamente por los n pares que se estén considerando para formarla. Esto es, dado el par (x_i, y_i) , muy posiblemente $\beta_1 \cdot x_i + \beta_0 \neq y_i$. Si llamamos a la estimación $\beta_1 \cdot x_i + \beta_0 = \hat{y}_i$, se evidencia que $y_i \neq \hat{y}_i$ para determinados valores. Este razonamiento se observa en el gráfico siguiente. Podemos ver que, para $x = 1$ se observa que $y = 0,9$ (dato con el cual se inició la estimación). Ahora, utilizando el análisis de regresión se llega a que $E(Y_i|x_i) = 0,9937 \cdot x_i + 0,6484$. Esto nos lleva a considerar que $E(Y|x=1) = 0,9937 + 0,6484 = 1,6421$. Se observa, entonces, lo postulado: $y_i \neq \hat{y}_i$ ($0,9 \neq 1,6421$). La diferencia entre estos valores constituye el error observado que, tal como se muestra en el gráfico, es de 0,7421, siendo la estimación por exceso.



La idea que se mantiene, entonces, para estimar los parámetros es elegir (β_0, β_1) de manera tal de minimizar la expresión $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, siendo n el número de observaciones con las que se cuenta.

En este contexto, vemos que $(y_i - \hat{y}_i)$ representa el error en la estimación, conforme a los valores observados. Este error lo elevamos al cuadrado dado que, en el caso contrario, se compensarían los desvíos positivos con los negativos y no podríamos medir en qué grado resulta adecuado el ajuste.

La expresión de la sumatoria puede reescribirse reemplazando $\hat{y}_i$ por la expresión funcional buscada:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_1 \cdot x_i - \beta_0)^2$$

Debemos tener presente que los datos con los que se cuentan son los pares (x_i, y_i) y que, por lo tanto, las variables de la ecuación a minimizar son β_1 y β_0.

Recordemos que una condición necesaria para encontrar el mínimo de la función es que su derivada primera respecto a las variables de interés sea igual a cero. En expresiones del tipo planteado, esta condición asegura la existencia de un mínimo en ese punto. En consecuencia, si se procede a minimizar la expresión de la sumatoria anterior, se obtienen los estimadores mínimo-cuadráticos de los coeficientes.

Cuando se realiza una regresión lineal del tipo $y = \beta_1 \cdot x + \beta_0 + \varepsilon$, se estiman los parámetros β_0 y β_1 minimizando los errores al cuadrado:

$$\min \sum_{i=1}^{n} \left(Y_i - \beta_1 \cdot X + \beta_0 \right)^2$$

Los **estimadores puntuales** que se obtienen son[61]:

$$\hat{\beta}_1 = \frac{n \cdot \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} X_i^2 \cdot \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i \cdot Y_i}{n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2} = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

Ejemplo 11

Tomemos los datos de la relación entre el incremento de las ventas y el gasto en publicidad, de manera tal de calcular la recta de regresión que mejor se ajuste a los mismos. La variable *"gasto"* será representada por X y la variable "*incremento de las ventas*" por Y.

En la tabla siguiente resumimos los cálculos necesarios para obtener los estimadores de los coeficientes. En la última fila, se expone la suma de cada una de las columnas.

	x	y	x^2	x·y
	5	20	25	100
	12	30	144	360
	20	55	400	1100
	35	70	1225	2450
	50	100	2500	5000
Suma	122	275	4294	9010

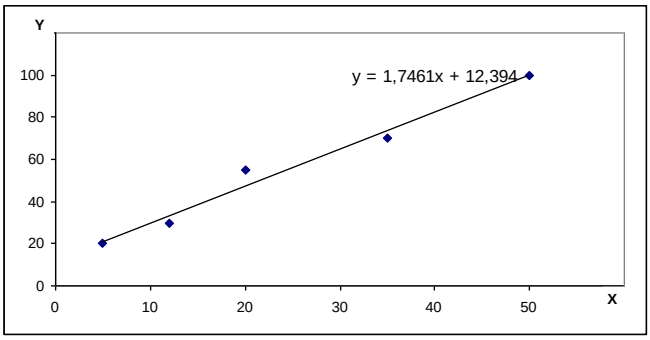
De acuerdo a lo deducido, entonces,

$$\hat{\beta}_1 = \frac{n \times \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \times \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} = \frac{5 \times 9010 - 275 \times 122}{5 \times 4294 - 122^2} = 1,74612$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \times \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i \times y_i}{n \times \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} = \frac{4294 \times 275 - 122 \times 9010}{5 \times 4294 - 122^2} = 12,3945$$

La recta de regresión es de la forma: $\hat{y}_i = 1,74612 \times x_i + 12,3945$

[61] El desarrollo se expone en el Apéndice



A partir de conocer esa recta, podríamos estimar cuál sería el incremento de las ventas si se destinan $30.000 a gastos de publicidad, simplemente reemplazando por estos valores en la ecuación:

$$E\left(Y \middle| X = 30\right) = 1,74612 \cdot 30 + 12,3945 = 64,7783$$

Dado que la regresión está realizada para miles de pesos y cientos de unidades, esto indica que podríamos esperar un incremento de 64.778 unidades si destinamos $30.000 a gastos de publicidad.

Es importante tener en cuenta que lo que hallamos mediante esta metodología y, a partir de la muestra, es una **estimación puntual** de los coeficientes, la cual surge a partir de una muestra particular. Uno de los problemas que pueden plantearse al realizar este tipo de regresión es la presencia de puntos de gran influencia o apalancamiento, que son *valores atípicos* en la relación que parece presentarse de acuerdo a los demás pares que conforman la muestra, generando un importante cambio en la pendiente de la recta estimada. No analizaremos en este caso el tratamiento de estos puntos, pero nos parece importante destacar la importante influencia de la muestra ante la presencia de valores atípicos.

6.2.2 Intervalos de confianza para los coeficientes

Tal como se mencionó en la sección anterior, los coeficientes que surgen de minimizar la suma de los desvíos son estimadores de los verdaderos coeficientes. En consecuencia, tienen asociada una función de distribución y podemos generar intervalos de confianza para los mismos.

Antes de iniciarnos en la distribución de los coeficientes recordemos que, de acuerdo con las hipótesis ya adoptadas:

$$y_i = \beta_1 \cdot x_i + \beta_0 + \varepsilon_i$$
$$= y_i + \varepsilon_i$$

donde, $\varepsilon_i \sim N(0; \sigma_\varepsilon)$ y $E\left(\varepsilon_i, \varepsilon_j\right) = 0 \quad \forall i, j \,/\, i \neq j$

Asimismo, puede demostrarse que los estimadores mínimo-cuadráticos de los coeficientes son **insesgados,** de **varianza mínima** y tienen **distribución normal**. Podemos, además, deducir la fórmula de la varianza para cada uno de ellos. Ambas demostraciones exceden el alcance de esta obra, el lector interesado puede consultar Gujarati (2004)

La varianza correspondiente a cada uno de los estimadores es, respectivamente:

$$Var\left(\hat{\beta}_0\right) = \sigma_\varepsilon^2 \cdot \left(\frac{1}{n} + \frac{\left(\bar{X}\right)^2}{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 \cdot \frac{1}{n}} \right)$$

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 \cdot \frac{1}{n}}$$

Si analizamos las expresiones anteriores de la varianza de los estimadores, puede verse que cuanto menor sea la varianza del error, σ_ε^2, más preciso será el estimador del coeficiente (menor será su error estándar). Asimismo, cuanto mayor es el tamaño de la muestra, n, también serán menores los desvíos de los coeficientes: una vez más, el trabajar con muestras mayores nos permite trabajar, también, con mayor precisión.

Hemos dicho que en ambos casos la varianza depende de la varianza del error σ_ε^2. El error es el residuo de la estimación, es decir, la diferencia entre el valor estimado de la variable explicada para un determinado valor de la variable independiente y el valor observado para ese caso.

Un nuevo problema que aparece entonces es el de estimar la varianza del error dado que, si bien podemos asumirla para un ejercicio dado, en la práctica no es un dato y debemos proceder a su cálculo.

El estimador de la varianza del error es de la forma siguiente:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n-2}$$

El numerador del estimador anterior es la suma de los residuos al cuadrado, dado que:

$$\left(y_i - \hat{y}_i\right)^2 = \left(\hat{\beta}_1 \cdot x_i + \hat{\beta}_0 + \varepsilon_i - \hat{\beta}_1 \cdot x_i - \hat{\beta}_0\right)^2 = \varepsilon_i^2$$

Asimismo, una de nuestra hipótesis es que la esperanza de los residuos es igual a cero, con lo cual:

$$\begin{aligned} \sigma_\varepsilon^2 &= Var\left(\varepsilon\right) \\ &= E\left\{ \left[\varepsilon - \underbrace{E\left(\varepsilon\right)}_{=0} \right]^2 \right\} \\ &= E\left(\varepsilon^2\right) \end{aligned}$$

Al dividir el cuadrado de los residuos por $n-2$ en lugar de n se logra que la estimación sea insesgada y, además, se refuerza la idea de que no podremos hacer análisis de regresión a menos que tengamos más de dos pares de datos como muestra.

Ahora, recordemos que el objetivo que nos planteamos en esta sección fue el de generar intervalos de confianza para los verdaderos valores de los coeficientes. El problema con el cual nos enfrentamos es la imposibilidad de conocer la varianza de estos coeficientes al desconocer cuál es la varianza del error. En consecuencia, a pesar de estar trabajando con variables Normales, al utilizar una estimación de la varianza en reemplazo de su verdadero valor, procederemos a estimar los intervalos con la distribución **t de Student** con $n-2$ grados de libertad.

