Universidad de la República
Regional Norte Sede Salto

CURSO DE ESTADISTICA BÁSICA

por

Luis Salvarrey
Profesor Adjunto de Estadística Social
Facultad de Ciencias Sociales
Profesor Adjunto de Bioestadística
Instituto Nacional de Enfermería
Encargado del Curso de Métodos Cuantitativos
Facultad de Agronomía

CURSO DE ESTADISTICA BÁSICA

por Luis Salvarrey
Regional Norte Sede Salto
Universidad de la República O. del Uruguay

UN NUEVO PROLOGO

Estamos poniendo en Internet una versión de nuestro curso básico de Estadística para estudiantes universitarios. Hace años que estamos utilizando este material como recurso didáctico y ha demostrado ser de utilidad. En 1977 se diseñó para estudiantes de Agronomía, en 1990 se adaptó a los cursos de Ciencias Sociales, a partir de 1997 se utilizaron además en Enfermería, Psicología y recientemente con alumnos de Profesorado de Matemática del CERP del Litoral. Habíamos tratado de tener versiones diferentes para cada uno de este grupo de alumnos, luego declaramos imposible la tarea y elaboramos esta versión que intenta tener en un solo material el interés de todos. Ponemos a disposición de los alumnos ejercicios específicos para cada grupo de interés por separado. Uds. dirán que tal resulta.

El interés original de este material se mantiene: poner a disposición del alumno notas que reemplacen a las suyas. No nos gusta que los alumnos estén anotando, porque muchas veces no piensan por anotar. El material no pretende reemplazar a los libros, aunque algunas veces las circunstancias han llevado a situaciones de ese tipo. Otro punto de nuestro objetivo es entonces que los alumnos que no pueden comprar un libro como Spiegel puedan disponer de un material, aunque no reemplace a Spiegel le permite estudiar, repasar y practicar. Finalmente se ha demostrado que los ejercicios que aquí planteamos son útiles, aunque no sean suficientes. Nos gusta que los estudiantes lean en libros, los estimulamos a ello, como parte de ese estímulo preguntamos cosas que exceden a lo que mostramos en estos apuntes. Como dicen los matemáticos: este material es necesario pero no suficiente.

Finalmente repetimos el lema que siempre hemos manejado: *Oigo y olvido, leo y recuerdo, hago y aprendo*. Esperamos que con esta recetita nuestros alumnos sigan aprendiendo como lo han hecho durante estos 24 años que venimos enseñando. ¿Seguimos en contacto?

Luis Salvarrey,
Regional Norte Universidad de la República,
Salto, Uruguay.

BIBLIOGRAFIA

Estas notas no pretenden ser un sustituto para libros de referencia y consulta profesional, sino un sustituto a los apuntes de los estudiantes, por tanto recomendamos consultar la literatura. Por otro lado el material no tiene propósito de originalidad, esto no lo inventé yo. Los principales libros tenidos en cuenta en la elaboración de las notas fueron:

Johnson, R. *Elementary statistics.* 2nd. ed. 1976. Duxbury Press. North Scituate, Massachusetts. 550p

Ott, L.; Mendenhall, W. y Larson, R. F. 1978. *Statistics: a tool for the social sciences*. 2nd. ed. Duxbury Press. North Scituate, Massachusetts. 531p

Koopmans, L. H. 1981. *An introduction to contemporary statistics.* Duxbury Press, Boston, Massachusetts. 599p

Tanur, J. M. (Ed) 1972. *Statistics: a guide to the unknown.* Holden-Day, San Francisco.430p

Bancroft, T. A. (Ed) *Statistical papers in honor of G. W. Snedecor.* Iowa State University Press, Iowa. 328p

Federer, W. T. 1973. *Statistics and Society.* M. Dekker. N. York. 399p

así como las guías de estudio que acompañan a los dos primeros libros mencionados.

Libros para ampliar conocimientos recomendables son los siguientes:

Snedecor, G. W & Cochran, W. G. 1956. *Statistical methods.* 5th ed. Iowa State University Press, Ames, Iowa. 534p Hay ediciones mas nuevas y traducción al español.

Steel, R. G. D. & Torrie, J. H. 1980. *Principles and procedures of statistics.* 2nd. ed. Mc Graw Hill, N. York. 633p También existe edición en español.

Pardell, H., Cobo, E. & Canela, J. 1986. *Manual de bioestadística.* Masson, Barcelona. 263p

Ferguson, G. A. 1976. *Statistical analysis in Psychology and Education.* 4th. ed. 529p Para psicología.

Para los estudiantes de Sociología otra bibliografía de interés incluye:

Blalock, H. M. 1972. *Estadística social*. 2nd. ed. Original publicado por McGraw-Hill, N. York, 1972. Traducción por FCE, México.

Padua, J. 1979. *Técnicas de investigación aplicadas a las ciencias sociales.* Fondo de Cultura Económica, México.

Sierra Bravo, R. 1981. *Análisis estadístico y modelos matemáticos*. Paraninfo, Madrid.

García Ferrando, M. 1985. *Socioestadística. Introducción a la Estadística en Sociología*. Alianza Editorial.

García Ferrando, M. 1988. *Técnicas de investigación social. Teoría y ejercicios.* Ed. Paraninfo, Madrid.

Cochran, W. G. 1952. *Sampling techniques*. J. Wiley, N. York.

Técnicas de muestreo. 12ª . reimpresión. 1996. Compañía Editora Continental, México. 513p

Des Raj, 1980. *Teoría del muestreo*. Fondo de Cultura Económica. México. 305p

Otras referencias bibliográficas citadas en las notas son:

González, N. 1998. *Estudio de morbilidad por cáncer de mama en Hospital Regional Salto y Centro de Asistencia Médica de Salto durante el 1/1/97 al 1/12/97.* INDE, Regional Norte Sede Salto.

Kirkwood, B. 1988. *Essentials of medical statistics*. Blackwell Scientific Publications, Oxford, UK. OPS.

Mood, A. M. & Graybill, F. A. 1964. *Introducción a la teoría de la estadística.*. 2nd. ed. Mc Graw Hill, NY. Hay varias versiones.

TABLA DE CONTENIDO

CLASE 1
INTRODUCCION A LA ESTADISTICA DESCRIPTIVA
TABLAS y GRAFICAS

Muchas veces nos vemos enfrentados a una masa de datos que necesita ser resumida e interpretada. El propósito de la estadística descriptiva es proveernos de herramientas gráficas y numéricas para esa tarea.

1.1.VARIABLES. La estadística trabaja con *datos* de característica variabilidad conocidos por ello como *variables*. Las variables pueden ser clasificadas en variables cuantitativas y variables cualitativas. Las variables cuantitativas también se conocen como variables propiamente dichas, mientras que las cualitativas se conocen como atributos, clases o categorías. Una posterior división de las variables cuantitativas es en continuas y discontinuas o discretas.

Variables Cualitativas o Atributos o Clases

Variables Cuantitativas { Discretos; Continuos

El sexo de una persona es un atributo, mientras que la altura es una variable cuantitativa. Las variables (cuantitativas) se miden, los atributos se cuentan. Por ejemplo, diremos que una clase de estadística tiene 19 estudiantes mujeres y 2 varones. El sexo de una persona es un atributo pero el número de estudiantes de determinado sexo en una clase es una variable cuantitativa discreta. Por esta razón el análisis de atributos a veces se llama análisis de conteos.

1.2.ESCALAS. Una clasificación de las escalas de medida que ha tenido gran aceptación en los últimos tiempos es:

i. Escala nominal. La escala más rudimentaria es la nominal, donde los objetos se distinguen en base a un nombre, muchas veces dado por un número. Por ejemplo en el sexo de personas, se puede acordar un número para simbolizar a cada sexo, pero ese número es arbitrario y un investigador puede definir hombre como 0 y mujer como 1, mientras que otro puede utilizar exactamente lo opuesto. Las escalas nominales se usan en atributos.

ii. Escala ordinal. Las mediciones en una escala ordinal solo indican orden ("ranking"). Los objetos en una escala ordinal se distinguen, pues, en base a la cantidad relativa de una característica que poseen. Ejemplos de ésto son los grados usados en la medición del estado de información de una población con las categorías (pobre, regular, buena, excelente). Una escala es: 0, 1, 2, 3, 4, y 5, pero puede haber otras diferentes que distingan igualmente el grado de información de las personas.

iii. Escala por intervalos. Cuando las diferencias entre objetos tiene sentido, es decir que la unidad de medida es fija. Generalmente tienen un cero, aunque este es arbitrario, como en el caso de la temperatura medida en grados centígrados, donde el cero no indica ausencia de temperatura. No tiene sentido acá decir que una temperatura de 60 grados es doble que una de 30.

iv. Escala racional. Cuando, además de lo anterior, los cocientes (razones) de valores tienen sentido la escala es racional. Un ejemplo es el peso, donde un objeto que pese 60 kg. pesa el doble de uno que pesa 30 kg. El cero es absoluto en esta escala.

Hay una jerarquía en la escala presentada, al bajar la escala se pierde potencia del análisis, por lo que se sugiere que de hacerse voluntariamente se haga con cuidado. Por otro lado, no siempre es fácil adjudicar inequívocamente una escala.

Ejemplo 1.1. Para cada una de las siguientes variables identifique si son cualitativas, cuantitativas discretas o cuantitativas continuas y que tipo de escala (nominal, ordinal, por intervalos o racional) las representa mejor: sexo de una persona, coeficiente de inteligencia, estado civil, número de autos robados en un dia, altura de una persona, temperatura.

El sexo de una persona y su estado civil son variables cualitativas típicas, y como tales se representan en escalas nominales. El coeficiente de inteligencia es una variable cuantitativa continua, aunque generalmente se expresa en unidades enteras como 60, 61, 62, etc. Vemos acá una característica de que las variables continuas se presentan como discretas en la práctica, generalmente por problemas de medida. La altura de una persona se expresa en centímetros, aunque es continua por naturaleza no se justifica la molestia de ir mas allá de los centímetros para medirla. Similarmente sucede con la temperatura, aunque es continua se la mide en grados. En cuanto a las escalas, la variable temperatura es típica de escalas por intervalos, ya fue comentado. La altura se puede considerar racional, ya que tiene sentido decir que una persona es el doble de alto que otra. No sucede lo mismo con el cociente de inteligencia ya que no tiene sentido decir que una persona tiene 0 de cociente de inteligencia, ni que es el doble de inteligente que otra, por lo tanto se representa por una escala de intervalos. Finalmente el número de autos robados en una ciudad en un determinado periodo es una variable cuantitativa discreta, que se representa en una escala racional, ya que tiene sentido hacer comparaciones a través de cocientes.

Otro autor (Hinde, 1995, com. pers.) presenta el siguiente cuadro con diferentes tipos de datos categóricos

Tipo de variable de respuesta	Ejemplo
Categórica con dos categorías	Vivo/muerto Presencia o ausencia de una enfermedad Empleado/desempleado
Categórica con mas de dos categorías no ordenadas	Causa de muerte Tipo de cáncer Partido político Afiliación religiosa
Categórica con categorías ordenadas	Fuerza de convicción a una actitud política Clase en la universidad Severidad de síntomas de una enfermedad
Conteo discretos	Número de hijos en una familia Número de accidentes en un cruce de calles Número de choques de aviones en un año
Discretas duración de tiempo datos históricos	Situación de desempleo para cada mes (tiempo desempleado) Estado diario de salud (duración de una enfermedad)

Ejemplo 1.2. Los siguientes datos son parte de un estudio realizado por González (1998) sobre la incidencia de cáncer de mama en Salto. Los datos se utilizarán a efectos de ilustrar el uso de las ideas de estadística en una situación del área de Enfermería. Las variables fueron codificadas asi:

EDA edad de la encuestada, vemos que está en años
EST estadio en que se descubrió el cáncer, según un código médico
ESC escolaridad o años de estudio * luego se codificaron
NSE nivel socio económico de una escala predeterminada
EC estado civil de la encuestada
OC ocupación, tambien situaciones predeterminadas o codificables a posteriori
AC consumo de alcohol regular
AN consumo anterior de alcohol
TAB habito de fumar
AFIS actividad física?
INFO fuentes de información sobre cáncer de mama
AEM si realiza autoexamen de mama
SITAV situaciones adversas vividas en los ultimos tiempos
CSM
MOVO motivo de consulta al médico
MNA edad de la menarca

```
data tres;
input eda EST$ EsC$ NSE$ EC$ OC$ AC$ AN$ TAB$ AFIS$ INFO$ AEM$ Sitav$ CSM$ Movo$ MNA$;
 if    eda<51 then edai=1;
 if 50<eda<61 then edai=2;
 if 60<eda<71 then edai=3;
 if 70<eda<81 then edai=4;
 if 80<eda    then edai=5;   drop eda;cards;
65 IIIBr . .  C .  .   .   .   .      .      .   .      .     PE     .
75 IIA   A b2 C a  NC  NO  No  NN     SM     no  nin    CCE   PE     12
42 IIIA  B c3 D b  S   SI  515 SScaTo SM     Reg ds     CCE   PEMF   12
77 IV    A b2 C a  NC  NO  No  NN     N      Av  Ots    CCE   PED    13
76 IIBr  B c3 C b  NC  NO  No  SScaDv SMtv   Av  nin    CCE   PED    11
52 I     B b2 C a  O   NO  No  SScaDv SMtv   Av  mfd    CCEU  PE     11
59 IV    B b2 C a  O   NO  S5  Ssca   SMtv   Av  ef     CCEmg PEPr   11
61 I     A b2 U a  NC  NO  No  SScaTo Stv    Av  Ots    CCE   SE     13
82 IIBr  A .  V .  .   .   .   .      .      .   .      .     .      .
48 IV    A b2 C a  O   NO  No  SScaTr SMtv   Av  Ots    CCEU  PED    12
80 IV    B d4 V c  NC  NO  No  NN     Stv    no  ef     CCEmg PE     14
72 IV    A b2 V a  NC  NO  No  NN     N      Av  Ots    CCE   PE     12
71 I     B c3 C c  NC  NO  No  NN     Stvre  Av  nin    CCE   PEDRP  13
56 IIA   B c3 C b  NC  NO  No  SScaTr SM     Reg nin    CCE   PEAx   14
74 I     A a1 V a  NC  NO  No  NN     SM     no  nin    CCE   SEPE   14
65 IV    . .  . .  .   .   .   .      SM     .   .      .     .      .
76 IV    . .  . .  .   .   .   .      SM     .   .      .     .      .
56 0     B b2 C a  O   NO  S5  SScaDv N      Av  ds     CCE   PEPr   11
52 I     A a1 C a  O   NO  No  NN     Stv    Av  nin    CCE   PEDRP  12
65 I     B c3 C c  O   NO  No  NN     Stvrae Reg Ots    CCEmg PEMFD  11
72 IV    B d4 S a  O   NO  No  NN     Stv    Av  nin    CCE   PEMF   13
70 IV    B b2 S a  NC  NO  No  NN     SM     Av  nin    CCE   PE     .
48 I     A a1 D a  NC  NO  S5  NN     N      no  efplds CCE   PED    12
49 IV    B b2 C c  O   NO  No  SScaDv SMtv   Reg nin    CCE   PEPr   11
58 IV    B c3 C a  O   NO  No  NN     Stv    Av  efmfd  CCEmg MFPrRP 10
```

EDA edad de la encuestada, en años, variable cuantitativa discreta, escala racional
EST estadio en que se descubrió el cáncer, código médico, escala ordinal, cuanti discreta o nominal?
ESC escolaridad o años de estudio, cuanti discreta escala racional * luego se codificaron y la escala pasa a ser ordinal.
NSE nivel socio económico de una escala predeterminada, escala ordinal, variable cuanti o cuali?
EC estado civil de la encuestada, escala nominal típica, variable cualitativa
OC ocupación, tambien situaciones predeterminadas o codificables a posteriori, tambien cuali nominal
AC consumo de alcohol regular
AN consumo anterior de alcohol
TAB habito de fumar
AFIS actividad física?
INFO fuentes de información sobre cáncer de mama
AEM si realiza autoexamen de mama
SITAV situaciones adversas vividas en los ultimos tiempos
CSM MOVO motivo de consulta al médicoMNA edad de la menarca

1.3.TABULACION. Muchas veces, al comienzo de un trabajo de análisis de datos se cuenta con un gran volúmen de información en bruto. Una de las primeras tareas es organizar esa información y tabularla. El propósito de la tabulación es resumir la información hasta llegar, a veces, a un par de valores (la media y la varianza por ejemplo) que encierran toda la utilidad de la información.

Ejemplo de enfermeria. ***Variables cualitativas.*** Ya dijimos que cuando la variable es cualitativa se cuenta. Por tanto los valores se presentan en una tabla de frecuencias. Supongamos que estamos trabajando con datos de estado civil de las encuestadas con la primera letra como codigo ¿cómo ven mejor los datos Uds. Asi? C C D C C C C U V C V V C C V V V V S S D C C o asi?

Estado Civil	Frecuencia
Casadas	11
Divorciadas	2
Unión libre	1
Viudas	7
Solteras	2

1.3.2.*Variables cuantitativas discretas.* Cuando la variable es cuantitativa discreta también los valores se tabulan naturalmente. Miremos estos datos de edad de muchachas en una encuesta (también de González, 1998): 12 12 13 11 11 11 13 12 14 12 13 14 14 11 12 11 13 12 11 10 No quedan mejor asi?

Edad	Frecuencia
10	1
11	6
12	6
13	4
14	3

En cambio cuando se pretendió trabajar con la edad en que las mujeres contrajeron cáncer, encontramos: 65 75 42 77 76 52 59 61 82 48 80 72 71 56 74 65 76 56 52 65 72 70 48 49 58
y por tanto la tabla quedó:

Edad al Contraer el Cáncer	Frecuencia
42	1
48	2
49	1
52	2
56	2
58	1
59	1
61	1
65	3
70	1
71	1
72	2
74	1
75	1
76	2
77	1
80	1
82	1

lo cual es impráctico: la tabla quedó casi tan grande como los datos que intenta resumir. Por lo tanto la investigadora resolvió usar categorías de edad. Retabulando los datos en clases quedaron asi:

Categorias de edad	Frecuencias
Menor a 50	4
51 a 60	6
61 a 70	5
71 a 80	9
Mas de 80	1

Notemos que transformamos una variable cuantitativa en cualitativa.

No creemos que haya grandes reglas para hacer tablas, pero Pardell et al. (1986) proporciona algunas. Para determinar el número de clases, generalmente se toma la observación más alta y más baja (la diferencia es el rango), se divide el rango en 5 a 20 clases.

Ejemplo 1.3. Consideremos los siguientes 60 pesos de animales reportados por Madalena (1973):

234 225 234 225 234 204 225 231 245 202 213 222 231 245 193 202 213 222 229 243
254 193 202 213 220 229 243 254 193 200 211 218 227 243 254 265 184 191 197 211
216 227 240 250 263 274 145 177 188 197 209 216 227 236 247 256 272 288 304 210

Tabla 1.2. Tabulación de los datos del ejemplo 1.3.