El estadístico que utilizaremos tiene la misma forma para cualquiera de los dos coeficientes:

$$t = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\left(\hat{\beta}_i\right)} \quad \square \quad t_{n-2}$$

Podemos, entonces, decir que si se realizan sucesivas estimaciones a partir de muestras de tamaño n, en el $(1-\alpha)\%$ de los casos el intervalo $\left[\hat{\beta}_i - t_{n-2;\,1-\alpha/2}\cdot\hat{\sigma}\left(\hat{\beta}_i\right); \hat{\beta}_i + t_{n-2;\,1-\alpha/2}\cdot\hat{\sigma}\left(\hat{\beta}_i\right)\right]$ contendrá el verdadero valor del parámetro. Entonces, podemos, definir la estimación por intervalos de los coeficientes de regresión:

Sea $\hat{\beta}_i$ estimador puntual mínimo cuadrático de uno de los coeficientes de la regresión lineal, el intervalo

$$\hat{\beta}_i - t_{n-2;\,1-\alpha/2}\cdot\hat{\sigma}\left(\hat{\beta}_i\right) \le \beta_i \le \hat{\beta}_i + t_{n-2;\,1-\alpha/2}\cdot\hat{\sigma}\left(\hat{\beta}_i\right)$$

contendrá al verdadero valor de β_i en el $(1-\alpha)\%$ de las muestras que se consideren procedentes para la estimación.

Ejemplo 12

Continuemos el análisis de regresión del ejemplo anterior. Para la relación entre gasto en publicidad e incremento en las ventas, obtuvimos que:

$$\hat{\beta}_1 = 1,74612 \qquad \hat{\beta}_0 = 12,3945$$

	x	y	ŷ	(ŷ - y) ^ 2
	5	20	21,1251	1,2659
	12	30	33,3480	11,2092
	20	55	47,3170	59,0279
	35	70	73,5090	12,3128
	50	100	99,7009	0,0895
Suma	122	275	275,0000	83,9053

Estimemos ahora la varianza del error. Para ello, mostramos en la tercera columna de la tabla el valor estimado de y, y, correspondiente a cada valor de x, de acuerdo con los coeficientes estimados. En la cuarta columna, entonces, mostramos el cuadrado de la diferencia entre el verdadero valor observado y su estimación.

Teniendo en cuenta que contamos con 5 observaciones, y utilizando los datos de la tabla, podemos calcular la estimación de la varianza del error:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2} = \frac{83,9053}{3} = 27,9684 \qquad \Rightarrow \ \hat{\sigma}_\varepsilon = 5,2885$$

Si utilizamos el cuadro del Ejemplo 11 para calcular los desvíos de los coeficientes, se tiene que:

$$Var\left(\hat{\beta}_0\right) = \sigma_\varepsilon^2 \cdot \left(\frac{1}{n} + \frac{\left(\overline{x}\right)^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \cdot \frac{1}{n}} \right) = 27,9684 \cdot \left(\frac{1}{5} + \frac{24,4^2}{4294 - \frac{122^2}{5}} \right) = 18,2351$$

$$\boxed{\sigma_{\hat{\beta}_0} = 4,27025}$$

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \cdot \frac{1}{n}} = \frac{27,9684}{4294 - \frac{122^2}{5}} = 0,02123$$

$$\boxed{\sigma_{\beta_1} = 0,1457}$$

El intervalo de confianza con una significatividad del 95% (es decir, $\alpha = 0,05$) para β_0 es:

$$12,394 - t_{3;\ 0,975} \times 4,27025 \leq \beta_0 \leq 12,394 + t_{3;\ 0,975} \times 4,27025$$

Buscando en la tabla o calculando con Microsoft® Excel el valor de la variable *t*, obtenemos que $t_{3;\ 0,975} = 3,182$, y por lo tanto el intervalo es:

$$-1,19349 \leq \beta_0 \leq 25,9824$$

Por lo tanto, la estimación para el término independiente no resulta ser muy precisa, fundamentalmente debido a que el tamaño de la muestra es pequeño.

Para el otro coeficiente, el intervalo es:

$$1,74612 - t_{3;0,975} \times 0,1457 \leq \beta_1 \leq 1,74612 + t_{3;0,975} \times 0,1457$$

$$1,283 \leq \beta_1 \leq 2,210$$

Si bien no lo presentamos en este apartado, conocer la distribución de estos estimadores es importante, no sólo para la construcción de los intervalos de confianza, sino también para analizar la significatividad de los coeficientes. Utilizando las técnicas de los tests de hipótesis presentados en el Capítulo 7, podemos plantearnos si cada uno de los dos coeficientes estimados resultan ser o no significativos para el análisis. A los fines del marco de una estimación lineal, resultará importante testar la significatividad de β_1, dado que es a través de este valor que se manifiesta la tasa de variación constante que implica la linealidad. Por lo tanto, este testeo tiene una relación muy fuerte con el testeo de ρ y hace a la presencia o no de una relación lineal. En cambio, que el valor del coeficiente β_0 pueda ser igual a cero, afecta la forma de la recta, pero no cambia el carácter de la linealidad. Es decir, si tenemos $E\left(Y / X = x\right) = \beta_1 \cdot x$ ó $E\left(Y / X = x\right) = \beta_1 \cdot x + \beta_0$, la relación entre X e Y es lineal de todas formas.

En caso de que deseáramos realizar este testeo, debemos plantear las siguientes hipótesis:

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

siendo el estadístico de prueba el estadístico t con $n-2$ g.l.

El rechazo de la hipótesis nula dejará evidencia, para el nivel de confianza con el cual se esté trabajando, de que el coeficiente β_i es significativo y, por lo tanto, no debería excluirse de la recta de regresión planteada.

Por el contrario, si los datos empíricos no permiten el rechazo, existe la posibilidad de que dicho coeficiente sea innecesario, al existir una alta probabilidad de que tome el valor cero.

6.2.3 Intervalos de confianza para Y

Es importante también observar qué es lo que sucede con el valor que estimamos de la variable independiente Y. El mismo es una variable aleatoria y, como tal, también tiene asociada una dispersión y podemos generar intervalos de confianza respecto a su verdadero valor medio.

Lo que necesitamos conocer es la dispersión de esta variable. Ya hemos visto en las primeras secciones que parte de la aleatoriedad de la estimación está dada por el término de perturbación estocástico ε. Otra parte de la misma está dada por la muestra utilizada y los consecuentes errores en la estimación.

El desvío estándar de la estimación es:

$$\hat{\sigma}_{\hat{y}_i} = \hat{\sigma}_\varepsilon \cdot \sqrt{\frac{1}{n} + \frac{\left(x_i - E(X)\right)^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{5} \cdot \left(\sum_{i=1}^{n} x_i\right)^2}}$$

Dado este desvío, el verdadero valor medio de Y ante un valor predeterminado de x estará incluido en el intervalo $\left[\hat{y} - t_{n-2;1-\alpha/2} \times \hat{\sigma}_y ; \hat{y} + t_{n-2;1-\alpha/2} \times \hat{\sigma}_y\right]$ el $100 \times (1-\alpha)\%$ de las veces. Es decir que si tomáramos infinitas muestras apareadas de las dos variables y estimáramos con cada una el intervalo anterior, el $100 \times (1-\alpha)\%$ de los intervalos contendría al verdadero valor esperado de Y.

El intervalo de confianza del valor esperado de Y dado el valor x de la variable X, con un $100 \times (1-\alpha)\%$ de confianza es:

$$\hat{y} - t_{n-2;1-\alpha/2} \times \hat{\sigma}_y \leq E(Y|X=x) \leq \hat{y} + t_{n-2;1-\alpha/2} \times \hat{\sigma}_y$$

Siendo $y = \beta_1 \times x + \beta_0$

Ejemplo 13

A los fines de mostrar el modo en el cual realizar un análisis de regresión de manera completa, continuaremos con el ejemplo de la relación entre el gasto en publicidad y el incremento de las ventas.

La estimación del incremento medio de ventas ante una inversión de $30.000 había resultado ser igual a 64.778 unidades. Construyamos ahora, un intervalo de confianza para el verdadero valor del valor medio con un 95% de confiabilidad.

El valor estimado, como ya dijimos, surgió de:

$$\hat{y} = 1,74612 \cdot 30 + 12,3945 = 64,7783$$

Procedamos al cálculo de $\hat{\sigma}_{\hat{y}}$. El único dato que no calculamos en ejercicios anteriores es $\left(x_i - \bar{X}\right)^2$.

El lector podrá fácilmente comprobar que $\bar{X} = 24,4$ y que $\left(x_i - \bar{X}\right)^2 = (30 - 24,4)^2 = 31,36$.