Límites de la clase		Marca de la clase	Frecuencia Absoluta n_i	Frecuencia Relativa f_i	Frecuencia Acumulada Absoluta N_i	Frecuencia Acumulada Relativa F_i	Media de la clase
136	145	140	1	0,0167	1	0,0167	145
146	155	150	0	0	1	0,0167	
156	165	160	0	0	1	0,0167	
166	175	170	0	0	1	0,0167	
176	185	180	2	0,0333	3	0,0500	180
186	195	190	5	0,0833	8	0,1333	191
196	205	200	7	0,1167	15	0,2500	200
206	215	210	7	0,1167	22	0,3667	211
216	225	220	9	0,1500	31	0,5167	221
226	235	230	10	0,1667	41	0,6833	230
236	245	240	7	0,1167	48	0,8000	242
246	255	250	5	0,0833	53	0,8833	251
256	265	260	3	0,0500	56	0,9333	261
266	275	270	2	0,0333	58	0,9667	273
276	285	280	0	0	58	0,9667	
286	295	290	1	0,0167	59	0,9833	288
296	305	300	1	0,0167	60	1,0000	304

En la tabla 1.2 se presenta una forma habitual de tabular datos como esos en clases. Los valores entre 176 y 185 se consideran una clase, los entre 186 y 195 otra y asi sucesivamente. Una columna muestra los *límites* de cada clase, una segunda con la *marca de la clase* (es decir el valor que representa la clase, generalmente el punto medio o semisuma de los límites de clase), y una tercera con la *frecuencia absoluta* n_i. Esta última es el número de observaciones comprendidas en cada clase. Un concepto relacionado es el de *frecuencias relativas*, simbolizado por f_i que es el número de observaciones de cada clase dividido por el total de observaciones. La *amplitud* (o *longitud*) *de clase* es la diferencia entre los límites de una clase. Muchas veces las clases son de igual amplitud, pero no tiene porque ser así.

Para determinar el número de clases, generalmente se toma la observación más alta y más baja (la diferencia es el *rango*), se divide el rango en 5 a 20 clases y finalmente se determina el número de observaciones en cada clase, la frecuencia absoluta. Tambien se utiliza el concepto de *frecuencia acumulada* de una clase, que es el número de valores menores o iguales a los de esa clase.

Límites de la clase		Marca de la clase	Frecuencia Absoluta n_i	Frecuencia Relativa f_i	Frecuencia Acumulada Absoluta N_i	Frecuencia Acumulada Relativa F_i
136	145	140	1	0,0167	1	0,0167
146	155	150	0	0	1	0,0167
156	165	160	0	0	1	0,0167
166	175	170	0	0	1	0,0167
176	185	180	2	0,0333	3	0,0500
186	195	190	5	0,0833	8	0,1333
196	205	200	7	0,1167	15	0,2500
206	215	210	7	0,1167	22	0,3667
216	225	220	9	0,1500	31	0,5167
226	235	230	10	0,1667	41	0,6833
236	245	240	7	0,1167	48	0,8000
246	255	250	5	0,0833	53	0,8833
256	265	260	7	0,0500	56	0,9333
266	275	270		0,0333	58	0,9667
276	285	280		0	58	0,9667
286	295	290		0,0167	59	0,9833
296	305	300		0,0167	60	1,0000

1.4.GRAFICAS E HISTOGRAMAS.

Generalmente cierto tipo de gráfica o figura ayudará a la interpretación de los datos. Una regla que manejamos frecuentemente con los estudiantes es que la tabla es mas exacta y la gráfica mas gráfica, digo mas demostrativa de una idea. Por tanto, generalmente aconsejamos utilizar mas los recursos gráficos en presentaciones orales y utilizar mas las tablas en los informes escritos. Pero alguna gráfica generalmente enriquece un informe y lo hace mas leible. Tambien conviene seguir la idea del libro de Neris y no repetir la información de la tabla en una gráfica, al menos no demasiado.

Existen diferentes tipos de gráficos (o diagramas como les llaman impropiamente Pardell et al., 1986).

Graficos de puntos. Son aquellos donde la frecuencia se representa por un punto. Si los puntos están unidos se conoce como poligono o poligonal. Ojiva es la gráfica correspondiente con los valores acumulados, especialmente en caso que la gráfica este suavizada. Un caso de gráficos de puntos que no corresponden con frecuencias es cuando graficamos dos variables cuantitativas entre si (ver ejercicio 3 del práctico). Ese tipo de gráficos se les llama X-Y por los autores de software como Excel.

Graficos de lineas. Si marcamos la frecuencia con una línea vertical queda mas visible, conceptualmente no tienen diferencia con los gráficos de puntos.

Graficos de barras. Con una barra queda aún mas visible y estetico. Las barras tridimensionales que se usan ahora son de lo mas elegantes. Y en colores ni les cuento.

Histogramas. Cuando la variable es continua se usan los histogramas. El histograma es una representación gráfica en la que la frecuencia (puede ser absoluta o relativa) de la clase está representada por el área de la barra. Si todas las clases tienen igual amplitud, la frecuencia de la clase esta representada por la altura de la barra y el gráfico se confunde como gráfico de barras. Pero si las clases tienen diferente amplitud los gráficos de barra y los histogramas difieren. Mucha gente no conserva las diferencias y llama histograma a los gráficos de barras. El hecho de que el área sea lo que represente la frecuencia tiene importancia a efectos del trabajo con probabilidades. Recién al final del curso veremos esos temas.

En la figura 1.1 se representa la frecuencia (absoluta o relativa) de cada clase con la altura de la barra. Estas gráficas se llaman gráficos de barras.

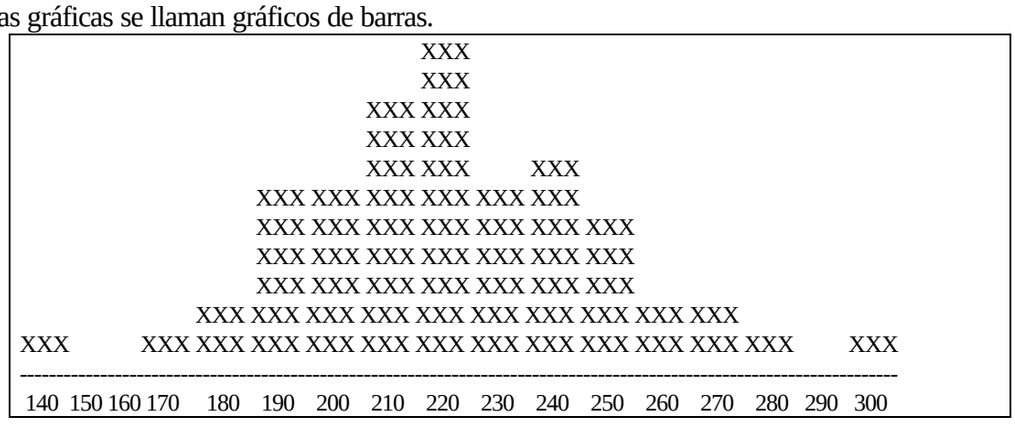


Figura 1.1. Gráfico de barras con los datos del ejemplo 1.3.

El histograma es una representación gráfica en la que la frecuencia (absoluta o relativa) de la clase está representada por el área de la barra. Si todas las clases tienen igual amplitud, la frecuencia de la clase esta representada por la altura de la barra y el gráfico se conoce como gráfico de barras. Mucha gente no conserva las diferencias y llama histograma a los gráficos de barras. Muchas veces se utilizan gráficos de puntos, donde la frecuencia se representa por un punto. Polígono de frecuencias se le llama a las gráficas donde los puntos están unidos. Ojiva es la gráfica correspondiente con los valores acumulados.

1.5.ELEMENTOS SEMIGRAFICOS. Existen varios elementos semi-gráficos

Steam-and-Leaf (Tallos y hojas). Si en lugar de representar cada valor por una marca cualquiera lo representamos por el dígito que lo identifica, no perdemos ese dato lo que puede ser de utilidad para el cálculo de ciertas cantidades como la media, tal cuál se verá mas adelante.

14	5
15	
16	
17	7
18	4 8
19	1 3 3 3 7 7
20	0 2 2 2 4 9
21	0 1 1 3 3 3 6 6 8
22	0 2 2 5 5 5 7 7 7 9 9
23	1 1 4 4 4 6
24	0 3 3 3 5 5 7
25	0 4 4 4 6
26	3 5
27	2 4
28	8
29	
30	4

Figura 1.3. "Steam-and-leaf" de los datos del ejemplo 1.3.

1.6. NIVELES DE MEDICIÓN Y PROCEDIMIENTOS ESTADÍSTICOS. En la siguiente tabla intentamos resumir como los diferentes procedimientos estadísticos se aplican a las escalas de medición:

Nivel de Medición	Estadisticas Descriptivas	Tablas de Frecuencia	Gráficos de Barras	EDA
Nominal		X	X	
Ordinal	X	X	X	
Intervalos	X	X	X	X
Racional	X	X	X	X

La X indica que el método es apropiado para ese nivel de medición. Se nota que las tablas de frecuencia y los gráficos de barra son aconsejados para todos los tipos de escalas, las estadísticas descriptivas que se explican en este capítulo se adaptan a escalas de rango ordinal o superior. El analisis exploratorio de datos (EDA - "exploratory data analysis") se adapta a escalas de intervalos o racional.

1.7. TIPOS DE GRAFICOS EN EXCEL. Si usamos la planilla electrónica Excel, que es un software muy popular en la actualidad, tenemos a nuestra disposición los siguientes tipos de gráficas:

Columnas o **barras** Si la barra es vertical le llama columna.

Circular también llamado gráfico de sectores o tortas para graficar porcentajes.

Líneas muy utilizados para graficos a traves del tiempo.
XY cuando las variables son de igual "orden de interés" en oposición a la situación en que graficamos una variable y la frecuencia con que ocurren cada uno de sus valores.

Las otras (**áreas**, **anillos**, etc.) son chiches no muy utilizados.

Práctico 1. Tablas y gráficas.

1.- Los siguientes datos corresponden a número de hijos por familia en un estudio. Grafique la frecuencia de hijos por familia.

Número de Hijos	Número de familias
0	8
1	16
2	38
3	22
4	10
5	6

2.- Presente los siguientes datos en una tabla con intervalos de 10 kgs. centrados en 55, 65, ...,185 kgs. Represente las frecuencias en un histograma.

103	133	111	184	127	124	117	102	124	115	153	122	105	104	115
140	115	113	117	125	135	127	125	121	84	87	108	85	101	117
90	144	106	111	97	70	113	113	110	64	55	90	93	107	93
89	94	100	126	119	82	98	57	100	134	111	113	93	117	122

3.- Grafique los siguientes valores en un sistema ortogonal de ejes:

X	0	1	2	3	4
Y	4	2	3,5	0,5	0

4.- Una fábrica tiene clientes en dos zonas del país. Con el fin de mejorar su política de ventas decide efectuar un estudio sobre las cantidades demandadas de su principal producto, llegando a los siguientes resultados:

Unidades Demandadas	Número de clientes Zona A	Zona B
0 - 100	20	30
100 - 200	30	35
200 - 300	35	50
300 - 500	25	40
500 - 800	10	15

Se pide extraer conclusiones primarias en base a la información proporcionada. Se sugiere para ello construir los histogramas de frecuencias relativas en un mismo gráfico.

5.- Los siguientes datos pertenecen a las notas obtenidas en exámenes de ingreso (X) y el promedio de notas de los estudiantes en su primer año de universidad (Y).

X	Y
37	97
66	30
97	57
27	77
55	63
84	87
14	96

Grafique los datos en un par de ejes ortogonales.

CONTESTE SI ES CIERTO O FALSO Y SI ES FALSO DIGA COMO CAMBIA LAS PALABRAS SUBRAYADAS PARA HACER VERDADERA LA FRASE.

1.- Estadística inductiva es el estudio y descripción de datos que resultan de una encuesta.

2.- Estadística descriptiva es el estudio de una muestra que nos permite hacer proyecciones o estimaciones acerca de la poblacion de la que se sacó la muestra.

3.- Una estadística es una medida calculada de alguna característica de una población.

4.- Un parámetro es una medida calculada de una muestra.

5.- En nuestra clase hay 20 personas, 17 mujeres, 3 hombres. El número de personas y el sexo son variables discretas.

6.- La altura de una persona es un atributo.

7.- El objetivo básico de la estadística es obtener una muestra, inspeccionarla, y hacer inferencia acerca de la población de la que se extrajo la muestra.

8.- De las siguientes variables indicar cuales son discretas y cuales continuas:

1. Número de libros en las estanterias de una biblioteca.
2. Temperaturas registradas cada hora.
3. Remuneraciones que se pagan en la industria.
4. Suma de puntos obtenidos en el lanzamiento de un par de dados.
5. Tiempo que cada día se dedica al estudio.
6. Número de pulsaciones por minuto.
7. Superficies de un conjunto de establecimientos.
8. Cantidad de lluvia caída.

CLASE 2
MEDIDAS DE POSICION

2.1. MEDIA ARITMETICA. La media es la suma de los valores dividido el número de valores. Si la media pertenece a una población se representa con la letra griega μ, si pertenece a una muestra con el símbolo de la variable con una barra encima[1]:

Muestral	Poblacional
$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$X = \frac{\sum_{i=1}^{n} X_i}{n}$

Ejemplo 1.2 (Cont.) La media de peso de los 60 datos del ejemplo 1.2 es 225,27 como puede comprobarse sumando los 60 valores y dividiendo por 60.

Ejemplo 1.3. Consideremos los siguientes datos encontrados en una investigación ya mencionada: 12 12 13 11 11 11 13 12 14 12 13 14 14 11 12 11 13 12 11 10. La media de estos valores es:
(12+12+13+11+11+11+13+12+14+12+13+14+14+11+12+ 11+13+12+11+10)/20=12,1

2.2. MEDIA DE DATOS AGRUPADOS

Ejemplo 1.3 (Cont.). Nótese que algunos de los valores se repiten. De hecho ya los habíamos tabulado así:

Edad X_i	Frecuencia n_i	$X_i\, n_i$
10	1	10
11	6	66
12	6	72
13	4	52
14	3	42
	20	242

En estos casos de datos repetidos no los sumamos sino que los multiplicamos:

$$\frac{10*1+ 11*6 + 12*6 + 13*4 +14*3}{20} = \frac{242}{20} = 12,1$$

La única particularidad del cálculo de la media si los datos están agrupados es, pues, que los valores deben multiplicarse por la frecuencia en que cada dato ocurre:

$$\bar{X} = \frac{\sum X_i\, n_i}{\sum n_i}$$

Para calcular la media a partir de las frecuencias relativas se usa la fórmula: $X = \sum X_i\ f_i$

[1] Los valores poblacionales se llaman parámetros, mientras que los valores muestrales se llaman estadísticos o estadígrafos.

2.3. MEDIA PONDERADA. La media de datos agrupados se considera una media ponderada por la frecuencia de las observaciones, pero no es el único caso de media ponderada.

Ejemplo 1.4. Las notas de un curso son el resultado de ponderar el promedio de los exámenes parciales por 0.4 y la nota del examen final por 0.6. Un estudiante que tuvo las siguientes notas: 1er parcial 50, 2o parcial 60, 3er parcial 100, Examen final 80. El promedio de los parciales es el siguiente (media simple): (50 + 60 + 100)/3 = 70. La nota final del curso es: 0.4*70 + 0.6*80 = 76

2.4. PROPIEDADES DE LA MEDIA. Tomando la convención x=X-X, llamada variable centrada,

i. La suma de los desvíos respecto de la media es cero: $\sum x = \sum (X - \bar{X}) = 0$

Ejemplo 1.2 (Cont.) Los desvíos acá son: (-52)(2) + (-32)(1) +(-12)(1) + (8)(1) + (28)(5)=0

ii. La suma de los cuadrados de los desvíos es menor con respecto a la media que con respecto a cualquier otro valor: $\sum x^2 \leq \sum (X - a)^2$ para cualquiera

iii. La media de una variable mas una constante es igual a la media de la variable mas la constante: $\overline{X} = a + \overline{d}$

Ejemplo 1.5. Un antiguo profesor mío proponía el siguiente ejercicio: calculemos la media de las edades de todos los que estamos en esta clase. Si nos encontramos dentro de 10 años para una fiesta de camaradería y se nos ocurre volver a calcular esa cantidad que habrá pasado?

iv. La media del producto de una constante por una variable es igual a la media de la variable por la constante:

$$b\bar{X} = b\bar{X}$$

Las propiedades iii y iv se pueden resumir así: $\overline{(a+bX)} = a + b\bar{X}$

v. La media de la suma de variables es igual a la suma de las medias:

$$\overline{(X+Y)} = \bar{X} + \bar{Y}$$

2.5. MEDIANA. La mediana es el valor de la variable que divide la distribución de tal modo que la mitad de los valores son iguales o menores que ella y la otra mitad son iguales o mayores. Si los datos no se repiten y no están agrupados para calcular la mediana basta con ordenarlos y contarlos: el que ocupe el lugar del medio es la mediana. Si hay un número par, muchos definen la mediana como el promedio de los dos valores intermedios.

Mediana para datos agrupados. Si los datos están agrupados aunque sea fácil identificar a la clase que contiene a la mediana, el valor no está unívocamente definido y se estila interpolar (ver Spiegel, 1970, pág.47-58). Para el caso de datos agrupados la mediana se puede calcular del siguiente modo:

$$Me = L + \frac{\frac{n}{2} - F_{j-1}}{f_j} b$$

donde L es el límite inferior de la clase que contiene a la mediana (llamada clase j), f_j es la frecuencia de la clase j, F_{j-1} es la frecuencia acumulada de la clase anterior a la clase j, y b es la longitud de clase. La mediana se usa en variables continuas y datos ordinales. No es sensible a valores extremos como la media.

2.6. OTRAS MEDIDAS DE POSICION.

Moda. La moda o modo es el valor más frecuente de la variable. Si los valores no se repiten no hay una moda única. Una distribución puede tener más de una moda, si tiene una sola es unimodal, de los contrario bimodal ,etc.

Ejemplo 1.3. (Cont.). En este caso la clase modal es la que va de 226 a 235, de modo que para muchos efectos se considera que la moda es 230. Opcionalmente, se puede interpolar como hace Spiegel (1969).

Ejemplo 1.3. (Cont.) Acá la moda es el valor 80 que es el más frecuente.
Notemos que la moda es especialmente valiosa en el caso de los datos nominales

Media Geométrica. La media geométrica es la nésima raíz del producto de los n valores. Eso equivale al antilogaritmo del promedio de los logaritmos.

Media Armónica. La media armónica es la inversa del promedio de las inversas. Es decir que se toman las inversas de las observaciones, se las promedia y se invierte el valor obtenido.

Media Cuadrática. La media cuadrática es la raíz cuadrada de la media de los cuadrados. Es decir que los valores se elevan al cuadrado, se promedian los cuadrados y luego se toma la raíz cuadrada.

Aunque estas últimas parecen artificiales complicaciones tienen aplicación en determinadas circunstancias aunque no con mucha frecuencia.

Ejemplo 1.7. Los siguientes valores muestran una característica de los dos padres y del promedio de los hijos. Como se observa la media geométrica se acerca mas al valor de la descendencia que la media aritmética.