Si reemplazamos en la expresión por los valores de las tablas ya calculadas:

$$\hat{\sigma}_{\hat{y}_i} = \hat{\sigma}_\varepsilon \cdot \sqrt{\frac{1}{n} + \frac{\left(x_i - \bar{X}\right)^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{5} \cdot \left(\sum_{i=1}^{n} x_i\right)^2}} = 5,2885 \cdot \sqrt{\frac{1}{5} + \frac{31,36}{4294 - \frac{122^2}{5}}} = 2,5019$$

Con lo cual, el intervalo será

$$64,7783 - 3,182 \cdot 2,5019 \leq E(Y|X=30) \leq 64,7783 - 3,182 \cdot 2,5019$$

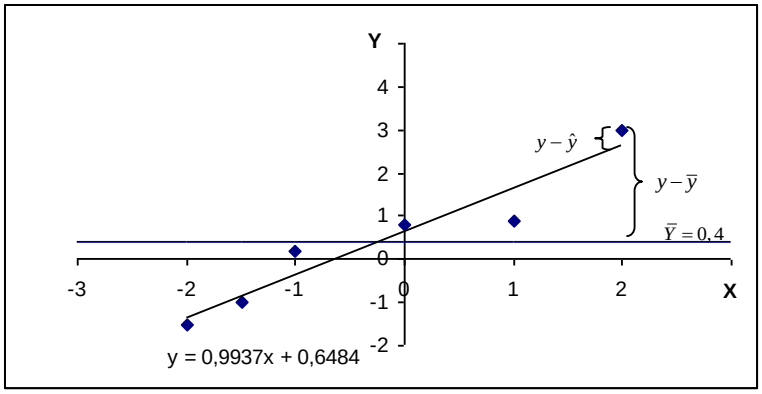
$$56,8219 \leq E(Y|X=30) \leq 72,7441$$

6.3 Bondad del Ajuste y Coeficiente de determinación

Cuando nos referimos a la existencia de una relación lineal entre variables, definimos dos medidas asociadas al grado de esta relación: la covarianza y el coeficiente de correlación. En este apartado definiremos al **coeficiente de determinación** como medida representativa de la **bondad del ajuste**, es decir, una medida que nos resume con qué grado de precisión la línea de regresión se ajusta a los datos.

Para poder entender un poco más respecto de esta medida, necesitamos comenzar con un **análisis de los residuos**, es decir, con un análisis de la diferencia entre el valor pronosticado de Y y el verdadero valor que toma la variable. Para medir qué tan significativa es la suma de estos errores, debe poder compararse con alguna otra.

Analicemos el gráfico siguiente:



Para cada uno de los datos que se tienen puede analizarse la diferencia de cada valor respecto de la media general (lo cual está relacionado con la varianza de la variable y obedece, claramente, a su característica de aleatoria) y, por otro lado, la diferencia entre la estimación y el verdadero valor.

Tomemos el caso de las variables que se exponen en la siguiente tabla y a las cuales corresponde el gráfico la regresión anterior.

X	Y
-2	-1,5
-1,5	-1
-1	0,2
0	0,8
1	0,9
2	3

Si desconociéramos la existencia de una relación entre ambas y quisiéramos pronosticar un valor de y, tomaríamos como medida el promedio de las observaciones, es decir, el valor $\bar{y}=0,4$. Sin embargo, si consideramos la relación lineal que existe entre las mismas, esperaríamos un valor distinto de y según el valor que se considere de la variable x. El lector podrá comprobar, a partir de la metodología expuesta, que la recta de regresión estimada por mínimos cuadrados es:

$$\hat{y}=0,9937\cdot x+0,6484$$

En consecuencia, para el valor $x=1,5$ tendremos la estimación $\hat{y}=2,1389$, siendo este pronóstico muy distinto al anterior calculado mediante el promedio simple de las observaciones de Y. La suma de los errores en uno y otro caso serán distintos y, si la regresión lineal es buena, los errores al aplicar la estimación deberían ser un porcentaje muy pequeño de los errores respecto a la media.

Un análisis de ese tipo es el que se realiza al calcular el coeficiente de determinación, al cual se lo denomina r^2. Para ello, primero realicemos algunas definiciones.

La Suma de Cuadrados Total es:

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Esta medida considera los desvíos de cada observación respecto del promedio de la variable Y (la variable a estimar), sin considerar la relación que ésta tiene con la variable X. Puede verse como la suma de los desvíos cuadráticos, ante el caso en el cual no trabajemos con la regresión para explicar los valores de la variable dependiente.

La Suma de Cuadrados Residual o Variación Residual es:

$$SCR = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

$$= \sum_{i=1}^{n} \varepsilon_i^2$$

En este caso, estamos considerando la suma de los cuadrados de los errores de la estimación realizada mediante la regresión, la misma sumatoria que utilizamos para estimar la varianza del error.

La Suma Explicada de Cuadrados o Suma Explicada por la Regresión es:

$$SCE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Esta última representa la variación de los valores estimados de Y, mediante la regresión respecto del promedio de las observaciones. Es decir, en cuánto diferenciamos nuestra estimación al utilizar la *regresión lineal,* en lugar del valor medio general de la variable.

Puede demostrarse que la Suma de Cuadrados Total es igual a la suma de las otras dos medidas de desvíos que han sido planteadas[62].

La Suma de Cuadrados Total es igual a la adición de la Suma de Cuadrados Explicada y la Suma de Cuadrados Residual:

$$SCT = SCE + SCR$$

Una vez expuesta las definiciones anteriores, podemos definir el coeficiente de determinación de la regresión.

El **coeficiente de determinación,** r^2, es el porcentaje de la variación total de Y que es explicada por la regresión lineal:

$$r^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

La interpretación surge de considerar:

[62] Ver Gujarati (2004) para una demostración

$$SCT \quad - \quad SRC \quad = \quad SCE$$

Variación total de Y — Residuos de la Regresión — Variación explicada por la estimación

Cuanto mayores son los residuos de la regresión, peor será el ajuste. En consecuencia, la Suma de Cuadrados Explicada será mayor, cuanto mejor sea el ajuste.

Podemos decir, además, que $\left(1-r^2\right)$ muestra la proporción de la variación total de Y que no es explicada por la regresión (aquella proporción del error de estimación que no pudimos disminuir mediante la regresión).

Dado que el coeficiente de determinación es una proporción, puede tomar valores únicamente dentro del rango $[0;1]$, indicando el valor $r^2=1$ la existencia de correlación perfecta, mientras que $r^2=0$ significa ausencia de correlación *lineal*.

Una forma sencilla de obtener este coeficiente es a través del coeficiente de correlación.

El coeficiente de determinación es el cuadrado del coeficiente de correlación de Pearson:

$$r^2 = (\hat{\rho}_{XY})^2$$

Los dos coeficientes utilizados en el análisis de regresión tienen su utilidad y difieren en cuanto a su interpretación. La diferencia entre ambos está en que el coeficiente de determinación tiene asociado un significado (representa el porcentaje de la variación de Y que es explicado por la regresión) mientras que el coeficiente de correlación sólo nos permite clasificar la relación lineal en "fuerte" o "débil".

Por otra parte, el coeficiente de correlación tiene la ventaja, por sobre el de determinación, de indicarnos el sentido de la relación, es decir, que nos permite saber si existe relación directa o inversa entre las variables.

Ejemplo 14

Si retomamos el caso planteado al introducir el tema en esta sección, podemos calcular qué porcentaje de la variación es explicada por la regresión. Los cálculos necesarios para hacerlo se resumen en la tabla siguiente:

X	Y	$\bar{Y}$	$\hat{Y}$	$(y-\bar{Y})^2$	$(\hat{Y}-\bar{Y})^2$
-2	-1,5	0,4	-1,3389	3,6100	3,0239
-1,5	-1	0,4	-0,8421	1,9600	1,5428
-1	0,2	0,4	-0,3453	0,0400	0,5554
0	0,8	0,4	0,6484	0,1600	0,0617
1	0,9	0,4	1,6421	0,2500	1,5428
2	3	0,4	2,6358	6,7600	4,9988
			Suma	12,7800	11,7255
				↓ SCR	↓ SCE

Entonces, $r^2 = \dfrac{SCE}{SCT} = \dfrac{11,7255}{12,78} = 0,9175$

El análisis de regresión propuesto explica el 91,75% de la variación de Y.

Ejemplo 15

Si consideramos nuevamente el análisis de la relación entre incremento de ventas y gastos de publicidad, podemos calcular el coeficiente de determinación simplemente elevando al cuadrado el coeficiente de correlación.

Si recordamos lo visto en las secciones anteriores, podemos calcular:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{460}{16,2308 \text{x} 28,6356} = 0,9897$$

Entonces, $r^2 = 0,9897^2 = 0,9795$

Este análisis de regresión explica el 97,95% de la variación.

6.4 Anexo: Demostraciones

6.4.1 Expresión alternativa de la Covarianza

$$\begin{aligned}
Cov(X,Y) &= \sum_x \sum_y (x-u_X)(y-u_Y) P_{XY}(x,y) \\
&= \sum_x \sum_y (x \cdot y - y \cdot \mu_X - x \cdot \mu_Y + \mu_X \mu_Y) P_{XY}(x,y) \\
&= \sum_x \sum_y x \cdot y \cdot P_{XY}(x,y) - \mu_X \underbrace{\sum_x \sum_y y \cdot P_{XY}(x,y)}_{\mu_Y} - \mu_Y \underbrace{\sum_x \sum_y x \cdot P_{XY}(x,y)}_{\mu_X} + \\
&\quad + \mu_X \mu_Y \underbrace{\sum_x \sum_y P_{XY}(x,y)}_{=1 \text{ por def. de probabilidad}} \\
&= \sum_x \sum_y x \cdot y \cdot P_{XY}(x,y) - 2 \cdot \mu_X \mu_Y + \mu_X \mu_Y \\
&= \left[\sum_x \sum_y x \cdot y \cdot P_{XY}(x,y) \right] - \mu_X \cdot \mu_Y
\end{aligned}$$

6.4.2 Covarianza de Variables Aleatorias Independientes

Por definición, si X e Y son variables independientes, entonces, $P_{XY}(x,y) = P_X(x) \cdot P_Y(y)$

Si, considerando esa definición, procedemos al cálculo de la covarianza, se tiene que:

$$\begin{aligned}
Cov(X,Y) &= \left[\sum_x \sum_y x \cdot y \cdot P_{XY}(x,y) \right] - \mu_X \cdot \mu_Y \\
&= \left[\sum_x \sum_y x \cdot y \cdot P_X(x) \cdot P_Y(y) \right] - \mu_X \cdot \mu_Y \\
&= \underbrace{\left(\sum_x x \cdot P_X(x) \right)}_{\mu_X} \underbrace{\left(\sum_y y \cdot P_Y(y) \right)}_{\mu_Y} - \mu_X \cdot \mu_Y \\
&= \mu_X \cdot \mu_Y - \mu_X \cdot \mu_Y \\
&= 0
\end{aligned}$$

6.4.3 Deducción de los Estimadores Mínimo Cuadráticos de los Coeficientes de la Recta de Regresión

Al momento de estimar los parámetros β_1 y β_0, minimizamos la suma de los cuadrados de las desviaciones de las observaciones respecto a la recta. Para ello, se deben dar, simultáneamente, las dos condiciones siguientes:

$$\begin{cases} \dfrac{\partial \sum_{i=1}^{n}\left(y_i-\hat{\beta}_1\cdot x_i-\hat{\beta}_0\right)^2}{\partial \hat{\beta}_1}=0 \\ \dfrac{\partial \sum_{i=1}^{n}\left(y_i-\hat{\beta}_1\cdot x_i-\hat{\beta}_0\right)^2}{\partial \hat{\beta}_0}=0 \end{cases}$$

Es decir, la derivada primera de la expresión respecto a cada uno de los parámetros debe ser igual a cero. Si desarrollamos las dos expresiones se tiene que:

$$\frac{\partial \sum_{i=1}^{n}\left(y_i-\hat{\beta}_1\cdot x_i-\hat{\beta}_0\right)^2}{\partial \hat{\beta}_1}=\sum_{i=1}^{n}\underbrace{2\cdot\left(y_i-\hat{\beta}_1\cdot x_i-\hat{\beta}_0\right)\cdot\left(-x_i\right)}_{\text{Importante: Recordar que aquí es } \beta_1 \text{ la variable y que los valores de x están dados por la muestra}} \quad \text{(regla de la cadena)}$$

Como la condición que estamos imponiendo es que esta derivada sea igual a cero, la primera restricción de nuestro sistema de ecuaciones será:

$$\sum_{i=1}^{n}2\cdot\left(y_i-\hat{\beta}_1\cdot x_i-\hat{\beta}_0\right)\cdot\left(-x_i\right)=0$$

$$-2\left[\sum_{i=1}^{n}y_i\cdot x_i-\hat{\beta}_1\sum_{i=1}^{n}x_i^2-\hat{\beta}_0\sum_{i=1}^{n}x_i\right]=0$$

$$\hat{\beta}_1\sum_{i=1}^{n}x_i^2+\hat{\beta}_0\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i\cdot x_i \qquad \text{(Condición 1)}$$

Realizando el mismo procedimiento, pero derivando respecto de β_0, se obtiene la segunda condición:

$$\frac{\partial \sum_{i=1}^{n}\left(y_i-\hat{\beta}_1\cdot x_i-\hat{\beta}_0\right)^2}{\partial \hat{\beta}_0}=\sum_{i=1}^{n}2\cdot\left(y_i-\hat{\beta}_1\cdot x_i-\hat{\beta}_0\right)\cdot\left(-1\right)$$

Al igualar a cero esta derivada:

$$\sum_{i=1}^{n}2\cdot\left(y_i-\hat{\beta}_1\cdot x_i-\hat{\beta}_0\right)\cdot\left(-1\right)=0$$

$$-2\cdot\left[\sum_{i=1}^{n}y_i-\hat{\beta}_1\cdot\sum_{i=1}^{n}x_i-n\cdot\hat{\beta}_0\right]=0$$

$$n\cdot\hat{\beta}_0+\hat{\beta}_1\cdot\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i \qquad \text{(Condición 2)}$$

El sistema queda así conformado con dos ecuaciones y dos incógnitas, pudiéndose expresar en forma matricial y resolviéndose por cualquiera de los métodos conocidos por el lector (utilizaremos Cramer en esta demostración)

$$\begin{bmatrix} \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & n \end{bmatrix}\cdot\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_0\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n}x_iy_i\\ \sum_{i=1}^{n}y_i\end{bmatrix}$$

Resolviendo el sistema:

$$\hat{\beta}_1 = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \qquad \hat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i \cdot y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} = E(Y) - \hat{\beta}_1 \cdot E(X)$$

7 Números Índice

Dario Bacchini
Lara Vazquez
Valeria Gogni

Seguramente, habrá escuchado hablar de la evolución de la inflación, de la evolución del costo de vida o de la modificación periódica del nivel de ventas de una empresa o sector. Uno de los principales intereses, cuando se analiza una variable determinada, está relacionado entonces con su evolución. En general, interesa conocer de qué manera se comporta la misma en distintos períodos del tiempo y observar estos valores comparativamente de manera clara. Una herramienta fundamental para este análisis es la construcción de índices, los cuales se tratarán en este capítulo.

Si bien el concepto de índice es sencillo y no le presentará mayor dificultad, sí es complejo el proceso de construcción de los mismos, de modo tal de asegurar que reflejen fielmente la variable que quiere medirse. Se mencionará también en este capítulo cuáles son los recaudos que deben tomarse y cuáles son los factores que pueden incidir en la construcción y generar resultados disímiles para la medición de un mismo fenómeno. Finalmente, acompañan a este capítulo, al igual que a los anteriores, una propuesta de ejercicios referidos al tema con el objetivo de afianzar los conceptos presentados.

7.1 Concepto

La utilización de los números índice se manifiesta en el análisis de series temporales. A partir de conocer el comportamiento de una variable a lo largo del tiempo, en muchas ocasiones pueden realizarse predicciones o estimaciones acerca de ellas. En otros casos, no nos interesa conocer el comportamiento para inferir, sino para explicar ciertos fenómenos o tomar algún tipo de decisión.

Como ya se ha dicho en capítulos anteriores, el análisis de series temporales es complejo y objeto de estudio de la Econometría. A pesar de la mencionada complejidad, un primer acercamiento al estudiar la evolución de las variables, se manifiesta a través de los números índice. Comprender estos indicadores es sencillo y, aún más sencillo, obtener a partir de los mismos una noción respecto al comportamiento de la variable.

En la tabla 1, se observa una evolución período a período, considerando la evolución de las toneladas exportadas por una determinada empresa agropecuaria.

Año	Toneladas
1990	310963,474
1991	354875,680
1992	360080,089
1993	325275,600
1994	342840,482
1995	418629,697

La variable que nos interesa ver, el comportamiento es, en este caso, las toneladas exportadas anualmente por nuestro país. Encontrar si esa cantidad aumentó o no con el correr de los años (o si tuvo un comportamiento errático) no presentará mayor dificultad que la de observar los seis dígitos (obviando los decimales) que caracterizan a la variable en cuestión. De ese modo, sin realizar cálculo alguno, podremos ver que el tonelaje exportado crece entre 1990 y 1992, cae en 1993 y luego hay una nueva tendencia de crecimiento hasta el año 2006.

Probablemente, si quisiéramos profundizar el análisis, buscaríamos encontrar las diferencias de año a año y la incidencia de estas diferencias medida en forma porcentual: recordemos que, en los análisis de variables, referirnos a montos absolutos puede prestarse a confusión (que la exportación crezca 20 kilos de año a año no generará diferencia, mientras que si una persona aumenta 20 kilos de año a año la situación de esta persona será muy diferente en uno y otro período. En ambas situaciones, el valor del incremento es de 20 kilos, pero el impacto en el análisis es muy distinto.

Por otra parte, incluimos también a modo de ejemplo, en la tabla 2, el Índice de Precios al Consumidor (IPC) publicado por el INDEC[63] para los meses del año 2006.

[63] www.indec.mecon.gov.ar

Año 2006 - IPC	
Enero	172,12
Febrero	172,80
Marzo	174,88
Abril	176,58
Mayo	177,41
Junio	178,27
Julio	179,37
Agosto	180,38
Septiembre	182,00
Octubre	183,56
Noviembre	184,86
Diciembre	186,67

Este índice muestra la evolución de los precios y suele utilizarse como indicador de la inflación. El carácter creciente para los meses considerados muestra la presencia de este último fenómeno. A diferencia del caso del tonelaje exportado, los números de la tabla no tienen un significado en sí mismo, dado que no se refieren a precios promedios, ni a suma de precios, ni a precio de ningún bien simple que podamos encontrar en el mercado. Sin embargo, como ya se ha mencionado, la observación de todos los datos en conjunto nos muestra un incremento en el valor del índice en forma continuada para todos los meses.

Antes de definir el concepto de índice, debemos tener en cuenta que los mismos constituyen un **indicador**. En consecuencia, decir que un índice toma un valor determinado tiene significado sólo en relación con su valor para otro momento de tiempo, pero no en sí mismo. En nuestro ejemplo, que el Índice de Precios al Consumidor tome el valor 180,38 en agosto de 2006 no significa nada a menos que lo comparemos con su valor en otro período. Así, si consideramos el valor que toma en diciembre del mismo año (186,67) el mayor valor del índice refleja un incremento general de los precios.

Un **número índice** es una relación en porcentaje que mide el cambio, para distintos períodos del tiempo, en precios, cantidades o alguna otra variable de interés.

Tal como lo expresa la definición, un número índice va a reflejar, porcentualmente, el cambio en una variable respecto de un momento particular, al cual se denomina **período base**. Así, una vez elegido el período respecto del cual se van a analizar las diferencias, será sencillo encontrar qué relación mantienen entre sí los valores a lo largo del tiempo.