Padre 1	Padre 2	Media Descendencia	Media Geométrica	Media Aritmética
54.1	1.1	7.4	7.7	27.6
57.0	1.1	7.1	7.9	29.1
173.6	1.1	8.3	13.8	87.4
53.0	5.1	23.0	16.4	29.1
150.0	12.4	47.5	43.1	81.2

Ejemplo 1.8. Un ejemplo de media armónica esta dado por el siguiente problema: Un coche recorre los 500 km (aproximadamente) entre Salto y Montevideo en 8 horas al ir y en 6 horas al volver. Cuál fue la velocidad media en el viaje de ida? Cuál fue la velocidad media en el viaje de vuelta? Cuál fue el promedio de la velocidad?

La velocidad en el viaje de ida es 500 km/8 h = 62,5 km/h. La velocidad en el viaje de vuelta fue de 500 km/6 h = 83,33 km/h. La velocidad promedio fué de 1000 km/14 h = 71,43 km/h. Esta velocidad no es la media aritmética de 62,5 y 83,33 (que es 72,92) sino la media armónica entre ellas:

$$\left[\frac{1}{62,5}+\frac{1}{83,33}\right]^{-1} = (0,016 + 0,012)^{-1} = 71,429$$

Aunque la diferencia no es muy grande se puede apreciar que representa estadísticas diferentes.

Ejemplo 1.9. Si se analizan los desvíos con respecto a la media (en el ejemplo 1.3. resulta fácil) se puede concluir que su media cuadrática tiene un significado muy especial, como veremos a continuación se le conoce como desviación estándar.

Práctico 2. Medidas de Posición.

1. Una población es un conjunto generalmente grande de individuos o medidas acerca del cual se desea información.

2. Una muestra aleatoria es aquella obtenida de tal modo que todos los individuos de la población tienen igual oportunidad de entrar en la muestra.

3. La media de una muestra siempre divide a los datos en dos partes iguales: una mayor o igual y otra menor o igual.

4. El promedio común, o media aritmética, es la medida de tendencia central mas comunmente usada y entendida.

5. Una medida de tendencia central es un valor cuantitativo que describe cuan dispersos están los datos en torno a un valor.

6.- Calcule el rango (diferencia entre menor y mayor), la media, la moda y la mediana de los siguientes datos: 5, 7, 6, 4, 2

7.- Haga lo mismo con los siguientes: 8, 3, 12, 7, 10, 6, 9, 9

8.- Se tomó una muestra de salarios de una empresa:

$ 380, $ 390, $ 420, $ 380, $ 370, $ 380, $ 480, $ 390, $ 380, $ 390

Construya una gráfica para representar la situación. Calcule la medida de tendencia central que los representa mejor y explique por qué la eligió. Encuentre la media para esos valores.

9.- Con los datos del ejercicio 1 del práctico anterior, determine la moda, la media; la varianza y la desviación estándar y la mediana y los percentiles 25 y 75.

10.- Con los datos del ejercicio 2 del práctico anterior. Calcule la media aritmética de los datos sin agrupar. Calcule la media aritmética de los datos agrupados. Cómo explica las diferencias?
Si se divide a los datos por 100, cuál es la media? Y si se les resta 20? Calcule la varianza usando los datos originales y la tabulación hecha y comente si le da diferencias. Calcule la mediana y el primer cuartil.

11.- Los tres tipos de animales tienen diferente peso:

A	B	C
14	12	6

11.1. Si las frecuencias con que ocurre cada caso son: 0,81; 0,18 y 0,01. ¿Cuál es la media de la población?

11.2. ¿Y si las frecuencias son: 0,64; 0,32 y 0,04?

CLASE 3
MEDIDAS DE DISPERSION

3.1.RANGO. Es la diferencia entre el mayor y el menor de los valores.

3.2.VARIANZA. La varianza de una muestra [2] se define como: $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$

mientras que la varianza de una población finita, de N elementos: $\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}$

El numerador de la varianza se le conoce como suma de cuadrados y se calcula generalmente como:

$$\Sigma(X - X)^2 = \Sigma x^2 = \sum_{1}^{n} X^2 - \frac{(\sum_{1}^{n} X)^2}{n}$$

A la cantidad (n-1) se le conoce como grados de libertad. En algunos casos se usa el valor (n-1) como denominador de la cuasi- varianza para no subestimar la varianza poblacional (ver clase 13). La varianza como medida de dispersión nos sirve para determinar si desviaciones observadas son usuales o notorias.

Ejemplo 1.2 (Cont.) La varianza de los datos del ejemplo 1.2 es 778,1955.

3.3. VARIANZA PARA DATOS AGRUPADOS. La varianza para datos agrupados se define como:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 n_i}{\Sigma n_i}$$

3.4. PROPIEDADES DE LA VARIANZA.

1. $V[a+bX] = b^2\ V[X]$

La varianza es invariante respecto a un cambio de origen (sumarle una cantidad igual a todos los valores), pero no es invariante respecto a un cambio de escala (multiplicar por una constante los valores).

Continuación del ejemplo 1.5.Qué pasara con la varianza de las edades de todos los que estamos en esta clase si nos encontramos dentro de 10 años para una fiesta de camaradería y se nos ocurre volver a calcular esa cantidad?

2. $V[X \pm Y] = V[X] + V[Y] \pm 2\ Cov[X,Y]$

3.5. DESVIACIÓN ESTÁNDAR. La desviación estándar o desviación típica es la raíz cuadrada (positiva) de la varianza: $S = +\sqrt{S^2}$. La desviación estándar tiene la ventaja de que se expresa en las mismas unidades que la variable en estudio, pero no tiene las propiedades matemáticas de la varianza, por lo que la consideramos un subproducto de la varianza.

Ejemplo 1.2. (Cont.): La desviación estándar de los datos del ejemplo 1.2. es: $S = \sqrt{778,1955} = 27,89615$

[2] Le daremos el nombre de cuasi-varianza cuando se divide por (n-1)

3.6. COEFICIENTE DE VARIACION. El coeficiente de variación es el cociente entre la desviación estándar y la media: CV = S_X/X

Muchas veces el coeficiente de variación se expresa en porcentaje: CV = S_X*100/X

Ejemplo 1.1 (Cont.). El coeficiente de variación del ejemplo 1.1 es:

$$\frac{27{,}89*100}{225} = 12{,}39\%$$

El coeficiente de variación se utiliza para comparar la variabilidad de características que tienen diferentes unidades de medidas. Supongamos que a un investigador le interesa saber si dos poblaciones varían más en poder adquisitivo (medido en pesos de ingresos) o en educación (medida a traves de los años de estudio). Resulta difícil comparar pesos contra años, por lo que puede acudir al coeficiente de variación.

3.7. CUANTILES. Los cuantiles, de los cuales los mas usados son los percentiles, son valores de la variable que dividen la distribución en determinadas partes, por ejemplo los percentiles en 100. Constituyen una extensión del concepto de la mediana, que divide la distribución en dos por lo que es el percentil 50. Por supuesto que también se puede decir que la mediana es un caso particular de percentil. Por la forma que definen la distribución constituyen medidas de dispersión al mismo tiempo que de posición.

3.8. OTROS. Otras medidas de dispersión incluyen el rango semi- intercuartlico, rango de percentil (diferencia entre el percentil 90 y el percentil 10) y la desviación media. Otros momentos incluyen los coeficientes de asimetría (no decir sesgo) y de curtosis (achatamiento). Los conceptos de variables reducidas y de variables estandarizadas tambien vienen al caso acá.

Práctico 3. Medidas de dispersión.

1.- Encuentre la mediana y la desviación estándar de los siguientes datos: 95, 86, 78, 90, 62, 73, 99

2.- Determine la media, la mediana y la moda de los siguientes datos: 10 7 14 19 17 17 16 16 16 20 15 14 12 15 8

3.- Encuentre la media y la varianza de la siguiente distribución:

Limites	Frecuencia
78 - 85	5
85 - 93	8
93 - 101	11
101 - 110	4
110 - 117	5
117 - 125	2

4.- Supongamos que en un estudio en la ciudad de Salto se obtuvieron los siguientes datos:

Hijos	Familias
0	32
1	17
2	21
3	14
4 o mas	8

a. Calcule el número promedio de hijos por familia.
b. Calcula la varianza.
c. Calcule la mediana.
d. Calcule la desviación estándar y el coeficiente de variación.
f. Grafique los datos de un modo adecuado.

5.- Con los siguientes datos: 86 87 56 93 94 93 73 79 80 79 58 91 77 82 74 66 83 75 49 68 74 63 58 72 96 98 74 86 88 91
a. Construya una tabla con 5 intervalos de igual amplitud.
b. Construya un histograma.
c. Calcule la media.
d. Calcule la mediana.
e. Calcule la desviación estándar.
f. Calcule el coeficiente de variación.

6.- Los siguientes datos representan la frecuencia con que se presentan diferentes alturas de cuatro poblaciones de plantas en centímetros.

Cm	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
P1	4	21	24	8													
P2									3	11	12	15	26	15	10	7	2
F1					1	12	12	14	17	9	4						
F2		1	10	19	26	47	73	68	68	39	25	15	9	1			

Calcule la media, varianza y coeficiente de variación para cada población (cada fila). ¿Cuál es la mas variable?

7.-Los siguientes son datos de presión de personas de diferentes edades.

Edad	100	110	120	130	140	150	160	170	180	190	200	210	TOTAL
Menos de 20	1	7	20	9	2								39
20-29	7	26	62	51	14								160
30-39	2	13	65	47	18	2	3	3	4				157
40-49	2	3	14	11	20	5	7	2	1			1	66
50 y más	.	1	4	4	6	6	5	3	10	1		2	42
TOTAL	12	50	165	122	60	13	15	8	15	1		3	464

CONTESTE SI ES CIERTO O FALSO Y SI ES FALSO DIGA COMO CAMBIA LAS PALABRAS SUBRAYADAS PARA HACER VERDADERA LA FRASE.

7.- La suma de cuadrados de los desvíos respecto a la media $\sum(X-X)$ puede ser negativa a veces.

8.- El 1er. y el 3er. cuartil encierran la mitad de las observaciones.

9.- Para cualquier distribución la suma de los desvíos con respecto a la media es cero.

10.- La desviación estándar de los siguientes números es 2: 2, 2, 2, 2, 2

11.- En un exámen Juan está en el percentil 25 y Pedro está en el percentil 50, por lo tanto Pedro sacó el doble de puntos que Juan.

12.- 25% de los datos están entre el primer y tercer percentil.

13.- Qué ventajas y desventajas tiene la media aritmética como indicador de los ingresos económicos de los habitantes del Uruguay? Con qué otra medida se puede solucionar esa desventaja?

14.- La media aritmética de las alturas de un grupo de personas es 160 cm. con desviación estándar 16 cm. Mientras que su peso medio es 70 kgs. con desviación estándar 7 kgs.

14.1. Qué medida presenta mayor variabilidad, el peso o la altura?

14.2. Si se descubre que un desperfecto de la balanza utilizada hizo que marcara 2 kgs. de mas de cada observación. Se mantiene la misma respuesta al punto 14.1.?

15.- Los precios pagados por los mayoristas a los productores avícolas tienen una media de $U 7,50 y una desviación estándar de 25 cts. Al presentarse ante el Organismo Regulador de los Precios, qué conviene ms a los productores: un aumento de 75cts. en los precios de todos sus productos, o uno del 10% sobre los precios actuales? Sugerencia: analice cómo se modifica el precio medio y la desviacin estndar. Complete el problema calculando el coeficiente de variación.

16.- Si cada uno de los valores de la variable del problema 2 se incrementa en dos unidades, qué sucede con el valor del coeficiente de variación? y si se incrementa en un 20%?.

CLASE 4
MEDIDAS DE COVARIACION y CORRELACION

4.1. COVARIANZA. Una parte importante de describir un conjunto de datos es proporcionar la relación que existe entre dos o mas variables cuantitativas. Este tema será discutido con mas detalle en el futuro pero acá presentamos a la covarianza.

Figura 4.1. Cambio de coordenadas y coeficiente de correlación.

Tomando: $x = \overline{X} - X$, $y = \overline{Y} - Y$ se logra un cambio de ejes coordenados, porque el nuevo sistema (x,y) tiene su origen en el punto (X,Y) del anterior. Los valores de X mayores que la media tendrán x mayor a cero, estando ubicados a la derecha de la gráfica; en tanto, los valores de Y mayores que su media tendrán valores de y positivos, estando por encima del eje x. Tomando los productos de ambas variables reducidas x.y, observamos que tienen signo positivo en el primer y tercer cuadrantes, mientras que tienen signo negativo en el segundo y cuarto; tomando la sumatoria de esos productos para cada par de valores X,Y se puede visualizar su alineación. La sumatoria será positiva en caso de alineación del primer al tercer cuadrante, será negativa en caso de alineación del segundo al cuarto, y nula si la distribución es uniforme. La estadística as obtenida presenta dos inconvenientes. El primero es que depende del tamaño de la muestra, lo que se soluciona tomando el cociente entre la sumatoria de productos y el tamaño de la muestra, con lo que se obtiene la covarianza muestral. El segundo inconveniente es la dependencia de las unidades de medida, lo que se soluciona dividiendo por las desviaciones estándar de ambas variables. El coeficiente de correlación así obtenido vara, pues, entre -1 y +1. Cuando los puntos se alinean perfectamente con pendiente negativa vale -1, cuando la alineación es perfecta con pendiente positiva es +1 y los casos intermedios corresponden a diagramas de dispersión elípticos.

Figura 4.2. Valores del coeficiente de correlación.

Ejemplo 1.7. Los datos que se muestran en la tabla corresponden a dos variables, llamemosle X y Y.

	X	Y	x=X-X	y=Y-Y	xy
	0	4	-2	2	-4
	1	2	-1	0	0
	2	3,5	0	1,5	0
	3	0,5	1	-0,5	-0,5
	4	0	2	-2	-4
TOTALES	10	10	0	0,0	-8,5

Las respectivas medias son: $\bar{X} = 2$ y $\bar{Y} = 2$. En la tercer y cuarta columnas se presentan los desvíos con respecto a las medias de los valores de X y de Y, se puede verificar que suman cero. Finalmente, en la quinta columna se presentan los productos. La covarianza es el promedio de esos productos de desvíos con respecto a la media:

$$S_{xy} = \frac{\sum xy}{n} = \frac{-8,5}{5} = -1,7$$

4.2. COEFICIENTE DE CORRELACION DE PEARSON

El coeficiente de correlación de Pearson es la covarianza dividida por el producto de las desviaciones estándares:

$$r = \frac{S_{xy}}{S_x * S_y}$$

En el ejemplo, $r = \frac{-1,7}{1,41 * 1,41} = -0.85$

Práctico 4. Medidas de correlación.

1.- Los siguientes datos corresponden valores de Y bajo diferentes niveles de X:

X	Y	x	y	x^2	y^2	x.y
0	1.220					
50	1.505					
100	1.565					
150	1.423					
200	1.438					
250	1.513					
TOTAL						

1.1. Complete el cuadro calculando los valores faltantes.

1.2. Calcule $\sum x^2 = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=}^{n} X_i)^2}{n}$

$$\sum y^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=}^{n} Y_i)^2}{n}$$

$$\sum xy = \sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}$$

Comprobando que el resultado es el mismo.

1.3. Calcular el coeficiente de correlación rectilínea de Pearson r y el coeficiente de regresión $b=\sum xy/\sum x^2$

2. La tabla siguiente muestra el CI (coeficiente intelectual) de un grupo de personas y su capacidad lectora (HL).

CI	HL
109	31,8
138	24,5
86	11,8
153	18,8
156	17,3
40	11,0
70	12,2
126	20,6
68	10,8
99	5,3
112	29,3
138	8,0
103	35,8
127	19,6
63	21,4

Calcule la correlación entre ambas variables.

3.- Los siguientes datos pertenecen a aumentos de peso en animales de distintas edades.

Dias de edad	Aumento diario promedio en kg

30	0.797
60	0.630
90	0.757
120	0.777
150	0.563
180	0.487
210	0.496

Calcule la media, varianza y desviación estándar para cada una de las dos columnas (variables) y el coeficiente de correlación.

4.- Los siguientes datos representan valores de un experimento.

	X1	X2	X3	X4
	4	6	-2	4
	7	8	-1	1
	6	7	-1	1
	4	3	1	1
	4	5	-1	1
	5	8	-3	9
	6	5	1	1
	4	7	-3	9
Total	40	49	-9	27

Calcule las medias, y desviaciones estándar. Cuál es la relación entre X1, X2 y X3? Y entre X3 y X4? Cómo afecta eso las medias y varianzas?

5.- En un estudio se registraron 12 diferentes valores:

Muestra	X1	X2	X3	X4	X5	X6
1	0.782	183	0.793	44	0.562	0.678
2	0.817	188	0.884	140	0.610	0.747
3	0.763	189	0.873	86	0.522	0.698
4	0.815	189	0.873	122	0.575	0.724
5	0.775	189	0.873	58	0.628	0.751
6	0.720	191	0.698	65	0.507	0.603
7	0.782	191	0.698	33	0.594	0.646
8	0.759	191	0.698	39	0.564	0.631
9	0.853	192	0.976	114	0.520	0.748
10	0.807	192	0.976	17	0.632	0.804
11	0.800	194	0.964	28	0.615	0.790
12	0.856	194	0.964	9	0.504	0.734

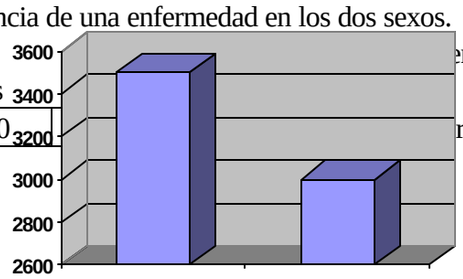
Calcule la media, varianza y desviación estándar de cada una de las variables (columnas) y la covarianza entre X1 y X2.

4.3. DESCRIPCIÓN DE UNA VARIABLE CUALITATIVA Y UNA CUANTITATIVA.

Ya vimos criterios para describir variables cuantitativas (media, varianza, correlaciones) y cualitativas (frecuencias). ¿Qué sucede si una de las variables es cuali y la otra cuanti? Por ejemplo peso al nacer (cuanti) en dos sexos (niñas y varones). En este caso, lo mas frecuente es que el investigador le interese describir la cuanti para cada sexo:

	Peso Nacer
Niñas	3.000
Varones	3.500

Asociación entre variables cualitativas. Vista la conveniencia del estudio de correlación entre variables cuantitativas, los investigadores se ven tentados de aplicar las mismas ideas a las variables cualitativas. Pero la situación aquí es más difícil. Supongamos una situación en la que estamos estudiando la presencia de una enfermedad en los dos sexos.