Para el período base el índice toma el valor 100, y el valor que tome para los otros momentos representará la relación que tengan con el valor del período base elegido. De esta manera, si para el año X el índice toma el valor 130, esto significa que la variable es un 30% mayor respecto del año tomado como base. Lo que se logra al trabajar con índices es que, mediante la resta de sus valores en forma directa con la del período base se obtienen las diferencias presentadas en el tiempo en términos porcentuales (relativos) y no absolutos, permitiendo ver de manera más clara las tendencias en la evolución de las variables, sobre todo en los casos en los cuales las cifras que se manejan son importantes.

Ejemplo1

Supongamos que el índice que se muestra en la tabla 3, se utiliza en una compañía para analizar el impacto de su campaña publicitaria en términos de la evolución del consumo de un bien determinado.

Período	Índice	Período	Índice
Ene-05	98,5	Ago-05	125
Feb-05	99	Sep-05	124
Mar-05	100	Oct-05	124
Abr-05	105	Nov-05	120
May-05	110	Dic-05	122
Jun-05	122	Ene-06	116
Jul-05	128	Feb-06	115

Dado que la campaña comenzó su vigencia en abril de 2005, deciden considerar al mes de marzo de ese año como base. Se dice entonces que el índice es base $Marzo\,2005 = 100$. Lo que nos resulta interesante es interpretar el significado de los valores que se muestran. A pesar de no conocer las cantidades vendidas a partir de la campaña, notamos que hubo un amplio crecimiento para los meses siguientes al inicio de la misma. El valor de $I_{\text{Abr }05} = 105$ muestra que las ventas se incrementaron en un 5% respecto a las del mes anterior, ya con el primer mes de campaña. Este porcentaje se calcula sencillamente sustrayendo del valor del índice del período, el del período base

$$(I_{\text{Abr }05} - I_{\text{Mar }05} = 105 - 100 = 5).$$

Para Julio, con el mismo razonamiento, podemos ver que las cantidades vendidas fueron un 28% mayores a las presentadas en marzo. Para el mes de agosto, las unidades vendidas continúan siendo superiores a las de marzo, pero inferiores a las del mes anterior, mostrando que la campaña va "perdiendo efecto". Si observamos la variación entre febrero y marzo de 2005 puede verse que febrero tuvo ventas un 1% menores a las de marzo

$$(I_{\text{Feb }05} - I_{\text{Mar }05} = 99 - 100 = -1).$$

Si queremos analizar el porcentaje en el que creció de un mes a otro, el cálculo debe modificarse dado que el porcentaje se calcula sobre el período anterior:

$$\frac{I_{Mar05} - I_{Feb05}}{I_{Fer05}} \times 100 = \frac{100}{99} = 1,01$$

El número calculado implica que las ventas de febrero a marzo crecieron en un 1,01%. Si comparamos el crecimiento mensual en las ventas "pre campaña" con el crecimiento del 5% presentado en el primer mes, podemos ver que la misma ha generado un impacto.

Otro análisis interesante que podemos hacer es considerar la variación porcentual entre dos períodos que no son base. Así, si queremos analizar en forma comparativa las unidades vendidas para enero de 2005 con las correspondientes a enero de 2006 podemos hacer:

$$\frac{I_{\text{Ene }06} - I_{\text{Ene }05}}{I_{\text{Ene }05}} \times 100 = \frac{116 - 98,5}{116} \times 100 = 15,09$$

Es decir, las ventas para enero de 2005 crecieron en un 15,09% respecto a las del mismo mes para el año anterior.

En base a lo observado en el ejemplo, podemos analizar el cambio porcentual en una variable de interés a través de la evolución de un índice de acuerdo con la siguiente fórmula:

$$\text{Variacion}\%\left(t_0;t_1\right) = \frac{I_1 - I_0}{I_0}$$

Como mostraremos más adelante, en este mismo capítulo, la utilización de índices nos va a permitir también analizar la evolución conjunta de un grupo de variables, como es el caso del precio de una canasta de bienes.

La observación de la serie de índices para diferentes períodos mostrará su evolución y, por lo tanto, es importante reflejarla de manera tal de no generar en quien la observa una percepción errónea. En consecuencia, al momento de la construcción de un índice es importante la determinación del período base, sobre el cual se va a reflejar la relación entre las variables. Si la realización de la variable para ese momento fue menor a la que usualmente se presenta, esperaríamos valores muy altos del índice para períodos siguientes. Lo contrario ocurriría en el

caso en que la realización sea mayor a la usual. Los índices nos harían suponer gran variabilidad respecto de ese momento y tenderíamos a considerarlo inestable a pesar de que, para el resto de las observaciones, la diferencia no sea tan notoria. En el caso del comportamiento de los precios de nuestro país, no deberíamos considerar como representativos los precios de 2002, dado que la crisis impactó fuertemente sobre los mismos y se presentaron períodos muy inestables. Asimismo, tampoco podríamos tomar como base años hiperinflacionarios, como 1989.

Ejemplo2

Observemos la irregularidad, en tabla 4, de los precios para 1989 observando el IPC con **Base 1989**.

Período	IPC
Jun-89	0,403757
Jul-89	1,197680
Ago-89	1,651138
Sep-89	1,805608
Oct-89	1,906637
Nov-89	2,030931
Dic-89	2,844785
Ene-90	5,0980

Podemos ver que, entre junio y Julio de 1989, el nivel de precios se incrementó en un 196,6%, es decir, casi se **triplican** los precios.

$$\frac{1,197 - 0,4037}{0,4037} \times 100 = 196,66$$

Dado el caso de ese mes como las variaciones de los meses siguientes, 1989 no sería un año que cumpliera con los requisitos deseables para constituir un período base.

A los fines de realizar un análisis objetivo, el período base debe ser elegido de manera tal que no existan argumentos para ser considerado poco representativo del comportamiento: en el caso de que analicemos las toneladas cosechadas de un determinado cultivo, no podremos elegir como base a un año en el cual las condiciones climáticas generaron un importante impacto negativo porque los índices tendrían valores muy elevados; en realidad, no sería adecuado reflejar un crecimiento tan grande cuando lo que se produjo fue un retorno a las capacidades normales de producción.

Un efecto similar aparece cuando un mismo período base se utiliza para una serie de tiempo muy larga. Muchos factores han incidido en ese caso en lo que hace a la evolución de la variable (tecnologías, estructura de gastos, etc.) y los valores pueden encontrarse distorsionados o tener valores muy altos o demasiado pequeños, perdiéndose la sencillez en el análisis.

En resumen, podemos decir que:

Al momento de seleccionar el **período base**, éste debe ser uno para el cual la variable no haya tenido valores excepcionales por medidas políticas transitorias, fenómenos climáticos poco frecuentes o crisis especiales. Asimismo, los períodos no deben ser muy prolongados a fin de que continúe siendo útil su uso.

El proceso de construcción de un índice difiere según el tipo del cual se trate. Básicamente, distinguimos entre índices simples e índices ponderados. En general, los mismos se refieren a precios o cantidades, pero la metodología puede también aplicarse a otra variable (como puede ser la construcción de un índice que refleje el comportamiento de otros). En las secciones siguientes analizamos los mismos.

7.2 Índices Simples y Ponderados

El caso más sencillo de números índice es el **índice simple** (o no ponderado). Este índice se corresponde con la primera intuición que podríamos hacernos de ellos en base a la definición dada en la sección anterior, pero tiene la desventaja de que puede considerarse representativo en un número limitado de casos.

Al tomar un índice simple, el valor de la variable analizada en un determinado momento será igual al porcentaje que la misma representa respecto del período base.

Ejemplo 3

Si consideramos la evolución de la cantidad vendida de lápices comunes en una librería en los meses de un año cualquiera, la misma podría reflejarse mediante la tabla 5 que presentamos en este ejemplo.

Mes	Cantidad
Enero	20
Febrero	300
Marzo	600
Abril	450
Mayo	200
Junio	125
Julio	101
Agosto	130
Septiembre	235
Octubre	100
Noviembre	150
Diciembre	10

Si quisiéramos construir un índice simple que muestre la evolución de las cantidades vendidas, lo primero que deberíamos hacer es seleccionar nuestro mes base. La elección dependerá del analista y, dado que no tenemos datos respecto a la situación del contexto en cada uno de los meses, podremos tomar cualquiera de ellos (asumimos que el valor máximo presentado en marzo, así como los valores bajos de diciembre y de enero se debe al ciclo lectivo y corresponde a variaciones regulares que se presentan en todos los años, no a hechos excepcionales). Tomemos como período base al mes de septiembre. La notación usual consiste en que el valor de la variable para el período base se simboliza con el subíndice cero. En consecuencia, la cantidad vendida (que es nuestra variable de interés) en el mes base se representa como $q_0 = 235$, y el valor del índice para este período es, por definición, igual a $Q_{Sep} = 100$.

Para el resto de las variables se tiene que el valor del índice en el mes "i" es el porcentaje de las ventas de ese mes respecto a las de septiembre, es decir:

$$Q_i = \frac{q_i}{q_0} \times 100$$

donde q_i es la cantidad vendida en el mes "i", q_0 es la cantidad del período base, y el cociente entre ambas, q_i / q_0, indica la proporción que representa la cantidad vendida del mes "i" respecto de la cantidad base.

Puede entonces construirse la tabla 6 que refleja, para cada mes, el valor del índice.