No independientes

	Sanos	Enfermos	Total
Hombres	100		

ndientes

	Sanos	Enfermos	Total
res	50	50	100

Mujeres	0	100	100
Total	100	100	200

Mujeres	50	50	100
Total	100	100	200

En la situación de la izquierda, los sanos son todos hombres y las personas enfermas son todas mujeres. En situaciones como esta hablamos de enfermedad ligada al sexo. En la situación de la derecha tenemos que las personas enfermas y las sanas se distribuyen uniformemente entre los dos sexos. Tenemos por tanto, una situación en la que el ataque de la enfermedad es independiente del sexo. En la primer situación la incidencia de la enfermedad depende del sexo. En la práctica las situaciones no son tan claras así. Por ejemplo tenemos

CLASE 5
PROBABILIDAD Y VARIABLES ALEATORIAS

En ésta sección haremos una resumida presentación de conceptos básicos de probabilidades.

5.1.CONCEPTO DE PROBABILIDAD

Definición a "priori". La probabilidad de un suceso es el número de casos favorables sobre el número de casos totales.

Ejemplo 5.1. La probabilidad de caer cara en una moneda es 1/2 pues es uno de los dos posibles resultados.

Definición a "posteriori". La probabilidad de un suceso es el límite (si existe) de la frecuencia relativa cuando el tamaño de muestra tiende a infinito.

Ejemplo 5.2. La probabilidad de germinar de semillas de una determinada población es 80%. Esto se sabe porque en una serie de pruebas se obtuvo ese porcentaje de germinación. La idea básica es que el investigador llega a la conclusión de que haciendo pruebas con cantidades cada vez más grandes el porcentaje de germinación que se obtendrá será de 80%.

Enfoque axiomático. Algunos autores objetan que ambas definiciones son criticables. La definición clásica define probabilidad en término de casos equiprobables, es decir de igual probabilidad. O sea que para decir lo que es probabilidad necesitamos ya saber de antemano lo que significa probabilidad. **C** La segunda es en realidad una forma de decir (como veremos mas adelante) que la probabilidad es un parámetro y su estimador (la frecuencia relativa) tiende a él. Una alternativa más rigurosa es encarar el concepto de probabilidad con un enfoque axiomático: es un número entre 0 y 1 que cumple con determinadas propiedades, llamadas leyes de la probabilidad.

C o **D** esa es la cuestión según Gary.

Ejemplo 5.3. ¿Cuál es la probabilidad de nacimiento de un varón al nacer un niño? Si razonamos que hay dos sexos posibles se puede decir que la probabilidad es 1/2. No obstante hay ciertos estudios que indican que es mas probable que nazca un varón que una niña, algunos autores dicen que la probabilidad de nacer varón es de 0,51, otros incluso más alta. Esos estudios se basan en análisis de frecuencias y encontraron que era más frecuente el nacimiento de varones. En este caso el razonamiento inicial falló debido a que los dos sexos no son "equiprobables".

5.2.LEYES DE LA PROBABILIDAD.

Dos sucesos son **excluyentes** si la ocurrencia de uno impide la ocurrencia del otro, es decir la probabilidad de que ambos ocurran al mismo tiempo (probabilidad de la intersección) es cero.

Ley de suma de probabilidades. La probabilidad de uno u otro de dos sucesos (probabilidad de la unión de ambos) es la suma de las probabilidades individuales menos la probabilidad de la intersección:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= P(A) + P(B) \quad \text{si son excluyentes}$$

Dos sucesos son **independientes** si la ocurrencia de uno no afecta para nada la ocurrencia del otro. Es decir que la probabilidad de A dado B (probabilidad condicional de A dado B), P[A|B], es igual a la probabilidad sin condición de B : P[A|B] = P[B].

$$P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{n(A \cap B)/\, n(\Omega)}{n(B) / n(\Omega)} = \frac{P(A \cap B)}{P(B)}$$

Ley del producto de probabilidades. La ley del producto de probabilidades dice que la probabilidad de A y B simultáneamente es el producto de la probabilidad de uno de ellos por la probabilidad condicional del otro: $P(A \cap B) = P(A) * P(B|A) = P(B) * P(A|B)$

Ejemplo 5.4. Se sorteará un estudiante como delegado de la Regional Norte a Montevideo. En la clase son 4 hombres de los cuales 3 trabajan, y 16 mujeres de las cuales 7 trabajan. ¿Cuál es la probabilidad de que por azar le toque ir a un hombre? 4 en 16.
Cuál es la probabilidad de que le toque ir a alguien que trabaja? 0,50 ya que son 10 los que trabajan en un total de 20.
Cuál es la probabilidad de que sea un hombre que trabaja?
Es 3 en 20 lo que no es igual que el producto de las probabilidades marginales: (1/2)(1/4). Por lo tanto los sucesos no son independientes. Esto también se puede observar si notamos que los hombres que trabajan son 3/4 (es decir 75%) mientras que las mujeres que trabajan son 7/16 (lo que es alrededor de 0,4375. Es decir que es más probable encontrar una persona que trabaja entre los hombres que en las mujeres, o sea que la característica (el suceso) "trabajar" no es independiente del sexo.

Probabilidad total. Si un suceso ocurre necesariamente asociado con otro, la probabilidad de ocurrencia es la suma de las probabilidades que ocurra con subconjuntos de su condición.

Por ejemplo si un artículo es producido por tres máquinas: M1, M2 y M3. La M1 produce el 20 de los artículos, la M2 el 30% y la M3 el restante 50%. La probabilidad de que la M1 produzca un artículo defectuoso es de 2%, la de M2 es de 1% y la de M3 es de 3%. La probabilidad de obtener un artículo defectuoso en todo el taller es la suma de las probabilidades de obtener artículos defectuosos en cada una de las máquinas:

P = P [Def en maq 1] + P [Def en maq 2] + P [Def en maq 3]
= P[Def y de la M1]+ P[Def y de la M2] + P[Def y de la M3]

en el ejemplo: P=(0,20)(0,02)+(0,30)(0,01)+(0,50)(0,03) = 0,004 + 0,003 + 0,015 = 0,022

Teorema de Bayes. La probabilidad de que un artículo defectuoso sea de una máquina en particular (la máquina 1 por ejemplo) en el caso anterior es una situación donde se aplica el teorema de Bayes:

$$P[\text{Maq 1}|\text{ Art Defectuoso}] = \frac{P[\text{art defectuoso y máquina 1}]}{P[\text{artículo defectuoso}]}$$

En el ejemplo, P[M1| D] = 0,004/0,022 = 0,1818

Aplicaciones de las ideas de probabilidad. Las ideas presentadas aqui son de gran aplicación en el área de la salud. Un ejemplo es para diagnóstico:

Situación ideal

	Sanos	Enfermos	Total
Negativo	100	0	100
Positivo	0	100	100
Total	100	100	200

Situación real

	Sanos	Enfermos	Total
Negativo	720	10	730
Positivo	180	90	270
Total	900	100	1000

Deseamos que el método de diagnóstico nos diga positivo siempre que el paciente esté enfermo y negativo siempre que esté sano (que ningún paciente sano aparezca como enfermo)
Sensibilidad del diagnóstico P[+|E] = 1 Probabilidad de diagnóstico positivo en un paciente enfermo sea igual a 1.
Especificidad P[-| S] = 1 Probabilidad de diagnóstico negativo en un paciente sano sea igual a 1.

En el presente ejemplo la sensibilidad es S=90/100=0,9 y la especificidad es E=720/900=0,8

Ejercicio 2. La probabilidad de que se diagnostique una cierta enfermedad cuando ésta existe es del 90% y de que no se diagnostique cuando no existe es el del 80%. La probabilidad de que una persona tenga la enfermedad es del 1%. Hallar la probabilidad que una persona con diagnóstico positivo tenga efectivamente la enfermedad.

Práctico 5. Probabilidad.

1.- ¿Qué es la probabilidad de un suceso?
2.- Si tenemos una caja con 1 bolilla azul, 2 rojas y 3 amarillas y extraemos una bolilla al azar cuál es la probabilidad de que sea roja? Si extraemos una roja, cuál es la probabilidad de que la segunda que extraemos también sea roja? Fundamente ambas respuestas.
3.- De cada 200.000 niños nacidos vivos en USA en años pasados vivían los siguientes a las edades indicadas:

EDAD	20	40	60
Varones	95.743	90.183	65.704
Mujeres	97.013	93.969	79.982

¿Cuál es la probabilidad de que una niña recién nacida viva hasta los 40 años?
4.- Si A es un suceso cualquiera, P(A) está siempre entre dos límites ¿Cuáles son?
5.- Cuánto suman P(A) y P(no-A) ?
6.- Cómo se llaman A y no -A (o $\overline{A}$ o A^C) ?
7.- Cuál es la probabilidad de que una persona de 20 años del sexo masculino viva hasta los 40, de acuerdo a los datos de la pregunta 2?
8.- Cómo se llama ese tipo de probabilidad?
9.- Qué son dos sucesos independientes?
10.- Si H_1, H_2, ..., H_n son sucesos mutuamente excluyentes cuya unión es el espacio muestral y si el suceso A debe presentarse con uno de ellos, cuánto vale P(A) en función de ello? Cómo se llama esa Ley?
11.- Cuál es la probabilidad de H_1 si se dió A? Cómo se llama esa ley?
12.- Qué es una variable aleatoria ?
13.- Cuáles son variables aleatorias contínuas y cuáles discontínuas o discretas?
14.- Se tiran dos dados uno rojo y uno verde: Cuál es la probabilidad de que caiga un 6 en el dado rojo?
15.- Cuál es la probabilidad de que caiga un número menor de 3 en el verde?
16.- Cuál es la probabilidad de que caiga un seis en el dado verde y un número menor de 3 en el rojo?
17.- Cuál es la probabilidad de que caigan dos números cuya suma sea menor de cuatro?
18.- Si cayeron dos números que suman menos de cuatro, cuál es la probabilidad de que el dado verde tenga un 1?

Ejercicio 20. Muchas enfermedades genéticas son recesivas, si dependen de un solo gen solos los animales doble recesivos adquieren la enfermedad. Si se cruzan dos animales heterozigotos,
2.1. ¿Cuál es la probabilidad de obtener un homocigoto doble recesivo (o sea enfermo)?
2.2. Si nacen diez animales hijos de heterocigotos ¿cuál es la probabilidad de que todos sean sanos? (Esto se usa mucho en pruebas de progenie)
Ejercicio 21. En una ciudad el 70% de los adultos escucha radio, el 40% lee el periódico y el 10% ve televisión. Entre los que escuchan radio, el 30% lee periódicos y el 4% ve televisión. El 90% de los que ven televisión lee el periódico y solo el 2% de la población total adulta lee el periódico, ve televisión y escucha radio. Si se elige una persona al azar se pide la probabilidad:

1.1. De que lea el periódico, escuche radio o vea televisión.
1.2. Sabiendo que lee el periódico, la de que vea televisión.

Ejercicio 22. En una zona de explotación papera la producci'on obtenida tiene determinado porcentaje de papas en mal estado. El detalle de los establecimientos agrícolas en el siguiente:

ESTABLECIMIENTO	AREA DE LA EXPLOTACION	PRODUCTO OBTENIDO	
		Normal	Malo
A	10 Há	50%	50%
B	15 Há	70%	30%
C	25 Há	85%	15%

Los tres establecimientos forman parte de una cooperativa de producción, la que embolsa el producto en bolsas de 50Kg. Se supone que la composición de cada bolsa (con respecto a

papas malas o normales) es la misma que la del establecimiento en que fueron producidas. La producción obtenida en cada uno de los establecimientos es proporcional al área de la explotación. Si todas las bolsas se envían al mercado, y un comprador abre una al azar y extrae de ella una papa, cuál es la probabilidad de que la bolsa provenga del establecimiento A, si la papa extraída resultó ser buena?.

Ejercicio 23. La probabilidad de que se diagnostique una cierta enfermedad cuando ésta existe es del 90% y de que no se diagnostique cuando no existe es el del 80%. La probabilidad de que una persona tenga la enfermedad es del 1%. Hallar la probabilidad que una persona con diagnóstico positivo tenga efectivamente la enfermedad.

Ejercicio 24. Muchas enfermedades genéticas son recesivas, si dependen de un solo gen solos los animales doble recesivos adquieren la enfermedad. Si se cruzan dos animales heterozigotos,
4.1. ¿Cuál es la probabilidad de obtener un homocigoto doble recesivo (o sea enfermo)?
4.2. Si nacen diez animales hijos de heterocigotos ¿cuál es la probabilidad de que todos sean sanos? (Esto se usa mucho en pruebas de progenie)

CLASE 6
VARIABLES ALEATORIAS

6.1. CONCEPTO DE VARIABLE ALEATORIA. Una variable aleatoria es una variable que toma valores al azar, es decir que cada valor de la variable tiene asociada una determinada probabilidad de ocurrir. Por lo tanto las variables aleatorias miden alguna característica de un experimento aleatorio: si muestreamos una serie de estudiantes el coeficiente de inteligencia de 1 estudiante tomado al azar es una variable aleatoria. En el ejemplo de la elección de las edades de los niños no hay experimento aleatorio en la elección de las X y por lo tanto la variable no es aleatoria.

6.2. VARIABLES ALEATORIAS DISCRETAS Y CONTINUAS. Como vimos antes las variables pueden ser discretas o continuas. Nos interesa distinguir las variables aleatorias en ese sentido pues algunas propiedades cambiarán de acuerdo con eso.

Función de cuantía. La probabilidad de cada valor de la variable es positivo o cero (pero no negativo) y la suma de todos los valores es 1.

Función de cuantía: $P(X = x) \geq 0$ y $\sum p(x) = 1$

Función de densidad de probabilidad. Si la variable X es continua, la probabilidad de cada valor x se simboliza con f(X=x) y sigue siendo un número no negativo. La suma de todos los valores, que ahora se representa por la integral de menos a mas infinito (es decir entre todos los valores posibles de la variable), es uno.

Función de densidad: $f(x) \geq 0$ y $\int_{-\infty}^{+\infty} f(X)dx = 1$

6.3.ESPERANZA MATEMATICA. La esperanza de una variable (que se simboliza con la letra E) es la suma del total de valores posibles de la variable multiplicados por la probabilidad de ocurrir que cada uno tiene: $E(x) = \sum x_i \, p(x_i)$. Si la variable es continua la sumatoria se reemplaza por la integral.

6.4.ESPERANZA DE UNA FUNCION. Similarmente, la esperanza de una función g(X) es la suma de: $E(x) = \sum g(x_i) \, . \, p(x_i)$

6.5.MOMENTOS. Los momentos son las esperanzas de las potencias de la variable. Por ejemplo el momento de orden v es la esperanza de la potencia v: $E[x^v] = \sum x^v \, . \, p[x_i]$. Los momentos centrados con respecto a la media son las esperanzas de las potencias de los desvíos respecto a la media. Por ejemplo el momento de orden v con respecto a la media es la esperanza de la potencia v del desvío con respecto a la media: $E[X-\mu]^v = \sum [X-\mu]^v . \, p[X]$

Para distinguir a los momentos centrales de los definidos previamente a éstos se les dice momentos ordinarios.

6.6. LA MEDIA Y LA VARIANZA COMO MOMENTOS. La esperanza es una idealización del concepto de media aritmética, de modo que se dice que la media es la esperanza de una variable. La varianza es el momento de segundo orden con respecto a la media, es decir es la esperanza del desvío cuadrático: $\sigma^2 = E(X - \mu)^2$

6.7. FUNCIÓN GENERADORA (GENERATRIZ) DE MOMENTOS. Si tenemos una función tal que su derivada n-ésima sea el n-ésimo momento de la variable se puede considerar que esa función "genera" los momentos de la variable. La función generadora de momentos no siempre existe, pero si existe es única. Por lo tanto se usa para identificar a la distribución. Por ejemplo, el Teorema del Límite Central se demuestra al comprobar que la función generadora de momentos de una suma de variables se aproxima a la de una normal.

Práctico 6. Concepto de Variables Aleatorias.

1. En un negocio de venta de frutillas tenemos la siguiente situación:

Ventas	Dias	Probabilidad
10	15	
11	50	
12	40	
13	25	
	130	

o sea que en 15 días se vendieron 10 cajas de frutillas, en 50 días se vendieron 11 cajas, etc. Calcule la probabilidad de vender 10 cajas, 11 cajas, etc.

2. La empresa compra las frutillas a $ 20 el cajón y lo vende a $50. Si todo sale bien gana $30 por cajón, si no vende el producto se pudre. Por tanto tenemos que las pérdidas por obsolescencia son de $20 y la pérdida de oportunidad $30. Con la ayuda del siguiente cuadro queremos calcular cuanto pierde el negocio, en diferentes situaciones. Completelo.

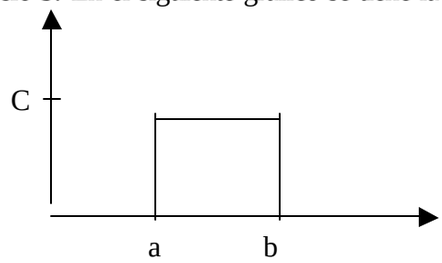
	Existencia			
Demanda	10	11	12	13
10				
11				
12				
13				

3. Si la empresa tiene en existencia 10 cajas tenemos este panorama:

Demanda	Pérdida Condicional	Probabilidad	Pérdida Esperada
10	0	0,15	
11	30	0,20	
12	60	0,40	
13	90	0,25	

4. Haga lo mismo para lo que ocurriría si la empresa tuviera en existencia 11 cajas, 12 cajas o 13 cajas. Grafique los resultados.

Ejercicio 5. En el siguiente gráfico se tiene la función de densidad f(X) de una variable aleatoria X:



5.1. Asigne valores a las constantes a, b y c de forma que se cumplan las propiedades de f(X).

5.2. ¿Cuál es la probabilidad de obtener valores de X tales que $a<X<(a+b)/2$?

Ejercicio 6. Una variable aleatoria toma valores entre 1 y 3:

$f(X)= aX$ para $1<X<3$

6.1. ¿Cuánto vale a?

6.2. ¿Cuánto vale la esperanza de X?

6.3. ¿Y la varianza de X?

Ejercicio 7. Una variable aleatoria tiene función de densidad:

$f(X) = ce^{-3X}$ si $x>0$

$= 0$ en caso contrario

7.1. Determine la constance c

7.2. $P(1<X<2)$

7.3. $P(X\geq 3)$

7.4. $P(X<1)$

haciendo el gráfico en cada caso.

Ejercicio 8. Una variable aleatoria tiene función de densidad de probabilidad:

$f(X) = cX^2$ para $1 \leq X \leq 2$

$= cX$ para $2 <X <3$

$= 0$ en caso contrario

8.1. Determine la constante c

8.2. Calcule la probabilidad de que X sea mayor a 2

8.3. Calcule $P(1/2 <X <3/2)$

FUNCIÓN GENERATRIZ DE MOMENTOS. La función generatriz de momentos es por definición: $E[e^{tX}]$
Recordemos que habíamos hablado de función generatriz de momentos como la función
Algún vivo se dio cuenta que derivando esa función r veces e igualando la variable t a cero se obtenían los momentos de la distribución que estamos estudiando.
∂

OTRAS DISTRIBUCIONES

BINOMIAL NEGATIVA

UNIFORME

GAMMA

BETA

CLASE 7
VARIABLES ALEATORIAS DISCRETAS: BINOMIAL

7.1. DISTRIBUCION DE BERNOULLI

Una variable que tiene una distribución de Bernoulli es una variable que tiene dos resultados posibles, generalmente uno se considera éxito y el otro fracaso, o se simbolizan con 1 y 0.