Mes	Índice
Enero	8,51
Febrero	127,66
Marzo	255,32
Abril	191,49
Mayo	85,11
Junio	53,19
Julio	42,98
Agosto	55,32
Septiembre	100
Octubre	42,55
Noviembre	63,83
Diciembre	4,26

A modo de ejemplo, mostramos el cálculo del mismo para el mes de octubre:

$$Q_{\text{Oct}} = \frac{q_{\text{Oct}}}{q_{\text{Sep}}} \times 100 = \frac{100}{235} \times 100 = 42,55$$

Este valor indica que las ventas de lápices para octubre son aproximadamente el 43% de las cantidades vendidas en septiembre. Si esperamos para el año siguiente un comportamiento similar al del año para el que se construyó la tabla, probablemente en octubre pidamos a nuestro proveedor la cantidad suficiente para tener en el local el 43% de los lápices vendidos en septiembre.

De la misma manera en la que trabajamos con cantidades en el ejemplo anterior, podríamos considerar el caso de un análisis de precios. En este caso, la metodología para la construcción del índice será la misma, pero estaremos trabajando con la variable p (precios) en lugar de con la variable q (cantidad).

Sea P_t la variable aleatoria precio de un bien en el período t y p_t el valor que la misma toma en ese período (la *realización* de la variable). Para cada uno de los momentos considerados, el valor del **índice de precios simple** es:

$$I_t^{(P)} = \frac{p_t}{p_0} \times 100$$

donde p_0 es el precio del bien en el período base.

De manera análoga, sea Q_t la variable aleatoria representativa de la cantidad de un concepto dado para el período t y q_t la realización de la misma. Puede calcularse para cada uno de los momentos el **índice de cantidades simple** mediante la fórmula siguiente:

$$I_t^{(Q)} = \frac{q_t}{q_0} \times 100$$

donde q_0 es la cantidad representativa del período base.

Ejemplo 4

Consideremos la evolución del precio de una entrada al cine, en tabla 7, para los años desde 2000 a 2006. Podemos también considerar el precio promedio de las entradas de teatro en Capital Federal para el mismo período y el precio de una cena para una persona.

Año	Pcio Cine	Pcio Teatro	Pcio Cena
2000	7	35	20
2001	7	35	22
2002	8,75	40	28
2003	9,5	39	28
2004	10,5	41	30
2005	11,5	42	32
2006	12,75	45	35

Si quisiéramos calcular un índice de precios simple para cada uno de los casos, lo único que deberíamos hacer es considerar, para cada variable analizada, el porcentaje que representa el cociente entre el precio en el período base (año 2000) y el precio en el período que se analice. Los índices se muestran en la tabla 8, el lector puede intentar hacer en este punto el desarrollo por sí solo y corroborar luego los resultados.

Año	Ip (cine)	Ip (teatro)	Ip (cena)
2000	100,00	100,00	100,00
2001	100,00	100,00	110,00
2002	125,00	114,29	140,00
2003	135,71	111,43	140,00
2004	150,00	117,14	150,00
2005	164,29	120,00	160,00
2006	182,14	128,57	175,00

Ahora, consideremos el caso en el cual la evolución que queremos conocer es la del precio de una salida. Es decir, queremos saber en cuánto aumentó el costo de una salida recreativa entre 2000 y 2006, tomando al año 2000 como período base. La alternativa que menores dificultades presenta es comparar, período a período, la suma de los conceptos definidos como representativos de la salida (teatro, cine y cena): esta metodología consiste, como veremos a continuación, en un **índice de precios agregado simple**. El índice, para cada uno de los años, se calculará entonces como:

$$I_t = \frac{p_t^{cine} + p_t^{teatro} + p_t^{cena}}{p_0^{cine} + p_0^{teatro} + p_0^{cena}} \times 100$$

De esta manera, el *"Índice de Precios de esparcimiento"* que estamos proponiendo para el año 2006 tomaría el valor:

$$I_{2006} = \frac{12,75 + 45 + 35}{7 + 35 + 20} \times 100 = 149,6$$

Como podemos comparar, el incremento de precios que manifestó la entrada al teatro es inferior al incremento general del costo de la salida: el índice simple para el teatro en 2006 indica un incremento del 28,57% respecto de 2000 (ver tabla 9) mientras que el índice agregado muestra que el incremento del costo general de la salida fue del 49,6% para ese mismo período.

Año	Suma Pcios	Ind. Agregado
2000	62	100,00
2001	64	103,23
2002	76,75	123,79
2003	76,5	123,39
2004	81,5	131,45
2005	85,5	137,90
2006	92,75	149,60

En el ejemplo anterior, hemos utilizado un índice de precios agregado simple. Se trabajó con un bien compuesto de tres bienes distintos y se analizó la evolución conjunta de los mismos. El análisis conjunto de los precios de distintos bienes resulta muy útil para medir la evolución de los costos o indicadores de la inflación, dado que no podemos evaluar el impacto de la inflación analizando el precio de un único bien. En el caso planteado, nos limitamos a analizar una canasta de tres bienes como representativos, pero debe prestarse especial cuidado al momento de

considerar los bienes contenidos en la misma. Por ejemplo, para el índice creado estamos omitiendo el costo del transporte o del combustible que puede generar también un importante cambio en los costos.

Un **índice de precios agregado simple** para una canasta de n bienes y para el período t se expresa como el cociente entre la suma de los precios de los n elementos considerados en el momento t respecto de la suma de los precios de los mismos bienes pero para el período base.

$$I_t^{Ag.Sim} = \frac{\sum_{k=1}^{n} p_{k,t}}{\sum_{k=1}^{n} p_{k,0}} \times 100$$

donde $p_{k,t}$ representa el precio del bien k en el momento t.

Si bien este tipo de índice puede considerarse una primera aproximación al cálculo de la evolución conjunta de los precios representativos de alguna canasta en particular, dista mucho de reflejar de forma consistente el impacto del incremento de precios.

Para el ejemplo que vimos, la cantidad de veces que una persona va a cenar en una salida, probablemente, sea distinta a la cantidad de veces que se dirige al cine o al teatro. En consecuencia, el impacto en el cambio de precios no será el mismo si el incremento es en uno u otro caso. Piénselo de este modo: si se considera que por cada cuatro veces que una persona sale, tres de ellas va a cenar, dos concurre al cine y sólo una va al teatro (está implícito en este caso planteado que habrá *"salidas importantes"* en las que hará más de una actividad) un aumento en el precio de la entrada del teatro generará un incremento general del costo de sus salidas, pero será pequeño en relación a sus gastos totales. Sin embargo, si el aumento se da en el precio de las cenas, el bolsillo de esta persona se verá mucho más afectado. Un aumento de $1 en el valor de la entrada del teatro va a influenciar en una de cada cuatro veces que salga, mientras que, si se presenta el mismo aumento en el cine, en cuatro salidas la persona gastará $2 más en lugar de $1 ¿Cómo medir estos diferentes impactos en las modificaciones de precios? A partir del **índice de precios agregado ponderado**.

Un **índice de precios agregado ponderado** para una canasta de n bienes y para el período t se expresa como el cociente entre la suma de los precios de los n elementos considerados en el momento t ponderados por una cantidad representativa del consumo de cada uno de ellos respecto de la suma de los precios de los mismos bienes ponderados por igual cantidad pero para el período base.

$$I_t^{Ag.Pond} = \frac{\sum_{k=1}^{n} q_k \cdot p_{k,t}}{\sum_{k=1}^{n} q_k \cdot p_{k,0}} \times 100$$

donde $p_{k,t}$ representa el precio del bien k en el momento t y q_k la cantidad consumida del bien *k*.

Se desprende claramente de la fórmula expuesta que, si las cantidades q_k son iguales para todos los productos que componen la canasta, el impacto del cambio del precio de un bien en $\$X$ impacta igual que un incremento del mismo monto en cualquiera de los otros bienes. En ese caso, entonces, no existe necesidad de realizar un índice ponderado dado que el concepto resultante es el mismo que para un índice simple: el impacto del cambio es independiente del bien en el cual se dé.

Ejemplo 5

Retomemos el ejemplo de nuestro "*Índice de esparcimiento"* (ver tabla 10 con la evolución de los precios), pero, esta vez, ponderando el impacto del cambio en los precios con las cantidades usualmente consumidas de cada uno de ellos.

Año	Pcio Cine	Pcio Teatro	Pcio Cena
2000	7	35	20
2001	7	35	22
2002	8,75	40	28
2003	9,5	39	28
2004	10,5	41	30
2005	11,5	42	32
2006	12,75	45	35

Tomemos como base lo expuesto en párrafos anteriores, por lo cual diremos que por cada cuatro salidas, tres veces va a cenar, dos concurre al cine y sólo una al teatro. Traducido a términos simbólicos podemos decir que:

$$q^{Cena} = \frac{3}{4}; \qquad q^{Cine} = \frac{2}{4}; \qquad q^{Teatro} = \frac{1}{4}$$

Como estas cantidades son utilizadas como ponderadores, su valor en sí mismo no importa, sino que lo significativo es la relación con las demás: la cantidad de veces que sale a cenar y al cine es el triple y el doble, respectivamente, de la cantidad que va al teatro. Esto sucede porque el valor final del índice no se alterará si se multiplica a todas las cantidades (ponderadores) por un mismo número, ya que ese número sería factor común en las sumatorias y se cancelarían en el cociente. Podemos entonces tomar:

$$q^{Cena} = 3\ ; \qquad q^{Cine} = 2\ ; \qquad q^{Teatro} = 1$$

El cálculo del índice se muestra en la tabla 11.