Ejemplo 7.1 Tirar una moneda es un experimento de Bernoulli.

7.2. BINOMIAL

Si un experimento de Bernoulli se repite n veces y la probabilidad de éxito no cambia, la suma de éxitos tiene una distribución binomial.

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

Ejemplo 7.2. Tirar 5 monedas constituye un experimento donde el número de caras sigue una distribución binomial pues en cada moneda la probabilidad es la misma.

Media y varianza de la Binomial. La media y la varianza de una distribución binomial con parámetros n y p, que se simboliza con B(n,p), es:

$$E(x) = np$$

$$V(x) = np(1 - p)$$

La distribución binomial está asociada a experimentos de muestreo con repetición o muestreo de poblaciones infinitas (casos en que la probabilidad de éxito no cambia).

7.3. POISSON. Si la probabilidad de éxito está dada por la siguiente expresión:

$$P[X=x] = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

la distribución se conoce como de Poisson, por el nombre del autor que la introdujo. El parámetro m es la media y la varianza simultáneamente de la distribución.

7.4. HIPERGEOMETRICA. Si la probabilidad es $p[X = x] = \dfrac{\binom{a}{x}\binom{n-a}{n-x}}{\binom{N}{n}}$

se dice que la variable X tiene una distribución hipergeométrica. La media y la varianza de la hipergeométrica son: E (X) = np

$$V(X) = = np(1-p)\frac{N-n}{N-1}$$

La distribución hipergeométrica está asociada al muestreo de poblaciones finitas sin reposición.

7.5.RELACIONES ENTRE LAS DISTRIBUCIONES DISCRETAS

p constante BINOMIAL
- p< 0,1 se aproxima por POISSON
- p > 0,1 se aproxima por NORMAL

b

p no constante: HIPERGEOMETRICA

Las aproximaciones se usan para valores grandes de n, digamos mayores a 50.

RESUMEN DE CARACTERISTICAS DE LAS DISTRIBUCIONES DISCRETAS.

Parmetros	Media	Varianza
Bernoulli	p	p(1-p)
Binomial	np	np(1-p)
Hipergeométrica N, n, p	np	$np(1-p)\frac{N-n}{N-1}$
Poisson	λ	λ

Práctico 7. Probabilidad Binomial.

1.- Hallar la probabilidad de que en el lanzamiento de tres monedas

1.1.Caigan tres caras.

1.2.En las dos primeras caiga cara y en la siguiente número.

1.3.En la primera caiga cara y en las siguientes número.

1.4.Caigan dos números y una cara, sin importar el orden.

1.5.Caigan dos caras y un número, sin importar el orden.

2.- Hallar la probabilidad de que una familia con 4 hijos tenga:

2.1.Al menos un varón.

2.2.Al menos un varón y una niña.

3.- De un total de 2.000 familias con 4 hijos cada una - En cuantas de ellas cabe esperar que haya

3.1.Al menos un niño?

3.2.dos niños

3.3.una o dos niñas?

3.4.ninguna niña?

4.- Cuál es el número esperado de varones en familias con 4 hijos?

5.- En general, ¿cuál es la media de una variable binomial? ¿Y la varianza?

6.- En un rodeo hay 200 animales de los cuales 2 están enfermos. Si un comprador se lleva 50 animales al azar -Cuál es la probabilidad de:

6.1.¿Llevarse los dos enfermos?

6.2.¿Llevarse algún enfermo?

7.- Ajuste distribuciones binomial y Poisson a los siguientes datos:

X	n_i
0	8
1	16
2	38
3	22
4	10
5	6

CLASE 8
VARIABLES ALEATORIAS CONTINUAS

8.1.DISTRIBUCION NORMAL

Se caracteriza por una medida de posición: la media y una medida de dispersión: la varianza o su raíz cuadrada la desviación estándar.

Ejemplo 8.1. Una generación de estudiantes tiene una distribución normal con media 600 y varianza 3.600.

Normal estandarizada. La distribución normal que tiene media 0 y varianza (por lo tanto desviación estándar) 1 se conoce como la normal estandarizada y se representa con z. Para estandarizar una variable basta con restarle la media y dividirla por la desviación estándar:

Si $X \sim N(\mu;\sigma^2) \Rightarrow z \sim N(0;1)$

Una propiedad importante de la distribución normal es que permanece ante una transformación lineal o sea que toda función lineal de una variable normal es normal.

Uso de tablas. La normal estandarizada viene en tablas, una de cuyas forma indica, para cada valor de z, la probabilidad de valores entre la media y ese valor. Nosotros asumimos que el lector dispone de tablas de ese tipo.

Importancia de la normal. La distribución normal es la base de la estadística llamada paramétrica. Tal vez el origen de su importancia esté en el Teorema del Límite Central que veremos mas adelante.

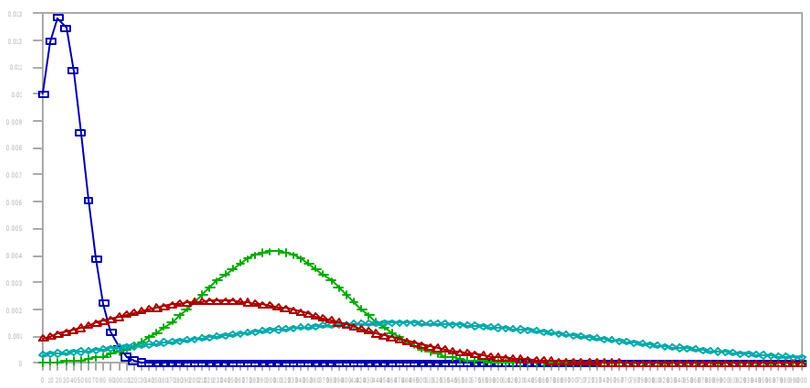


Figura 8.1. Cuatro distribuciones normales con diferentes medias y varianzas.

8.2.DERIVADAS DE LA NORMAL

Chi-Cuadrado. La suma de n normales estandarizadas independientes es una distribución χ^2 con n grados de libertad:

$$\chi^2 = \sum_{i=1}^{n} z_i^2$$

T de Student. La t de Student es el cociente entre una z (normal estandarizada) y una X2 dividida por sus grados de libertad:

$$t = \frac{z}{\sqrt{\frac{\chi^2_{(v)}}{v}}}$$

F de Snedecor y Fisher. La distribución F es el cociente entre dos χ^2 independientes divididas por sus respectivos grados de libertad: $F = \dfrac{\frac{\chi^2_{(a)}}{a}}{\frac{\chi^2_{(b)}}{b}}$

$$F_b^a$$

$$F_b^{a=1} = t^2_{(b)}$$

$$F^a_{b=\infty} = \chi^2_{(a)}$$

$$F^{a=1}_{b=\infty} = z^2$$

Figura 8.2. Relación entre las variables derivadas de la normal.

Práctico 8. Distribución Normal.

1.- Busque el área debajo de la curva normal a la derecha de z=1,52.
2.- Busque el área a la izquierda de z=1,52
3.- Calcular el área entre la media (z=0) y z=-2,1
4.- Calcular el área a la izquierda de z=-1,35
5.- Calcular el área entre z=1,5 y z=2,1.
6.- Calcular el área entre z=0,7 y z=2,1
7.- Cuál es el registro z con el percentil 75?
8.- ¿Qué valores de z encierran el 95% central de la distribución?
9.- El cociente de inteligencia (CI) se distribuye normal con media 100 y desviación estándar 10. Si una persona se elige al azar. ¿Cuál es la probabilidad de que su CI se encuentre entre 100 y 115: P [100 < CI <115] ?
10.- Encuentre el percentil 33.
11.- Cuál es la probabilidad de que una persona elegida al azar tenga un CI = 125?
12.- En un exámen las notas se distribuyeron normalmente con media 70 y desviación 15, dos estudiantes obtuvieron 60 y 93 puntos respectivamente.
    12.1.- Estandarice (es decir tipifique) los valores.
    12.2.- Encuentre el área de la curva normal entre z=0 y z=1,4
    12.3.- Encuentre el porcentaje de estudiantes que obtuvo nota superior al que obtuvo 60
    12.4.- Qué nota obtuvo el estudiante que integró el 25% superior con la nota más baja.?
    12.5.- Si eran 500 estudiantes y la nota de promoción era 70 puntos, cuántos salvaron?
13.- Al clasificar animales cuyos pesos están distribuidos normalmente, un 20% es pequeño, 55% mediano, 15% grande y 10% extra grande. Si el peso promedio es 680 kg y la desviación estándar 17 kg. Cuáles son los pesos mínimo y máximo entre los que un animal se considera mediano?
14.- Se lanza 500 veces una moneda, cuál es la probabilidad de:
    14.1.- ¿Obtener más de 400 caras?
    14.2.- ¿Obtener un número de caras que difiera de 250 en menos de 10?

Ejercicio 15. Si las estaturas de 10 000 estudiantes universitarios tienen una distribución normal con media 175 cm. y con desviación estándar de 6,25 cm.
    3.1. cuántos estudiantes tendrán por lo menos 180 cm. de estatura?
    3.2. entre qué valores se encuentra el 75% central de las mediciones?

Ejercicio 14. La gráfica corresponde a una función de una v.a. X normalmente distribuida en la población.
a) Hallar
b) Hallar la mediana.
c) Calcular el área rayada.
d) ¿ Qué porcentaje de individuos hay entre
e) ¿ Qué valor es superado poe el 90% de la población?

Ejercicio 15. La variable peso al nacer tiene una distribución normal con μ=3400 gr y xxx
a) ¿ Cuál es el intervalo central en el que se encuentra el 90% de la población?
b) ¿ Cuál es el intervalo central del 95 %?
c) ¿ A qué peso corresponde el percentil 10?
d) ¿Cuál es el porcentaje de niños con peso al nacer mayor de 4800?
e) Se considera de bajo peso al niño que al nacer pesa menos de 2500gr. Si un niño es de bajo peso, cuál es la probabilidad de que pese menos de 2100 gr al nacer.

Ejercicio 16. Se consideran normales los valores de hierro en sangre (sideremia) entre 40 mg/dl y 160 mg/dl (correspondiente a xx desvíos) teniendo la sideremia una distribución normal.
Se estudia una muestra de 2000 personas aparentemente normales.
a) ¿ Cuántas personas se espera encontrar con valores entre 30 y 100 mg/dl?
b) ¿ Cuántas personas se espera encontrar con valores mayores al percentil 90?

Ejercicio 17. En una población de lactantes varones de 3 meses de edad se estudia la distribución del peso corporal, considerada como normal con media 5,720 kg y varianza 0,8464 kg x.

a) Si se desea estudiar la población de lactantes cuyo peso no supere los 5.350 kg ¿qué porcentaje de niños sera estudiado?
b) ¿Cuál es el peso límite que es superado por el 90% de la población?
c) ¿Entre qué percentiles se encuentra un lactante cuyo peso es de 6 kg?

Tablas de la distribución normal, t, χ^2 y F.

CLASE 9
MUESTREO

9.1. INTRODUCCION AL MUESTREO. Muchas veces el universo de estudio en una investigación consiste en una población demasiado numerosa o no se cuenta con suficientes recursos para estudiarla en su totalidad. En esos casos es muy frecuente que se recurra al muestreo. **Universo o población** es el conjunto de individuos objeto de estudio, por lo tanto estará en función del objetivo de la investigación. **Muestra** es un subconjunto de la población que se pretende que represente a esta. En el proceso de sacar conclusiones para una población a partir de una muestra se cometen errores, estos pueden ser de dos tipos: sistemáticos y aleatorios. Los primeros, también llamados sesgos, se deben minimizar y los segundos se cuantifican. Serían errores evitables e inevitables, con los segundos hay que aprender a convivir. Uno de los objetivos del muestreo estadístico es conocer el grado de incertidumbre que tiene lo que estamos diciendo. Por ejemplo no conocemos el precio que tendrá un producto el año que viene, pero puede ser de utilidad decir: "estará entre 80 y 100 con un 95% de probabilidad". Generalmente no se dirá entre 80 y 100 sino 90 ±10. El valor 10 es lo que se conoce como márgen de error. Se intenta que el márgen de error sea pequeño. La precisión de la inferencia será mayor cuanto mas pequeño sea el márgen de error. Conviene distinguir entre **población muestreada y población objetivo.** La inferencia estadística proporciona herramientas para sacar conclusiones de la muestra hacia la población muestreada, la extrapolación a la población objetivo (si ambas no coinciden) es exclusiva responsabilidad del investigador.

9.2. TIPOS DE MUESTRAS. Los distintos tipos de muestras pueden ser descritos como:

- No probabilisticas o Finalistas
  - Casual
  - Intencional
  - Por cuotas
- Probabilisticas
  - Simple al azar
  - Sistematica
  - Estratificada
    - proporcional
    - no proporcional
  - Por conglomerados

9.3. MUESTRAS NO PROBABILÍSTICAS. En las **muestras no probabilísticas** los elementos de la población tienen una probabilidad desconocida de integrar la muestra. No tienen valor desde el punto de vista estadístico. Las muestras no probabilísticas se dividen en:

Muestras casuales, por ejemplo cuando un periodista entrevista a una de cada 10 personas que pasan por una calle. Aunque no las elija no pasan por una calle todos los integrantes de una población por lo que hay un sesgo desconocido.

Muestras intencionales, son aquellas en las que el investigador interroga solamente a ciertos informantes claves elegidos por el. Parece que tiene utilidad en investigaciones de tipo exploratorio.

Muestra por cuotas, utilizadas en investigaciones de mercado. A una serie de investigadores le es fijada una cuota de individuos a entrevistar y ellos seleccionan por su cuenta a los entrevistados. Generalmente se les proporciona alguna característica que los entrevistados deben reunir (mayores de edad, casados, etc)

9.4. MUESTRAS PROBABILÍSTICAS. En las **muestras probabilísticas**, cada elemento de la población tiene una probabilidad conocida de integrar la muestra. Los distintos tipos de muestras probabilísticas son: simple al azar, sistematica, estratificada y por conglomerados

Simple al azar, es la muestra en la que se eligen los integrantes al azar entre el total de la población. Requiere de un listado de los elementos de la población, su numeración y elegir al azar (por ejemplo usando una tabla de números aleatorios los que integraran la muestra). Sus ventajas son: es una metodología muy simple desde el punto de vista estadístico, tanto para llevar a cabo como para interpretar y utilizar; es insesgada, especialmente esta libre de los sesgos que introducirían las ponderaciones incorrectas que se puedan utilizar, no supone un conocimiento previo de la población de la cual se va a extraer la muestra; y, como consecuencia de esto, tiende a reflejar todas las características del universo. No obstante la simplicidad conceptual, puede ser muy difícil de llevar a la practica a veces y entonces aparecen las otras.

Muestreo sistemático es el que se sigue cuando se elige según un orden determinado, por ejemplo cada 10, se elige el primero el 11, el 21, etc. Se menciona la ventaja en la selección de la muestra y la desventaja es que si hay un gradiente ("trend") en el orden esta sesgando los resultados.

Muestreo estratificado es cuando la población se divide en estratos y se hace un muestreo aleatorio simple dentro de cada estrato. Ventajas: el estrato necesita una muestra más pequeña que el muestreo aleatorio simple. Desventaja: hay que saber hacer bien los estratos. Hay dos variantes acá: proporcional o no. En el primer caso el tamaño de la muestra de cada estrato es proporcional al tamaño del estrato ("la fracción de muestreo es igual para cada estrato"), en el segundo no.

Muestreo por conglomerados, utilizado en aquellos casos donde el universo a estudiar esta disperso a lo largo de áreas geográficas extensas o situaciones similares. Luego se elige un conglomerado, es decir uno de los grupos formados. Por ejemplo se elige una manzana de casas y en ella se entrevista a todas las personas que habitan en las casas de la manzana.

Enfatizamos la importancia que tiene desde el punto de vista estadístico el uso de muestras probabilísticas. Estas son las únicas en las que se puede aplicar la inferencia estadística que se verá más adelante.

Práctico 9. Distribución en el Muestreo.

1.- Dados los números 2, 4 y 6

1.1.- Grafique la distribución con un gráfico de barras.

1.2.- Determine la media y la varianza.

1.3.- Qué tipo de población se puede considerar que constituyen esos números? Qué muestreo se puede hacer en ella?

1.4.-Efectuando muestreo con reposición determine todas las muestras posibles de tamaño 2. Cuántas son? Cómo lo encuentra?Calcule las medias de esas muestras y grafique la distribución que tienen. Calcule la media y la varianza de esa distribución.

1.5.- Efectúe lo anterior para muestras de tamaño 3.

1.6.- Lo mismo para muestras de tamaño 9

1.7.- Efectúe los pasos 4 y 5 para muestreo sin reposición.

1.8.- ¿A qué concluye que es igual el promedio de las medias muestrales cuando el muestreo es con reposición? ¿Y cuándo sin reposición?

1.9.- ¿A qué concluye que es igual la varianza de las medias muestrales cuando el muestreo es con reposición? ¿Y sin reposición?

1.10.- A qué distribución se aproxima la de las medias muestrales?

2.- En la tabla adjunta se muestra una población normal simulada por un conjunto de 100 datos con media $\mu = 40$ y varianza $\sigma^2 = 144$. Sacando muestras de diversos tamaños con reposición.

2.1.- Calcule sus medias y grafique la distribución, calculando el promedio de ellas y su varianza ¿Qué distribución teórica se aproxima a la que tienen las medias?

2.2.- Calcule las varianzas y grafique la distribución que tienen calculando la media y la varianza.

2.3.- Grafique la distribución que tiene la suma de cuadrados de las muestras ¿A qué distribución teórica se aproxima?

100 valores tomados de una distribución normal.

34.6962	23.6426	27.6656	25.4427	36.8052	35.4854	13.6441	60.3879	49.4893	22.6428
37.0004	29.0456	40.7507	28.0678	40.9672	50.4255	40.9633	43.1284	44.3241	37.7500
42.6091	27.4148	58.8120	51.1407	8.5553	44.1250	26.9158	63.0216	31.4607	47.6411
40.8416	63.7941	39.6712	32.7833	39.6975	41.3788	38.1368	55.7218	38.4241	39.5010
55.4357	31.0474	26.8504	43.6270	37.4011	58.0712	50.0406	55.6076	48.7257	37.0704
40.7898	62.2072	48.5709	23.7161	53.4557	57.4617	52.2480	31.3830	54.5322	20.4457
67.0281	27.6250	31.7087	41.0106	59.0268	56.4032	44.6187	32.7486	44.9472	23.5218
57.9688	43.6631	36.7368	42.0180	21.5186	46.3358	47.0021	32.4908	48.3352	37.4687
17.4754	29.4158	44.5722	33.9291	46.0771	42.6079	60.1864	52.2163	43.1610	37.0984
5.1300	26.5412	39.9371	77.7104	30.5114	37.4290	36.3371	54.3501	34.9023	34.2470

3.- Conteste las siguientes preguntas:

3.1.- Una F con 1 y v2 grados de libertad es una...? Con V e infinitos? Con 1 e infinitos?

3.2.- Una normal (o,1) dividido χ / v es una ....?

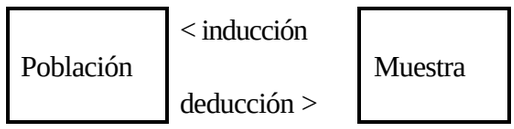
3.3. $\dfrac{\overline{Y} - \mu}{\widehat{\sigma}_{\overline{Y}}}$ se distribuye?

3.4.- ¿Cómo se distribuye el cociente de dos varianzas?

4.- Hay gente que confunde parámetros con variables. ¿Cómo las distinguirá Ud.?

CLASE 10
INFERENCIA ESTADISTICA

10.1.POBLACION Y MUESTRA. Los datos tienen diferentes distribuciones de las que nos interesa principalmente la distribución normal. La distribución que sigue el conjunto de todos los datos en consideración se denomina DISTRIBUCION POBLACIONAL y un subconjunto es una MUESTRA.



Sacar conclusiones de la población para la muestra (de lo general a lo particular) es hacer deducción, mientras que sacar conclusiones de la muestra para la población (de lo particular a lo general) es inducción o inferencia. La parte de la estadística que describe como hacer inferencia es la inferencia estadística. Nosotros trabajaremos con muestras aleatorias (obtenidas al azar) como representativas de la población. Se considera aleatoria la muestra en la que todos los integrantes de la población tienen igual probabilidad de ser elegidos. Si la muestra no es aleatoria los resultados de la teora estadística no son vlidos. Considerando los ejemplos 1 y 2 del curso de Estadística I, vemos que en el primer caso se puede intentar sacar inferencias sobre todos los animales de esa raza y condición (la población), mientras que en el segundo caso no tiene sentido decir que esos valores son una muestra representativa de alguna población, ya que si el investigador hubiese querido podra haber usado otras fertilizaciones a voluntad.

MODELOS. Como las poblaciones son muchas veces conceptuales (no reales) o infinitas se las define en un modelo. En el caso de los posibles rendimientos que pueden proporcionar parcelas del cultivo constituyen una población infinita o imaginaria. Un conjunto de supuestos con una estructura de predicción constituye un modelo. Al decir que el coeficiente de inteligencia de un grupo de estudiantes tiene una distribución normal con media 100 y varianza 600, estamos adoptando un modelo.

Modelo lineal aditivo. Muchas veces se postula que cada observación es la suma de una media mas un error aleatorio: $Y_i = \mu + \varepsilon_i = Y + e_i$

Este tipo de modelo se conoce como aditivo porque la variable Y se explica por la suma de μ y ε. Se le llama lineal debido a que ninguno de los parámetros está sometidos a multiplicaciones con otros parámetros. Cuando usemos modelos mas complejos la característica de lineal (en oposición a cuadrtáico, exponencial, etc.) aparecerá mas clara.

10.2.ESTIMACION DE PARAMETROS. Los modelos incluyen parámetros, como la media, la varianza y la proporción, que resultan desconocidos por lo que se intenta estimarlos a través de estadísticos, llamados estimadores por tal razón.

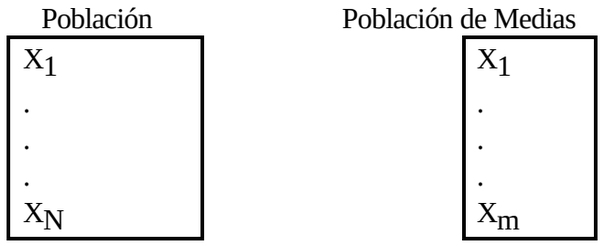
ESTIMADORES PUNTUALES Y POR INTERVALOS. Las estimaciones pueden ser puntuales o por intervalos. En las estimaciones puntuales se toma un valor para el parámetro, por ejemplo el valor mas probable. En las estimaciones por intervalos se toma un intervalo en el que se estima que el parámetro estará comprendido. Las primeras tienen mayor facilidad de uso en ciertas ocasiones, las segundas tienen una probabilidad conocida de ser correctas.
Veremos en este ejemplo que tenemos que la media de la variedad Población 1 puede ser estimada puntualmente por la media obtenida 6,63 o en intervalo diciendo que está entre 5,82 y 7,44 (como se calcula más adelante).

PROPIEDADES DE LOS ESTIMADORES. Existen una serie de propiedades deseables en un estimador:
1 - INSESGAMIENTO. Asegura que los investigadores que usan este método no se equivocan en promedio.
2 - EFICIENCIA. Dice que si se usa un método A que, por ejemplo, tiene 110% la eficiencia de B, entonces B necesita 110% el número de observaciones que necesita A para tener igual precisión.
3 - CONSISTENCIA. Indica que el estimación mejora con el aumento del tamaño de muestra.
4 - DISTRIBUCION CONOCIDA. Posibilita construir estimadores por intervalos.

METODOS DE ESTIMACION. Existen diferentes métodos de estimación, es decir métodos de encontrar estimadores de los que mencionaremos:
1 - ***Método de los momentos.*** Consiste en igualar los momentos de la población con los momentos muestrales. Por ejemplo se puede decidir estimar la media de la población por la media de la muestra.
2 - ***Método De La MAXIMA VEROSIMILITUD.*** Propone considerar como estimaciones los valores más probables del parámetro.
3 - ***Método de los MINIMOS CUADRADOS***, uno de los más importantes a nuestros efectos. Propone como estimador el valor que haga mínima la suma de cuadrados de los desvíos.
4 - ***Mínimo c^2.*** Para algunas situaciones en que se usa χ^2, se puede proponer como estimador el valor que minimice el χ^2.

10.3.DISTRIBUCIONES EN EL MUESTREO

Supongamos que de la población extraemos sucesivamente muestras todas de tamaño n, como se observa en la figura 10.1.



$$\mu_{\bar{x}} = \mu_x$$

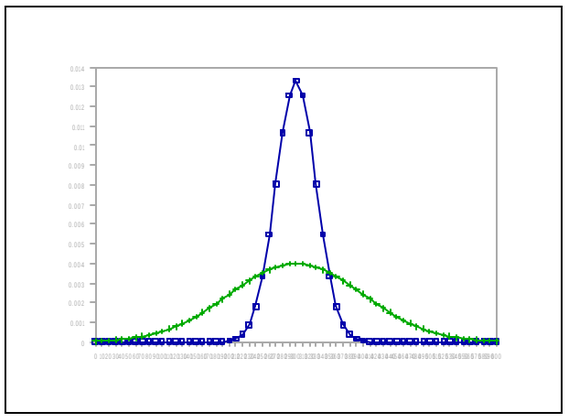
$$2\sigma_{\bar{x}} = \frac{\sum(\bar{x} - \mu x)^2}{m}$$

Figura 10.1. Población y población de medias muestrales.

Figura 10.2. Población original y población de medias.

$$\mu_{\bar{x}} = \mu_x$$

$$\sigma^2_{\bar{x}} = \frac{\sigma^2_{\bar{x}}}{n}$$



DISTRIBUCION DE LAS MEDIAS MUESTRALES.TEOREMA DEL LIMITE CENTRAL. Las medias de muestras aleatorias tienen distribución normal si provienen de poblaciones normales o tienden a distribuirse normalmente al aumentar el tamaño de las muestras si la distribución no es normal. La media de la población de medias muestrales es la media de la población, y la varianza es una enésima parte de la varianza poblacional.

Estandarización de la distribución de medias. Como toda distribución normal, la de las medias se puede estandarizar: $\frac{\overline{X}-\mu}{\sigma_{\overline{X}}}=\frac{\overline{X}-\mu}{\frac{\sigma_X}{\sqrt{n}}}$ donde $\sigma_{\overline{X}}$ es la desviación estándar de la variable medias muestrales llamada también el error estándar de la media.

Distribución t de Student. Si no conocemos o no podemos utilizar la distribución normal pero W. S. Gosset ("Student") construyó tablas con la distribución que tiene el cociente $(X- \mu)\sqrt{n}/s$ denominado por el ello con el seudónimo que él utilizó t de Student. La distribución de Student tiene un nuevo parámetro, los grados de libertad, y al aumentar éstos tiende a la distribución normal. Por lo tanto se puede considerar a la normal una t con infinitos grados de libertad.

OTRAS DISTRIBUCIONES EN EL MUESTREO. La varianza no tiene una distribución conocida, pero la suma de cuadrados (el numerador de la varianza) si la tiene:

$$\chi^2_{(n)}=\sum_{i=1}^{n} z_i^2=\sum_{i=1}^{n}\left[\frac{X_i-\mu}{\sigma}\right]^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2}+\frac{n(\overline{X}-\mu)^2}{\sigma^2}=\chi^2_{(n-1)}+\chi^2_{(1)}$$

DISTRIBUCION c2. La distribución χ^2 tiene una propiedad, llamada reproductiva, de que una variable con distribución χ^2 y n grados de libertad, más otra con la misma distribución y m grados de libertad tiene distribución χ^2 con m+n grados de libertad.

10.4. PRECISION Y EXACTITUD DE UNA ESTIMACION. La exactitud de una estimación se refiere a la cercana entre la cantidad que se desea estimar, por ejemplo μ, y su estimador, en este caso X. La precisión de una estimación, X en este caso, se mide por el error estándar del estimador. Generalmente se escribe (ver Mood y Graybill [1976]):

$$E[\theta-\theta]^2 = E[\theta - E\{\theta\}]^2 + (E\{\theta\} + \theta)^2$$

Exactitud = Precisión + Sesgo^2

La exactitud se mide por el error cuadrático medio.

CLASE 11
INTERVALO DE CONFIANZA PARA LA MEDIA POBLACIONAL

11.0.ESTIMADOR PUNTUAL DE LA MEDIA. No vamos explorar demasiado en la idea de que el mejor estimador puntual de la media de una población es la media de una muestra tomada al azar de esa población, ya que es intuitivamente claro.

11.1. INTERVALO DE CONFIANZA CON LA DISTRIBUCION z

Observando: $P[-z_\alpha < z < z_\alpha] = 1-\alpha$ podemos reescribir:

$$P\left[-z_\alpha < \frac{\overline{X}-\mu}{\sigma_X/\sqrt{n}} < z_\alpha\right] = P\left[X - z_\alpha\frac{\sigma}{\sqrt{n}} < \mu < X + z_\alpha\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

Esto se puede interpretar diciendo que esos límites aleatorios encierran la media poblacional un $(1-\alpha)$% de las veces, de modo que un par dado la encierran con una confianza del $1-\alpha$. A ese par de valores determinado se le conoce como los límites de confianza para la media, y al intervalo que encierran se le dice intervalo de confianza para la media. Al valor $(1-\alpha)$ se le conoce como nivel de confianza del intervalo.

Ejemplo 11.1. Para conocer el peso promedio de un grupo de personas se tomo una muestra de 38 personas. La media muestral resultó ser de 74,3 kg. Construya un intervalo de confianza del 98% para la media de la población, si la desviación estándar es 14 kg.
Como el valor de tablas que encierra el 98% de la distribución z es 2,33 tenemos: $74{,}3 \pm (2{,}33)\ 14/\sqrt{38}$ lo que es 74,3 5,29 es decir que el intervalo es (69,01-79,59).

Ej. 1.11 (Cont.) Un intervalo de confianza para la media de P1 en el ejemplo 1.11 estará entre $6.63 \pm 1.96\ (0{,}81/\sqrt{57})$

11.2. INTERVALO DE CONFIANZA CON LA DISTRIBUCION t. Intervalo de confianza para μ en caso de σ^2 desconocida. Cuando la varianza es desconocida la fórmula anterior no se puede utilizar pero una expresión adecuada es: $P\left[-t_\alpha < \frac{\overline{X}-\mu}{\sigma_X/\sqrt{n}} < t_\alpha\right] = P\left[X - t_\alpha\frac{\sigma}{\sqrt{n}} < \mu < X + t_\alpha\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$

de modo que la inicia diferencia está en que en este caso se debe utilizar la variable t en lugar de la z.

11.3.TAMAÑO DE LA MUESTRA. Llamamos error máximo de la estimación a la mitad del ancho (el radio) del intervalo de confianza, y lo podemos simbolizar con d. También se puede entender que d es la diferencia entre la media muestral y la media poblacional d=|X-u|.

Tomando la expresión anterior, podemos escribir: $z = \frac{(\overline{X} - \mu)}{\sigma_X}\sqrt{n}$, de donde despejamos

$n \geq z_{\alpha} 2 \frac{\sigma^2}{d^2}$

De modo que para obtener una precisión (es decir una diferencia máxima entre la media y su estimación) d, la muestra tiene que tener un tamaño mínimo dado por la expresión anterior. Nótese que si el cálculo proporciona un valor fraccionario (caso frecuente en la práctica) se tiene que utilizar el número entero inmediato mayor para asegurar la precisión deseada.

Tamaño de muestra en caso de varianza desconocida. Del mismo modo la expresión para el cálculo del tamaño de muestra mínimo se debe ajustar al uso de la variable t. Pero como en este caso necesitamos saber el tamaño de la muestra para definir los grados de libertad de la t a utilizar, debe recurrirse a un proceso interactivo.

Práctico 11. Inferencia sobre la media poblacional.

1) Si las notas de un grupo de estudiantes de Ciencias Sociales se distribuyen normalmente con media μ =1.000 y varianza σ^2 = 400

1.1. -Entre que valores estará el 95% de las medias de muestras aleatorias de tamaño 9?

1.2.- -Y si el tamaño es 1?

1.3.- -Cómo se construye un intervalo de confianza del 95% para la media poblacional?

1.4.- -Si no conoce la distribución poblacional. -Qué resultados obtendría?

2) Los siguientes datos pertenecen a la velocidad máxima de automóviles 0 km., variable con distribucin supuestamente normal: 163 208 154 183 169

2.1.- Construya un intervalo de confianza del 95% suponiendo que la varianza fuera conocida igual a 676.

2.2.- Construya un intervalo de confianza del 95% suponiendo que no conoce la varianza poblacional.

2.3.- Prediga la velocidad de un auto 0 km., puntualmente y con un 95% de confianza basandose en lo anterior.

2.4.- -Qué tamaño de muestra mínimo es necesario para estimar la media poblacional anterior con un margen de error no mayor de 20 kgs.?

3. Se tomo una muestra de 64 medidas de una población contínua y la media de la muestra es 32,0. La desviación normal de la población es conocida por ser 2,4. Se va a calcular un intervalo de confianza del 0,90. Calcule las siguientes cantidades:

(a) $\overline{X}$ (b) α

(c) n (d) $1 - \alpha$

(e) $z_{(\alpha/2)}$ (f) $\sigma_{\overline{X}}$

(g) E (margen de error)

(h) límite de confianza superior

(i) límite inferior de confianza

(a) Explicar hasta que punto la siguiente afirmación es verdadera: "El nivel de confianza para un intervalo estimado es una probabilidad."

(b) Explique porque "El nivel de confianza no es una probabilidad cuando miramos al intervalo estimado después que este ha sido obtenido."

4. Suponga que a un intervalo de confianza se le asigna un nivel de confianza de $1 - \alpha = 0,95$. Como se usa el 0,95 en la construcción del intervalo de confianza?

5. Se cree que el tiempo en la cola de la caja de un supermercado se distribuye aproximadamente normalmente con una varianza de 2,25. (a) Una muestra de 20 clientes reveló un promedio de tiempo de compra de 15,2 minutos. Construya un intervalo de confianza del 95% para la estimación de la media de la población.

(b) Si la media de 15,2 minutos resultó frente a una muestra de 32 clientes, encuentre un intervalo de confianza del 95%.

(c) Que efecto tiene un tamaño de muestra mas grande en el intervalo de confianza.

Ejercicio 6. Una muestra arrojó los siguientes valores: 18,5 20,6 12,9 14,6 19,8 15,0

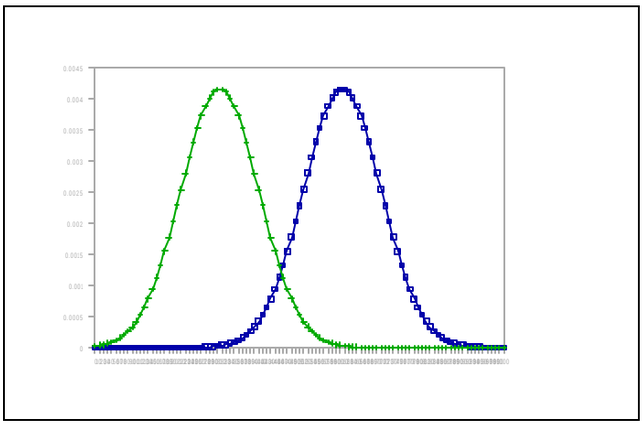
Determine los límites de 95% de confianza para la media de la población de la que se extrajo la muestra.

CLASE 12
PRUEBA DE HIPOTESIS SOBRE LA MEDIA.

12.1. CONCEPTO DE PRUEBA DE HIPOTESIS. La mecánica usual de una prueba de hipótesis $\mu_0 = \mu$ versus $\mu = \mu_1$ se muestra como sigue.

Supongamos por ejemplo que tenemos que decidir entre:

Figura 12.1.



H_0 : μ = 60

H_1 : μ = 80

sabiendo que ambas poblaciones tienen varianza igual a 36, con muestras de tamaño 4.

Una solución es el siguiente criterio: Si μ es 60 la probabilidad de que una media muestral, proveniente de una muestra aleatoria, sea mayor de 67,725 es menor de 0,05. Por lo tanto, considerando imposible se suceso se incurre en una probabilidad de error del 0,05. En base a esto podemos proponer como criterio de decisión: si la muestra tiene media mayor a 67.725 no aceptaremos que provenga de la población con media μ=60. Si la única opción es que provenga de una población con μ=80, entonces puede suceder que aceptemos Ho (que μ = 60) cuando en realidad la muestra provenga de la segunda población. Ese error se llama de tipo II y tiene una probabilidad de ocurrencia de

Supongamos que un investigador tiene un dato que puede provenir de dos poblaciones: ambas son normales con varianza 144 pero una tiene media igual a 40 y la otra media 60.
Tomemos a la primera población como base (llamémosle *hipótesis nula o básica*). Supongamos que nos preguntamos: Cuál es el valor de la variable que tiene un 95% de probabilidad de ser superado por azar, si efectivamente la observación proviene de una población con media 40? El valor de z que cumple con ello es 1,645 por lo tanto el valor de la variables es: 59,74 = 1.645*12+40.
Ahora preguntémonos (*hipótesis alternativa*): si la observación fuera en realidad de la población con media 60, cual es la probabilidad de que una observación sea por azar menor al valor 59,74?

$$P\left[z < \frac{59{,}74-60}{12}\right] = P[z < -0{,}0217] = 0{,}5 - \text{área}\,(0{,}0217) = 0{,}5 - 0{,}008 = 0{,}492$$

Si el investigador quiere tomar una decisión acerca de cual población provienen los datos puede tomar la siguiente regla o criterio de decisión: considero de la población con media 40 (es decir acepto la hipótesis nula o hipótesis base) si el valor es menor que 59,74 y de la otra (es decir acepto la hipótesis alternativa) si es mayor a 59,74.
Ese investigador puede cometer dos tipos de error. Eso generalmente se observa mejor en la siguiente tabla:

Decisión	Ho es cierta	Ho es falsa
Aceptar H_0	No hay error	Error tipo II
Rechazar H_0	Error tipo I	No hay error

A la probabilidad de cometer error de tipo I se le conoce como nivel de significación de la prueba de hipótesis y se simboliza usualmente con la letra griega alfa (α). A la probabilidad de cometer error tipo II se la simboliza con beta (ß). Al valor 1-ß se le conoce como potencia de la prueba de hipótesis. En el presente caso su probabilidad de equivocarse (que sea de la población con media 60 y diga que la media es 40) es de 0,05 y la probabilidad de que se decida que es de la población 2 siendo que era de la 1 es de 0,49.