Año	Pcine	Pteatro	Pcena	Qci x Pci	Qt x Pt	Qce x Pce	Sum. Prod	Ind. Ag. Pond
2000	7	35	20	14	35	60	109	100,00
2001	7	35	22	14	35	66	115	105,50
2002	8,75	40	28	17,5	40	84	141,5	129,82
2003	9,5	39	28	19	39	84	142	130,28
2004	10,5	41	30	21	41	90	152	139,45
2005	11,5	42	32	23	42	96	161	147,71
2006	12,75	45	35	25,5	45	105	175,5	161,01

Por ejemplo, el valor para 2006, tomando como base el 2000, se obtiene de la siguiente manera:

$$I_{2006} = \frac{12,75 \times 2 + 45 \times 1 + 35 \times 3}{7 \times 2 + 35 \times 1 + 20 \times 3} \times 100 = 161,01$$

Este valor refleja que el costo de la salida es un 61% superior en el 2006 respecto al costo en el 2000. Si recordamos el ejemplo anterior, el valor resultante del índice de precios agregado simple reflejaba un incremento del casi 50%, es decir, un incremento menor. Si analizamos cada componente por separado veremos que la causa de esa diferencia es que el costo del teatro, que es el que menor cambio presentó (se modificó en un 29%) es también el que menor impacto tiene, cobrando importancia en la ponderación los otros dos conceptos que sufrieron mayores aumentos.

Conocido el concepto de índice se hace más fácil medir la evolución conjunta de una canasta de bienes, a través de un análisis ponderado de los precios, siguiendo una metodología similar. Así, dependiendo de cuál sea la variable a analizar, se utilizarán criterios similares a los presentados para comparar, por ejemplo, la evolución de una cartera de contratos derivados o una cartera de acciones. Los índices, en estos casos, son también indicadores que no significan nada por sí mismos, pero permiten analizar el comportamiento en el tiempo al comparar los valores para

distintos momentos. Dos casos de índices de este tipo que podemos mencionar son el Índice Merval Argentina y el Índice ICA MATba[64] .

El primero de estos índices refleja la rentabilidad de una inversión en acciones de las empresas argentinas, con lo cual será representativo de la evolución de este rendimiento. Para su construcción se parte de una cartera teórica (que determina la cantidad de acciones para cada empresa, no dándole una ponderación superior al 20% de la cartera a cada una) y se considera el nivel de precios agregado de esta cartera en base a los criterios de ponderación determinados. Dado el carácter cambiante del Mercado de Valores, el período base y las cantidades que componen el índice se modifica trimestralmente, conforme al criterio de que el mismo no puede encontrarse muy alejado del período de análisis. El Índice ICA MATba se basa en un concepto similar, pero las ponderaciones se modifican mensualmente. Este índice representa el valor de mercado de una cartera de futuros determinada ponderando el valor de cada futuro con la proporción de la cantidad de operaciones abiertas del mismo respecto del total de operaciones del mercado. Una dificultad que presenta este índice es que la vigencia de un contrato de futuro es mucho más corta que la de una acción y, en consecuencia, la base se modifica mensualmente y debe tenerse en cuenta al momento de crear la cartera representativa que ninguno de los contratos venzan a lo largo de ese mes. A partir de los conceptos introducidos en este capítulo, el lector no encontrará dificultad en comprender la interpretación y construcción de cualquiera de los dos índices, ambos detallados en las páginas web referenciadas.

7.3 Índices de Laspeyres y Paasche

Dos de los índices usualmente utilizados son los índices de Laspeyres y de Paasche. Los mismos no le presentarán dificultad, dado que se trata de índices agregados ponderados, en lo que se modifica entre uno y otro es el criterio de ponderación.

Hemos visto en la sección anterior que, a los fines de analizar la correcta evolución de precios, tenemos que ponderar la influencia de cada bien sobre el total de la canasta a analizar. De esta manera, si analizamos la evolución del precio de los alimentos, podemos decir que un cambio en el precio de la leche o del pan va a afectar en mayor medida que el cambio en el precio del caviar. Esto es así porque, en un análisis general de precios, las cantidades consumidas de los dos primeros artículos son muy superiores respecto a la cantidad consumida del tercero. Ahora, ¿cómo elegir de manera correcta las ponderaciones? Si bien esto queda fuera del alcance del objetivo de este manual, tenemos que tener presente que, para cada concepto analizado, se toma una **canasta representativa** en la cual se incluyen las cantidades de cada bien en función de lo consumido en cada período.

Si en todos los períodos se hubiera consumido igual cantidad de cada uno de los bienes, no tendríamos problema en elegir la canasta representativa dado que la estructura de los gastos se ha mantenido constante. En ese caso, la ponderación se realiza, como en el caso del Ejemplo 5, multiplicando los precios por las cantidades representativas.

Ejemplo 6

Consideremos, en la tabla 12, la evolución de los gastos para un estudiante de nivel universitario entre 2002 y 2006.

Año	Pcio Apuntes	Pcio Cuadernos	Pcio Lapiceras	Pcio Libros
2002	5	2	2	45
2003	5	2,8	2,1	50
2004	7	3,2	2,5	52
2005	6	3,5	2,65	55
2006	8	4,5	3	60
Cant. Consum.	20	12	16	6

[64] Para mayor información respecto al modo de construcción de estos índices puede consultar las páginas www.merval.sba.com.ar y www.matba.com.ar.

Si consideramos que las cantidades detalladas en la última fila son las representativas de su consumo anual y que el año base es el 2003, podemos calcular un índice ponderado de precios para cada uno de los años haciendo:

$$I_t = \frac{p_t^{Apunte} \times 20 + p_t^{Cuad} \times 12 + p_t^{Lapic} \times 16 + p_t^{Libro} \times 6}{\underbrace{5 \times 20 + 2{,}8 \times 12 + 2{,}1 \times 16 + 50 \times 6}_{= 467{,}20}} \times 100$$

De esta manera, obtenemos los siguientes valores:

Año	Cantidad x Precio					Índice
	Apunt	Cuad	Lapic	Libros	Suma	
2002	100	24	32	270	426	91,18
2003	100	33,6	33,6	300	467,2	100,00
2004	140	38,4	40	312	530,4	113,53
2005	120	42	42,4	330	534,4	114,38
2006	160	54	48	360	622	133,13

Así, por ejemplo, vemos que el gasto de un estudiante universitario entre 2003 y 2005 aumentó en un 14,38%.

Ahora, ¿qué pasa si las cantidades consumidas en cada período no son las mismas? En estos casos, tendríamos que decidir cuál es la canasta que nos interesa. Si para cada período ponderáramos, respecto a la cantidad consumida en el mismo, los cambios que estaríamos representando, no serían los cambios en los precios, únicamente, sino también en las cantidades.

El Índice de precios de Laspeyres considera el cambio de precio que se manifiesta respecto del año base y para una canasta construida en ese mismo período base. Es decir, Laspeyres va a analizar el cambio en los precios teniendo en cuenta cuánto más (o menos) me cuesta en el período t consumir **la misma canasta que consumía en el período base**: esto es, la ponderación está dada por las cantidades representativas del período base.

El **Índice de Precios de Laspeyres** es un índice agregado ponderado en donde la ponderación está dada por las cantidades representativas de una canasta construida en el período base.

$$IL_t = \frac{\sum_{i=1}^{n} p_{i,t} \times q_{i,0}}{\sum_{i=1}^{n} p_{i,0} \times q_{i,0}} \times 100$$

donde $p_{i,t}$ es el precio del bien i en el período t y $q_{i,0}$ la cantidad consumida del bien i en el período base.

Por otro lado, el criterio que sigue Paasche para considerar la evolución de los precios es ver en cuánto se hubiera visto modificado el precio de la canasta que se consume en el momento t, entre el período base y dicho momento. La ponderación, entonces, está hecha por las cantidades **representativas de la canasta del momento para el que se construye el índice**.

El **Índice de Precios de Paasche** es un índice agregado ponderado en donde la ponderación está dada por las cantidades representativas de una canasta construida en el período de análisis.

$$IP_t = \frac{\sum_{i=1}^{n} p_{i,t} \times q_{i,t}}{\sum_{i=1}^{n} p_{i,0} \times q_{i,t}} \times 100$$

donde $p_{i,t}$ es el precio del bien i en el período t y $q_{i,t}$ la cantidad consumida del bien i en el período t.

La idea en ambos índices es la de ver la evolución del precio de **una misma canasta**, difieren únicamente en la canasta utilizada.

Ejemplo 7

Consideremos nuevamente los precios reflejados en el Ejemplo 6, pero asumamos que las cantidades consumidas en cada año fueron distintas. En la la tabla 14 se presenta la cantidad representativa de cada uno de ellos.

	Cantidades			
Año	Apunt	Cuad	Lapic	Libros
2002	20	12	16	6
2003	22	12	15	5
2004	24	11	16	4
2005	25	10	14	4
2006	27	10	12	2

Tomemos nuevamente como período base el año 2003 y calculemos el índice para 2004 y 2006, considerando tanto el índice de Laspeyres como el de Paasche.