El elemento básico de una prueba de hipótesis es la decisión que se toma, con lo cual se vuelve posible cometer error. Si se plantea la hipótesis de que la media de la población tiene un

valor dado y este valor hipotético resulta comprendido en el intervalo de confianza para la media poblacional, se puede aceptar la hipótesis. Si el valor hipotético de la media poblacional no está entre los valores posibles (intervalo de confianza de la media poblacional) se rechaza la hipótesis. Pero mientras que en un intervalo de confianza no hay posibilidad de error ya que no se toma decisión ninguna, al probar una hipótesis y tomar una decisión se incurre en la probabilidad de error. Por este motivo algunos autores dicen que el proceso es más que probar una hipótesis, es probar y decidir, por lo que proponen el término "docimasia" de hipótesis.

12.2.PRUEBAS UNI Y BILATERALES. A la prueba de hipótesis precedente se la describe como una prueba de hipótesis simple contra una alternativa simple. En los casos más comunes la hipótesis alternativa no es simple sino compuesta, por ejemplo decidir entre $\mu=60$ y $\mu>60$. En este tipo de hipótesis no se puede conocer ß; la prueba tiene una potencia que dependerá del verdadero valor de μ, el cual es desconocido.

12.3. PASOS PARA PROBAR UNA HIPOTESIS

1. Definir las hipótesis (H_0 e H_A)

*2. Elegir el nivel de significación (**a**)* Nos indica el porcentaje de probabilidad de acierto.
Ej. Un $\alpha=0.05$ nos indica un 95% de probabilidad. Este α es el más frecuente.

3. Elegir la variable, llamada a veces variable pivot (z o t) con la que vamos a trabajar. Recordemos que t se utiliza cuando no conocemos la varianza.

4. Definir el valor crítico y la región crítica. Valores que limitarán la Región de Aceptación. En caso de que la $H_A= \mu>60$ por ej. nos interesa un solo un valor para determinar la Región Crítica. que variará con respecto a una $H_A= \mu$ distinto de 60.

5. Hacer los cálculos.

6. Tomar la decisión rechazando o no rechazando la H_o.

Práctico 12. Prueba de hipótesis sobre la media.

1. La oficina de admisión de un colegio le dice a los aspirantes a estudiantes que sus libros cuestan en promedio $ 10.000, con una desviación de $ 2.500. Un grupo de estudiantes piensa que el costo promedio ha superado los $ 10.000 y decide probar la afirmación de la oficina de admisiones contra la alternativa de ellos, usando una muestra de tamaño 35.
(a) Si ellos usan $\alpha = 0.05$ ¿Cuál es el valor crítico para X que apoyará la teoría de ellos?
(b) Los datos de ellos se resumen por n = 35 y x = 393,25. ¿Es esta evidencia suficiente para rechazar la demanda de la oficina de admisiones?

2. El Sr. Pérez, un desilusionado pasajero que usa el sistema de ómnibus urbano, decidió llevar un registro de la cantidad de tiempo que perdía como resultado de la tardanza del ómnibus. X es el período de tardanza de su ómnibus en minutos. Después de varias semanas, sus registros mostraron el siguiente resumen: 80 ómnibus circulan, 223 fue el total de minutos de tardanza. Cuando el Sr. Pérez se quejó a la empresa de ómnibus, ellos respondieron que sus ómnibus no tardan más de dos minutos en promedio, con una varianza igual a 20. ¿Es la evidencia que presenta el Sr. Pérez suficiente para que la empresa de ómnibus rechaze la demanda? Usar $\alpha = 0.05$.

3. En un gran supermercado, el tiempo de espera de los clientes para comprar se distribuyen aproximadamente normalmente, con una desviación estándar de 2,5 minutos. Una muestra de 24 clientes perdiendo tiempo produjo una media de 10,6 minutos. Es esta suficiente evidencia para rechazar la afirmación del supermercado de que el tiempo de compras de sus clientes promedia en no más que 8 minutos? Completar esta prueba de hipótesis usando el nivel 0.02 de significancia.
4. El tiempo requerido por los estudiantes para registrarse en nuestra Facultad ha sido de 60 minutos, con una desviación normal de 10 minutos. Este semestre se introdujo un procedimiento nuevo más rápido de registración. Los estudiantes piensan que la gente de registros le erró a las cuentas y recogieron una muestra; las estadísticas que obtuvieron son n=40 y X=63,7. ¿Será su muestra suficiente evidencia, en el nivel 0.02 de significancia, para mantener su idea?
5. Un fabricante de cierto cigarrillo afirma que en promedio su cigarrillo no contiene más que 17,5 miligramos de nicotina con una desviación estándar de 2,5 miligramos. Una muestra de 32 cigarrillos seleccionados al azar fue examinada y se observó una media de 18,3 miligramos. Al nivel de 0,05 de significancia, tenemos evidencia suficiente para rechazar la información del industrial?

Ejercicio 6. Una variedad de trigo solo se tendrá en cuenta para posteriores ensayos si produce mas de 500 kg/há. Se plantaron 9 parcelas al azar y se obtuvo una media de 600 kg y una desviación estándar de 2.500 kg. ¿Se desecha la variedad?

Ejercicio 7. Un metalurgico realizó 4 determinaciones del punto de fusión del manganeso: 1269, 1271, 1263 y 1265 grados. ¿Está esto de acuerdo con el valor hipotético de 1260 grados?. Explique detallamente como resuelve el problema.

Ejercicio 8. En un cruzamiento de variedades de poroto se espera de acuerdo a la teoría genética que la mitad de las semillas producidas sean rugosas y la mitad lisa. Se tomó una muestra al azar de 40 semillas que consistía en 30 rugosas y 10 lisas. Pruebe la hipótesis mencionada con 10% de nivel de significación.

CONTESTE SI ES CIERTO O FALSO Y SI ES FALSO DIGA COMO CAMBIA LAS PALABRAS SUBRAYADAS PARA HACER VERDADERA LA FRASE

1. Beta es la probabilidad de un error de tipo I.
2. 1 - α se conoce como nivel de significación de una prueba de hipótesis.
3. El error estándar de la media es la desviación estándar de la muestra
4. El margen de error de una estimación es controlado por tres factores: nivel de confianza, tamaño de la muestra y desviación estándar.
5. Alfa es la medida del área de la curva de la variable que abarca la región de rechazo para Ho.
6. El riesgo de cometer un error de tipo I se controla en una prueba de hipótesis estableciendo un nivel para α.
7. El fracaso en rechazar la hipótesis nula cuando es falsa es una decisión correcta.
8. Si la región de aceptación de una prueba de hipótesis se hace más ancha (asumiendo que σ y n permanecen constantes) α se hace mas grande.
9. Rechazar una hipótesis nula que es falsa es un error de tipo II.
10. Para poder concluir que la media es mayor (o menor) que un valor postulado el valor de la variable pivot (variable usada en la prueba de hipótesis) debe caer en la región de aceptación.
11. La distribución t de Student es más dispersa que la distribución normal.
12. La distribución chi-cuadrado es usada para inferencias acerca de la media cuando la varianza poblacional es desconocida.
13. La distribución t de Student se usa para toda inferencia acerca de la varianza de una población.
14. Si el valor de la variable pivot (la variable usada para la pueba) cae en la región crítica, la hipótesis nula ha demostrado ser verdadera.
15. Cuándo la variable pivot es t y el número de grados de libertad es mayor de 30, el valor crítico de t e muy cercano al valor de z.?
16. Cuándo se hace inferencia acerca de una media en la que no se conoce el valor de σ (sigma), la variable que se usa como pivot es z ?
17. Sacar conclusiones de la población para la muestra es..... y de la muestra para la población es.....
18. Las únicas muestras que la estadística acepta son las muestras .....
19. La inferencia basada en muestras aleatorias es llamada .....
20. Las constantes poblacionales se llaman ..... y las muestrales .....
21. Los parámetros son normalmente desconocidos y por lo tanto se busca .....
22. La estimación de parámetros puede ser..... o .....
23. Las 4 propiedades deseables de los estimadores puntuales nombradas en clase son:
24. De los 4 métodos de estimación explicados en clase los dos más importantes son:
25. El método de los mínimos cuadrados se usa en las siguientes condiciones:.....
26. El método de máxima verosimilitud dice lo siguiente.....
27. Las medias muestrales se distribuyen..... Esa propiedad se conoce como .....
28. La media de las medias muestrales es..... y su varianza es .....
29. Por lo mencionado en 12 las medias muestrales se pueden estandarizar. ¿Cuando aparece la distribución t de Student? ¿Qué parámetros tiene la distribución t?
30. ¿Qué es la exactitud de una estimación, como se mide y como se relaciona con la precisión?
31. ¿Cómo se construye un intervalo de confianza para la media de una población con la distribución z?
32. ¿Y con la distribución t de Student?
33. ¿Cuál es mejor?
34. ¿Cómo se calcula el tamaño de muestra?
35. ¿Qué es el margen de error?
36. ¿Qué complicación aparece si queremos determinar un tamaño de muestra y no conocemos la varianza poblacional?

CLASE 13
COMPARACIÓN DE MEDIAS

13.1. MUESTRAS NO INDEPENDIENTES: OBSERVACIONES APAREADAS. Si deseamos comparar las medias de observaciones son apareadas el problema se transforma en uno de una sola muestra usando la variable diferencia: $d_i = X_i - Y_i$. La media de las diferencias es la diferencia de las medias: $\overline{d} = \overline{X} - \overline{Y}$ y la varianza de las diferencias es:

$s_L^2 = s_{\overline{X}}^2 + s_{\overline{Y}}^2 - 2\mathrm{Cov}[\overline{X}, \overline{Y}] = \dfrac{\sum_{i=1}^{m}\left(d_i - \overline{d}\right)^2}{m-1}$. El problema se redujo a una sola muestra de diferencias y se analiza como tal.

Ejemplo 13.1. Analizaremos la diferencia entre las medias de dos muestras dadas abajo.

Muestra 1	Muestra 2	d	d^2
161,30	149,64	11,66	135,96
148,26	163,30	-15,04	226,20
142,99	152,68	- 9,69	93,90
184,47	161,64	22,83	521,21
146,69	157,69	-11,00	121,00
164,11	146,08	18,03	325,08
162,31	170,04	- 7,73	59,75
171,22	173,27	- 2,05	4,20
170,08	146,70	23,38	546,62
161,27	157,89		

$t_{(7)} = \dfrac{\overline{d} - \mu_d}{\sigma_{\overline{d}}} = = 0.65197$ (ns) ya que:

$\hat{\sigma}_{\overline{d}} = \sqrt{\dfrac{\sum_{i=1}^{m}\left(d_i - \overline{d}\right)^2}{m(m-1)}} = \sqrt{\dfrac{\sum d_i^2 - \left(\sum d_i\right)^2}{m(m-1)}}$ = [2033.922-(30,39)^2/9]/9x8 = √ 26.82369

13.2. CONTRASTE DE 2 MEDIAS INDEPENDIENTES.

Generalmente en estadística no interesa tanto el efecto de un tratamiento como la comparación de efectos de dos o más tratamientos. En esta sección estudiaremos las comparaciones o contrastes entre medias. Para realizar este tipo de comparación nos valdremos de la propiedad reproductiva de la distribución normal que nos dice que toda función lineal de variables normales es normal. Se aplica como vemos:

$Y_1 \sim N(\mu_1 ; \sigma_1^2) \rightarrow Y \sim N(\mu_1 ; \sigma_{Y1})$ |

$> Y_1 - Y_2 \sim N(\mu_1 - \mu_2; \sigma_{Y1-Y2})$

$Y_2 \sim N(\mu_2 ; \sigma_2^2)$ $Y_2 \sim N(\mu_2 ; \sigma_{Y2})$ |

En el caso que las muestras sean independientes (es decir que la covarianza es cero) y si suponemos que tienen igual varianza poblacional:

$s_L^2 = s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ donde: $s_p^2 = \frac{\sum_{i=1}^{n1} x^2 + \sum_{j=1}^{n2} y^2}{n_1 + n_2 - 2} = \frac{s_1^2 + s_2^2}{2}$ la última igualdad vale si las muestras tienen igual tamaño. Si $n_1 = n_2$, entonces $\hat{\sigma}_L^2 = \frac{2\sigma^2}{n}$. Si, por el contrario, $n_1 \neq n_2$, entonces $\frac{1}{n} = \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ la media armónica representa bien al promedio de los tamaños muestrales.

Ejemplo 1.9 (Cont.) Supongamos que deseamos comparar las medias de los dos tratamientos mencionados antes, el tratamiento 1 y el 2. Tenemos: t =(104.4 - 177.5714)/

Ejemplo 14.1. Supongamos que estamos estudiando los tratamientos cuyos datos se proporcionan abajo y se intenta comparar la varianza del tratamiento 1 con la del 2.

Tratamiento	1	2
	62	163
	86	208
	117	154
	125	154
	132	183
		212
		169
Totales	522	1243
Medias	104,40	177,57

Las sumas de cuadrados son: SC = SC sin corregir - C. Para el tratamiento 1 tenemos 57978 -522^2/5 = 3481.2 por lo tanto la estimacion de la varianza es: 3481,2/5 = 870,3. Similarmente para el tratamiento 2, SC = 224259 - 1243^2/7 = 3537,714; con lo que la estimación de la varianza es: 3537,714/7 = 589,62. Un intervalo de confianza para la varianza del tratamiento 1 es: P[3481,2/11,14 < σ^2 < 3481,2/0,484] = P[312,4955 < σ^2 < 7192,561] =0,95. Similarmente para el tratamiento 2: P[251,4366 < σ^2 < 2093,322] = 0,95

Pruebas de Hipótesis Ejemplo 1.15 (Cont.) Supongamos que queremos probar la hipótesis de que la varianza de la población 1 es 1000. Según los pasos de la sección 7.3 tenemos:

1. *Definir las hipótesis*.
2. *Elegir el nivel de significación*, supongamos $\alpha = 0{,}05$
3. *Elegir la variable pivot*, en este caso será $\chi^2_{(4)} = \dfrac{\sum_{i=1}^{5}(X_i - \overline{X})^2}{\sigma^2}$
4. *Determinar la región crítica*: $\chi^2_c > \chi^2_t$
5. *Hacer los cálculos* $\chi^2_{(4)} = \dfrac{3481{,}20}{1000} = 3{,}4812$
6. *Tomar la decisión*. χ^2_c no pertenece a la región crítica, por tanto no rechazamos H_0.

Práctico 13. Comparación de medias.

Ejercicio 1. Los siguientes datos provienen de un experimento donde se aplicó una sustancia química en corderos mellizos a los efectos de determinar si aumentaba la población de folículos secundarios en la piel

Par	Tratados	Control
1	29,10	28,59
2	46,31	37,93
3	39,26	31,36
4	40,04	31,28
5	30,50	37,26
6	36,54	34,21
7	23,18	21,42

1.1. Pruebe la hipótesis que interesa al experimentador.
1.2. Construya un intervalo de 95% de confianza para el contraste.
1.3. Postule un modelo para los datos.
1.4. Qué ventaja tiene ese diseño experimental?

Ejercicio 2. Con los datos siguientes, provenientes del estudio de rendimiento de un cultivo bajo dos tratamientos:

Tratamiento	A	B
Nro. de plantas	44	36
Altura media	15,6	14,1
Suma de cuadrados	167,52	158,89

2.1. Pruebe la hipótesis de que provienen de tratamientos con igual rendimiento.
2.2. Calcule un intervalo de confianza para el contraste.

Ejercicio 3. Las producciones de dos variedades de maíz son las siguientes:
Variedad A. 1300 1350 1100 1400
Variedad B. 1800 1600 1900 1850 1750
Estimar las medias, varianzas y desvíos estándar.
Comparar las medias por la prueba t.
Obtener un intervalo de confianza para las medias al nivel de 95% de confianza.
Comparar las varianzas del ejercicio anterior por medio de la prueba F.

Ejercicio 4. Para comparar la vida útil media de dos marcas de pilas de 9 volts, se selecciona una muestra de 100 pilas de cada marca. La muestra de la primera marca tiene una vida útil media de 47 horas y una desviación estándar de 4 horas en tanto que la muestra de la segunda marca tiene una vida útil media de 48 horas y una desviación estándar de 3 horas. ¿ Es significativa la diferencia entre las dos medias muestrales al nivel de 0,05?

Ejercicio 5. Los siguientes datos son de dos tratamientos en parcelas al azar. Haga el análisis de varianza.

T1	T2
189	170
202	179
220	203
207	192
194	172
177	161
193	174
202	187
208	186
233	204

Ejercicio 6. Los siguientes datos son de dos muestras independientes.

	Muestra A	Muestra B
X	124	120
n	50	36
$\sum(X-X)^2$	5512	5184

Pruebe si la media de la muestra A es igual o mayor a la de la muestra B. Explique los pasos y el razonamiento.

Ejercicio 7. Tenemos los datos de dos variedades de un cultivo. Pruebe la diferencia en rendimiento entre las variedades, explicando cuidadosamente el razonamiento.

Variedad 1: 9 4 10 7 9 10
Variedad 2: 14 9 13 12 13 8 10

CONTESTE SI ES CIERTO O FALSO Y SI ES FALSO DIGA COMO CAMBIA LAS PALABRAS SUBRAYADAS PARA HACER VERDADERA LA FRASE

1. Muy a menudo la preocupación al estudiar la varianza es mantenerla bajo control, es decir relativamente chica. Por lo tanto, muchas de las hipótesis acerca de la varianza serán a una sola cola.
2. Cuando las medias de dos muestras no relacionadas se usan para comparar dos poblaciones estamos trabajando con dos medias relacionadas.
3. El uso de datos apareados (muestras dependientes) permite a menudo el control de variables no medibles o confundidas porque cada par esta sujeto a esos efectos confundidos igualmente.
4. La distribución chi-cuadrado se usa para hacer inferencia sobre el cociente de varianzas de dos poblaciones.
5. La distribución F se usa cuando se comparan dos medias dependientes.
6. Al comparar dos medias independientes cuando las varianzas son desconocidas necesitamos realizar una prueba F en sus varianzas para determinar la fórmula apropiada para usar.
7. La normal estandarizada se usa para toda inferencia concerniente a proporciones poblacionales.
8. La distribución F tiene media cero y es simétrica con respecto a la media.
9. El número de grados de libertad para el valor crítico de t es igual a el menor de n_1-1 y n_2-1 cuando se hace inferencia sobre la diferencia entre dos medias independientes para el caso que las varianzas sean desconocidas, pero se suponga que son iguales, y los tamaños de muestra sean pequeños.
10. Una estimación conjunta de cualquier estadística en un problema de dos poblaciones es un valor al que se llega combinando las estadísticas de dos muestras separadas para lograr el mejor estimador puntual posible.