Comencemos por el índice de Laspeyres. En este caso, tomamos como cantidades representativas a las del año 2003:

$$IL_{2004} = \frac{7 \times 22 + 3,2 \times 12 + 2,5 \times 15 + 52 \times 5}{5 \times 22 + 2,8 \times 12 + 2,1 \times 15 + 50 \times 5} \times 100$$
$$= \frac{489,9}{425,10} \times 100 = 115,24$$

Para el año 2006 utilizar el índice de Laspeyres nos permite una ventaja: como las ponderaciones son respecto de las cantidades del año base la canasta representativa no se modifica y tampoco lo hace el denominador. Entonces:

$$IL_{2006} = \frac{8 \times 22 + 4,50 \times 12 + 3 \times 15 + 60 \times 5}{425,10} \times 100$$
$$= \frac{575}{425,10} \times 100 = 135,26$$

Consideremos, ahora, el valor del índice siguiendo el criterio de ponderación de Paasche:

$$IP_{2004} = \frac{7 \times 24 + 3,2 \times 11 + 2,5 \times 16 + 52 \times 4}{5 \times 24 + 2,8 \times 11 + 2,1 \times 16 + 50 \times 4} \times 100$$
$$= \frac{451,20}{384,40} \times 100 = 117,38$$

Para el año 2006 se modifican las cantidades tanto en el denominador como en el numerador, con lo cual volvemos a realizar el cálculo:

$$IP_{2006} = \frac{8 \times 27 + 4,5 \times 10 + 3 \times 12 + 60 \times 2}{5 \times 27 + 2,8 \times 10 + 2,1 \times 12 + 50 \times 2} \times 100$$
$$= \frac{417}{288,20} \times 100 = 144,70$$

Comparemos los resultados obtenidos. Para ambos años considerados el Índice de Laspeyres muestra un incremento de precios inferior al señalado en el índice de Paasche. Si analizamos, por ejemplo, el caso del año 2006, podemos ver que, si bien el precio que más aumentó respecto del 2006 es el de los cuadernos ($\frac{4,5-2,8}{2,8} = 60,71\%$), la cantidad consumida de los mismos bajó un 17%: el impacto de este aumento será menor en la canasta de Laspeyres (donde los cuadernos son un 22% de todos los bienes consumidos) que en la canasta de Paasche (donde representa un 20% del total de artículos). Al momento del análisis entonces deben tenerse en cuenta estas situaciones que hacen que los valores obtenidos sean diferentes de acuerdo al criterio que se utilice.

Probablemente, preferiremos utilizar para el caso de este ejemplo el cálculo del Índice de Paasche dado que refleja las relaciones para la estructura de consumo en cada uno de los años. La desventaja que presenta respecto al de Laspeyres es que el denominador de la ponderación se modifica período a período y, en consecuencia, se requiere un mayor número de cálculos. El cálculo del Índice de Laspeyres es más sencillo, pero corremos con la desventaja de estar mostrando la evolución de una canasta que ya no es representativa. Supongamos, por ejemplo, el caso de una serie de precios más larga. Si consideramos en los últimos diez años una canasta compuesta por los bienes utilizados para la recreación de adolescentes, veremos que los mismos difieren bastante de un período a otro: probablemente en 1995 los cassettes seguían utilizándose con una ponderación media, la cual no es comparable con la de hoy (prácticamente nula) en donde los CDs y los MP3 tienen una importancia mucho mayor. Este cambio en la estructura de cantidades hace que el Índice de Laspeyres pierda importancia a medida que la distancia con el período base es mayor.

Si nosotros trabajamos con ejemplos pequeños como lo venimos haciendo, no encontraríamos elevado el costo de utilizar el índice de Paasche y obtener las carteras representativas de cada uno de los diferentes momentos. Ahora, en el caso de indicadores económicos como lo es el Índice de Precios Mayoristas o el Índice de Precios al Consumidor, la situación es muy distinta.

El Índice de Precios al Consumidor con base 1999, publicado por el INDEC, contiene las ponderaciones que se muestran en la tabla 15.

Agrupamiento principal	1999
Alimentos y bebidas	31,3
Indumentaria y calzado	5,2
Vivienda	12,7
Transporte y comunicaciones	17,0
Otros gastos	33,9
Gastos para la salud	10,0
Esparcimiento y educación	12,9
Esparcimiento	8,7
Educación	4,2
Bienes y servicios diversos	10,9
Equipamiento y funcionamiento del hogar	6,5
Bienes y servicios varios	4,4

Las mismas se calcularon teniendo en cuenta una encuesta realizada entre febrero de 1996 y enero de 1997 respecto a los gastos de cada rubro en hogares residentes en Capital Federal y 24 partidos de Gran Buenos Aires. Si consideramos que se requiere de al menos un año para construir una canasta representativa, ¿resulta válido el esfuerzo y costo de calcularla para cada uno de los años? Dada la relación costo - beneficio que se manifiesta, el Índice de Precios al

Consumidor que usualmente escuchamos como representativo de la inflación es un Índice de Precios de Laspeyres, cuya canasta representativa es la de la estructura de consumo de 1999 (estructura que se asimila al tratamiento de los datos de 1996/1997). Lo que sí debe realizarse para cada año es el cálculo del precio medio de los bienes considerados en el índice. En la página del INDEC se encuentra detallado todo el procedimiento de construcción. También el Índice de Precios Mayoristas es un Índice Laspeyres, fundamentándose su uso por las mismas causas.

Vista la aplicación de cada uno de los índices, a nivel teórico existe otro índice que se presenta como alternativa al considerar un punto intermedio entre Paasche y Laspeyres: el Índice de Fisher. En la práctica pocas veces se calcula y su contenido, generalmente, se corresponde a aplicaciones teóricas.

El **Índice de Fisher** se corresponde con la media geométrica entre el Índice de Paasche y el de Laspeyres:

$$IF_t = \sqrt{IL_t \times IP_t}$$

Debe prestarse importante atención al hecho de que tanto el Índice de Paasche como el de Laspeyres, debe haber sido confeccionado con el mismo período base.

Ejemplo 8

Si continuamos con el ejemplo del costo del estudiante, el valor del índice de Fisher para 2006, será la media geométrica entre los calculados en el Ejemplo 7:

$IF_{2006} = \sqrt{135,26 \times 144,70} = 139,9$. Tal como se ha mencionado, este valor es un valor intermedio entre los otros obtenidos.

Así como el índice de Laspeyres y de Paasche toman como ponderadores las canastas de bienes de determinados momentos del tiempo, de modo tal de comparar el precio de una canasta determinada, existen índices más complejos que se guían por otro criterio. Un caso de ello es el Índice Económicamente Significativo (Samuelson, 1983)[65], en donde el criterio para la comparación de precios es el de comparar dos canastas que proporcionen la misma "utilidad" en ambos momentos del tiempo.

7.4 Cambios en la Base

A los fines de comparar dos series de índices cuyo período base es distinto debemos adoptar un criterio homogéneo para hacer más sencilla la comparación. Tomemos el *"Índice de precios de esparcimiento"* creado en el Ejemplo 5 y el de la evolución de los gastos de un estudiante para el Ejemplo 6. Asumamos que queremos analizar en qué rubro existieron mayores cambios. Los índices que habíamos calculado son expuestos en las tablas siguientes.

Año	Ind Esp.
2000	100,00
2001	105,50
2002	129,82
2003	130,28
2004	139,45
2005	147,71
2006	161,01

Año	Ind. Est.
2002	91,18
2003	100,00
2004	113,53
2005	114,38
2006	133,13

Los mismos tienen base 2000 y 2003 respectivamente. A los fines de compararlos sería útil considerar la misma base. Tomemos, por ejemplo, base 2003 para ambos. En el Índice de costos para el estudiante universitario no se presentarían cambios, pero sí deberíamos modificar los correspondientes a los precios de esparcimiento con base 2000. Los cálculos en este caso se asemejan a la construcción de un "*índice de índices"* donde el valor del índice de 2003 será considerado como base y el valor para los demás años será el porcentaje que representen del índice 2003 original. Es decir:

[65] Foundations of Economic Analysis - Enlarged Edition. Harvard University Press

$$I_{Base\,03,t} = \frac{I_{Base\,00,t}}{I_{Base\,00,03}} \times 100$$

Si existe una serie de índices con base en el año X y se la quiere re-expresar en base al año Y para los distintos momentos del tiempo t, el cálculo a realizar es muy sencillo y obedece a:

$$I_{Base\,Y,t} = \frac{I_{Base\,X,t}}{I_{Base\,X,Y}} \times 100$$

Siguiendo la definición, para el año 2004 el valor del índice de esparcimiento será igual a:

$$I_{Base\,03,2004} = \frac{I_{Base\,00,2004}}{I_{Base\,00,2003}} \times 100 = \frac{139,45}{130,28} = 107,04$$

Ahora, los porcentajes de variación están expresados en función del año 2003. La tabla 16 resume el valor de esos índices, ahora comparables con los correspondientes al costo estudiantil.

Año	Ind Esp.
2000	76,76
2001	80,99
2002	99,65
2003	100,00
2004	107,04
2005	113,38
2006	123,59

De este modo, podemos observar que el precio de actividades recreativas creció un 23,59% y el costo de estudios se incrementó un 33,13%, y en base a ello afirmar que el impacto se sufrió más en este último concepto.

Bibliografía

Berenson, M. y Levine, D. (1991), *Estadística para Administración y Economía*. McGraw-Hill.

Bayes, T. (1763), "An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S."

Canavos, G. (1997), *Probabilidad y Estadística*, McGraw-Hill.

Enders, W. (1995), *Applied Econometric Time Series*, John Wiley & Sons.

Gujarati, D. N. (2004), *Econometría*, McGraw-Hill.

Hamilton, D. (1994), *Time Series Analysis*, Princeton University Press.

Hilderbrand, D. y Ott, R.L. (1991), *Estadística Aplicada a la administración y la economía*, Addison-Wesley Iberoamericana.

Kolmogorov, A. N. (1933). "Grundbegriffe der Wahrscheinlichkeitsrechnung" Springer, Berlín.

Levin, R. y Rubin, D. (2004), *Estadística para Administración y Economía*, Pearson Educación.

Landro, A. H. (2002), *Acerca de la Probabilidad – 2da. Edición*, Ediciones Cooperativas.

Mason, R. y Lind, D. (1998), *Estadística para administración y economía- Octava Edición*, Alfaomega, México

Novales Cinca, A. (1997), *Estadística y Econometría*, McGraw-Hill.

Novales Cinca, A. (2000), *Econometría*, McGraw-Hill.

Samuelson, P. (1983), *Foundations of Economic Analysis – Enlarged Edition*, Harvard University Press.