CLASE 14
INFERENCIA SOBRE VARIANZAS

14.1.INFERENCIA SOBRE UNA VARIANZA

Al igual que con respecto a la media se pueden realizar inferencias sobre la varianza de una población. En esto tiene importancia fundamental la propiedad (Pg. 18): $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{X-\overline{X}}{\sigma}\right)^2+n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2$

Estimadores de la varianza. Como $E(S^2) = [(n-1)/n]\sigma^2$ la varianza muestral es un estimador sesgado de la poblacional por lo que se propone el nuevo estimador[3] que es insesgado:

$$\hat{\sigma}^2=\left(\frac{n}{n-1}\right)s^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$$

Intervalos de Confianza para s^2 . En la siguiente reformulación de la variable χ^2 se puede construír intervalos de confianza para σ^2 (Recordemos que $\sum x_i^2=\sum(X_i-\overline{X})^2$).

$$P\left[\chi_I^2<\chi^2<\chi_S^2\right]=P[\chi^2_{\alpha/2}<\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}<\chi^2_{1-\alpha/2}]=p\left[\frac{\sum(X-\mu)^2}{\chi^2_{(1-\alpha/2)}}<\sigma^2<\frac{\sum(X-\mu)^2}{\chi^2_{\alpha/2}}\right]=1-\alpha$$

Ejemplo 14.1. Supongamos que estamos estudiando los tratamientos cuyos datos se proporcionan abajo y se intenta comparar la varianza del tratamiento 1 con la del 2.

Tratamiento	1	2
	62	163
	86	208
	117	154
	125	154
	132	183
		212
		169
Totales	522	1243
Medias	104,40	177,57

Las sumas de cuadrados son: SC = SC sin corregir - C. Para el tratamiento 1 tenemos 57978 -522^2/5 = 3481.2 por lo tanto la estimacion de la varianza es: 3481,2/5 = 870,3. Similarmente para el tratamiento 2, SC = 224259 - 1243^2/7 = 3537,714; con lo que la estimación de la varianza es: 3537,714/7 = 589,62. Un intervalo de confianza para la varianza del tratamiento 1 es: P[3481,2/11,14 < σ^2 < 3481,2/0,484] = P[312,4955 < σ^2 < 7192,561] =0,95. Similarmente para el tratamiento 2: P[251,4366 < σ^2 < 2093,322] = 0,95

Pruebas de Hipótesis Ejemplo 1.15 (Cont.) Supongamos que queremos probar la hipótesis de que la varianza de la población 1 es 1000. Según los pasos de la sección 7.3 tenemos:

1. *Definir las hipótesis*.
2. *Elegir el nivel de significación*, supongamos α= 0,05
3. *Elegir la variable pivot*, en este caso será $\chi^2_{(4)}=\frac{\sum_{i=1}^{5}(X_i-\overline{X})^2}{\sigma^2}$
4. *Determinar la región crítica*: $\chi^2_c > \chi^2_t$
5. *Hacer los cálculos* $\chi^2_{(4)}=\frac{3481,20}{1000}=3,4812$
6. *Tomar la decisión*. χ^2_c no pertenece a la región crítica, por tanto no rechazamos H_0.

[3] Llamada a veces la cuasi-varianza.

14.2.PRUEBA DE HOMOGENEIDAD DE VARIANZAS.

El cociente de dos estimaciones de la varianza de una población tiene una distribución F. $F = \frac{\sigma^2}{\sigma^2}$. La distribución F tiene dos parámetros que son los grados de libertad del numerador y los del denominador.

Ejemplo 14.1 (Cont.) En el ejemplo si queremos comparar las varianzas de los dos tratamientos, tenemos: F = 870,3 / 589,62 = 1,476037; como el valor de tablas para la F con 4 grados de libertad en el numerador y 6 en el denominador es: F(4,6) = 4.53 se concluye que el cociente de varianzas no es significativo, es decir que a los efectos prácticos tomamos las varianzas como iguales.

Práctico 14. Inferencia sobre varianzas.

1. Una persona podría usar la varianza o la desviación estándar del cambio diario en el precio de las acciones del mercado como una medida de estabilidad. Nosotros deseamos comparar la estabilidad de las acciones de una compañía este año con la del año pasado. Le proporcionan los siguientes resultados de muestras tomadas al azar de las ganancias y pérdidas diarias de los últimos dos años. Este año: n=25; s=1,57. El año pasado: n=25; s=0,26

1.1. Construya un intervalo de confianza del 95% para la razón de la desviación estándar del año pasado respecto a la de este año.

1.2. ¿Es la ganancia diaria mas o menos estable este año comparada con el año pasado?

2. Para estimar la razón de las varianzas entre los pesos de latas de duraznos en almíbar de la marca líder con respecto a su propia marca, el gerente de un almacén tomó muestras obteniendo los siguientes resultados: Marca líder: n=16; σ^2 =1,968. Su marca: n=25; σ^2 =2,834

2.1. Proporcione un estimador puntual para el cociente de varianzas entre la "marca líder" con la varianza de la marca del caballero.

2.2. Construya un intervalo de confianza del 90% para la misma razón de varianzas.

2.3. Encuentre el intervalo de confianza del 90% para la el cociente de las desviaciones estándar.

CLASE 15
ANALISIS DE LA VARIANZA

La técnica del análisis de varianza es muy utilizada con varios propósitos. Acá la introducimos como una manera de probar la hipótesis de que varias medias son iguales, contra la alternativa de que no lo son. Para ello haremos referencia al siguiente ejemplo:

Ejemplo 15.1. Ejemplo de análisis de varianza

Tratamiento	1	2	3	4
	62	163	60	137
	86	208	62	137
	117	154	72	159
	125	154	75	132
	132	183	52	126
Medias	104.4	177.57	64.2	135.57

El análisis de varianza se basa en estimar la variación que las medias definen e, independientemente, estimar la variación "natural" de la población y comparar ambas. La variación natural de los datos se mide por la "intravarianza" o "variación del error". Si las muestras son de la misma población, las medias tienen (por azar) una varianza que es la enesima parte de la varianza de la población: $s^2_{\bar{Y}} = \frac{s^2_Y}{n}$, por lo que $n s^2_{\bar{Y}} = s^2_Y$ deberá ser igual a σ^2. Si se descarta esa situación es porque, aparte del azar, algo más diferencia a las muestras. En la práctica lo anterior implica que descomponemos la varianza existente (del total de los datos) en dos partes, le llamamos entre y dentro de muestras. La varianza dentro de muestras, al no tener causa aparente de variación, decimos que es causada por el azar o error experimental. La varianza entre muestras puede ser debida al azar o no. Si las diferencias entre los tratamientos son solamente debidas al azar las dos variaciones son del mismo orden y su cociente vale mas o menos 1. El cociente de varianzas tiene distribución F, por lo que se busca en tablas de la distribución F si las relaciones obtenidas son aceptables como cercanas a 1 o no. El resultado se expresa generalmente en un cuadro de análisis de varianza. Existen fórmulas que facilitan el cálculo en el análisis de varianza del siguiente modo:

$$SC = \sum_{i=1}^{n}\sum_{j=1}^{k}\left(Y_{ij} - \overline{Y}_{..}\right)^2 = \sum_{i=1}\sum\left(Y_{ij} - \overline{Y}_{.j}\right)^2 + \sum\sum\left(\overline{Y}_{.j} - \overline{Y}_{..}\right)^2 = SCE + SCT$$

Fuente Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F	$F_{crític}$
Tratamientos		3			
Error o residuo		16			
TOTAL		19			

Este procedimiento tiene la diferencia con la prueba t de Student en que es aplicable a mas de dos medias simultáneamente. Pero existe una relación entre ambos y si aplicamos el análisis de varianza a la comparación de dos medias, obtenemos resultados concordantes $t^2 = F$. Con estos comentarios, por supuesto, no agotamos todo lo que hay para decir acerca del análisis de varianza, una de las técnicas mas usadas e importantes en análisis de datos. En clases posteriores (10 y 14) desarrollaremos otras visiones y otras aplicaciones del análisis de varianza.

Práctico 15. Análisis de la varianza.

Ejercicio 1. Los siguientes datos experimentales son de tres grupos independientes.

Grupo	Datos	
1	4 16 49 64 81	
2	49 121 144 169 196	
3	16 36 81 100 121	

Haga el análisis de varianza. Explique el razonamiento y las conclusiones.

Ejercicio 2.
```
DATA CAP5;    INPUT TRT Y;   CARDS;
1 37
1 31
1 38
1 34
2 20
2 16
2 15
2 29
3 41
3 44
3 48
3 47
4 25
4 22
4 15
4 18
```

Ejercicio 3. Dadas las medias de 10 individuos en cada uno de 5 grupos que son 30, 42, 43, 63 y 38 y la varianza del error que es 12, realice el análisis de varianza y la separación de medias. Explique por que elige el método de separación de medias que utilizo.

Ejercicio 4. Los siguientes datos son de dos muestras independientes.

	Muestra A	Muestra B
X	124	120
N	50	36
$\sum(X-X)^2$	5512	5184

Pruebe si la media de la muestra A es igual o mayor a la de la muestra B. Explique los pasos y el razonamiento.

Ejercicio 5. Los siguientes datos experimentales son de tres grupos independientes.

Grupo	Datos	Medias
1	4 16 49 64 81	42.80
2	49 121 144 169 196	135.80
3	16 36 81 100 121	70.80

Haga el análisis de varianza. Explique el razonamiento y las conclusiones.

Ejercicio 6. Se muestran datos de conteo de malezas. Analicelos por medio de un análisis de varianza. Comente los supuestos y limitantes que tiene el método. Aca hay un problema ¿Cuál es?

A	B	C	D	E
28	7	6	177	184
22	11	9	151	146
54	30	26	110	131
19	6	7	117	110
32	11	7	135	134

Ejercicio 4. Los datos en la página siguiente son de tres grupos independientes.

Grupo	Datos	Medias
1		
2		
3		

Explique y haga el análisis del experimento. Explique los principales puntos del razonamiento y de las conclusiones, así como todo lo que aprendió de la interpretación de experimentos como esos.

Datos de la tesis de Invernizzi y Marziotte (1998) de peso de animales sometidos a tres tratamientos:

Trata	Animal	Ganancia de Peso
1	1	-0.52
1	2	0.00
1	3	0.15
1	4	-0.43
1	5	0.25
1	6	0.14
1	7	0.04
1	8	0.33
2	1	0.39
2	2	-0.09
2	3	0.88
2	4	0.46
2	5	0.08
2	6	0.53
2	7	0.65
2	8	0.38
3	1	0.42
3	2	0.38
3	3	0.09
3	4	0.17
3	5	0.43
3	6	0.45
3	8	-0.05

EJERCICIO 1. Un estudiante obtuvo los siguientes datos en su tesis. El experimento consistía en comparar variedades de frutilla. Las variables son:

bloque	tratamiento	pc	gp	pg	pmf;
1	1	12570	85	11942	17.8
2	1	9726	82	8920	16.1
3	1	7628	80	7183	16.8
4	1	7575	82	7222	17.3
1	2	22711	85	21719	18.4
2	2	25913	90	25060	19
3	2	25107	90	24345	19.8
4	2	24321	89	23472	18.8
1	4	33576	89	31912	15.7
2	4	36621	92	34926	15.9
3	4	38418	92	36332	16.2
4	4	31375	90	29363	15.3
1	5	30120	93	29629	20.6
2	5	27298	94	26882	19.7
3	5	27549	95	27112	20.1
4	5	24202	93	23796	19.1
1	6	34134	91	31613	15.9
2	6	37491	91	34724	16.2
3	6	39972	90	36760	15.6
4	6	34447	90	31385	15.5

1	7	36194	94	35039	18
2	7	34191	91	32429	17.5
3	7	34138	92	32540	16.8
4	7	33999	90	31722	17.7

No recordamos que es cada variable, pero le solicitamos que haga el análisis de varianza para pmf (creemos que es peso medio a la floración). Note que son 24 observaciones en bloques.

EJERCICIO 2. Calcule la media general, el coeficiente de variación, y el R2 si recuerda lo que es. Estime los efectos de cada tratamiento. Explique brevemente lo que hace.

EJERCICIO 3. Haga la separación de medias usando la prueba que considere mas adecuada. Explique porque la elijió, ventajas e inconvenientes que puede tener la estrategia que Ud. utilizó.

EJERCICIO 4. Los siguientes datos son de dos tratamientos en parcelas al azar. Haga el análisis de varianza.

T1	T2
189	170
202	179
220	203
207	192
194	172
177	161
193	174
202	187
208	186
233	204

CLASE 16
ESTUDIO DE PROPORCIONES

16.1. ESTUDIO DE PROPORCIONES. Se ha comentado en las clases iniciales que las variables pueden ser cuantitativas o cualitativas también llamadas atributos (o categorías o clases). Decíamos que las variables se miden y los atributos se cuentan. Por esta razón el análisis de atributos muchas veces se denomina "análisis de conteos". En un estudio de atributos (pero no necesariamente ahí) se necesita en realidad estudiar las proporciones (o porcentajes) de observaciones con una determinada característica. Por ejemplo, puede interesar estudiar el sexo de los estudiantes de la Universidad de Uruguay. Si nos interesa saber que porcentaje de los estudiantes son mujeres disponemos de dos caminos: un estudio exhaustivo (censo) o un muestreo. Para estimar por muestreo el porcentaje de estudiantes que son mujeres, se toma una muestra al azar de estudiantes y se cuenta cuantos son mujeres. Supongamos que se tomaron 11 personas al azar y hay 6 mujeres: el mejor estimador del porcentaje de mujeres es 6/11. De este modo, casi intuitivamente tenemos un estimador puntual de la proporción poblacional. Este estimador se obtuvo por el método de los momentos: el estimador del parámetro es la correspondiente estadística.

Habíamos comentado que un característica deseable en un estimador era conocer su distribución. También habíamos analizado la idea de aproximar la binomial por la normal. La variable no de casos tiene una distribución binomial que se puede aproximar por una normal.

Las proporciones se distribuyen normal con media en la proporción poblacional P y varianza PQ/n, donde Q=1-P

$$p \sim N\left(P; \frac{PQ}{n}\right)$$

(Nota: usaremos indistintamente P o p minúscula proporción estimada o sea una proporción muestral. También se puede usar la expresión x/n donde x es el numero de casos con una determinada característica.)

16.2. INTERVALOS DE CONFIANZA PARA PROPORCIONES. Para los intervalos de confianza se presentan dos posibilidades: un intervalo aproximado usando p en lugar de P

$$p \pm z\sqrt{\frac{PQ}{n}}$$

Supongamos que queremos hacer un intervalo de 95% de confianza para la proporción de mujeres entre los estudiantes de la Universidad de la República con los datos manejados anteriormente:

$\hat{p} = \hat{P} = 6/11 = 0{,}545$ por lo tanto $\hat{Q} = 1 - 0{,}545 = 0{,}455$

por lo tanto: $0{,}545 \pm 1{,}96 * \sqrt{\frac{(0{,}545)(0{,}455)}{11}}$, o sea $0{,}545 \pm 1{,}96 * 0{,}150$ lo que da: (0,251; 0,839), es decir que la verdadera proporción de mujeres entre los estudiantes de la Universidad de la república esta entre 0,251 y 0,839 con un 95% de confianza.

Si observamos que PQ en la formula anterior depende de la P desconocida, vemos que el intervalo de confianza es aproximado. Un procedimiento para construir intervalos exactos seria más complejo.

16.3. PRUEBA DE HIPOTESIS SOBRE PROPORCIONES. Supongamos que interesa probar la hipótesis de que la proporción de mujeres en la Universidad es del 50%. El procedimiento será el siguiente:

H_0: P = 0,50

H_A: P ≠ 0,50

Elegimos un α = 0,05 por ejemplo

Elegimos la variable: $Z = \dfrac{P - P}{\sqrt{\dfrac{PQ}{n}}}$

Determinamos la región crítica: para el α=0,05 es |z|=1,96. O sea que se rechazaran los valores de z que sean mayores que 1,96 o menores que -1,96

Cálculo de valores $Z = \dfrac{0,545 - 0,500}{\sqrt{\dfrac{(0,500)(0,500)}{11}}} = \dfrac{0,045}{0,151} = 0,298$

como 0,298 no pertenece a la región crítica no se rechaza la hipótesis nula. Bien puede ser cierto que la mitad de los estudiantes de la Universidad sean mujeres de acuerdo a la información que nos proporciona la muestra.

Notemos que la discrepancia entre los valores observados y los postulados es mínima: es decir que en un total de 11 personas 6 sean mujeres es lo más cerca de la mitad que se puede pedir. Por lo tanto es fácil ver aun sin ir a tablas que la hipótesis no se rechazaría.

16.4. TAMAÑO DE MUESTRA PARA ESTUDIO DE PROPORCIONES. Una objeción que se puede hacer a la situación anterior es que la muestra puede ser muy chica para detectar una discrepancia con la hipótesis. Otra manera de decirlo es que el intervalo de confianza es muy amplio es decir poco preciso. Entonces se nos puede decir que interesa calcular el tamaño de muestra necesario para estimar la proporción con un margen de error por ejemplo del 0,10:

$Z\sqrt{\dfrac{PQ}{n}} = 0,10$ por lo tanto $n = z^2PQ/d^2$ lo que es una manera adaptada de la forma dada en la sección 11.8. Notemos que necesitamos conocer P para calcular el tamaño de muestra, lo que no tenemos. El razonamiento mas comúnmente seguido es que el máximo de PQ se da cuando P=Q=0,5 y PQ=0,25 por lo que en el peor de los casos $n = z^2/4d^2$. En el presente caso, con α=0,95 z=1,96 y d=0,10: $n = 1,96^2/4*0,10$.

Por lo tanto el tamaño mínimo que cumple con ese requisito es 97 observaciones.

16.5. ESTUDIO DE PROPORCIONES CON MUESTRAS CHICAS. Se puede ver que el numerador y el denominador no son independientes por lo tanto no tiene sentido plantearse una prueba t para proporciones. Eso, a su vez, implica que no se puede hacer un estudio de proporciones para muestras chicas con este procedimiento. Recordemos que muchos autores consideran muestras chicas a las que son mayores a 30 observaciones, mientras que otros llevan a 100 el límite para considerar grande a la muestra.

16.6. DIFERENCIA DE PROPORCIONES. Si las proporciones se distribuyen normal la diferencia también lo hace, es decir: La diferencia de proporciones se distribuye normal con media en la diferencia de proporciones poblacional y con varianza que es la suma de las varianzas:

$$(p_1 - p_2) \sim N(P_1 - P_2 ; \sigma^2_{P_1 - P_2}) \quad \text{donde} \quad \sigma^2_{P_1 - P_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

16.7. PRUEBA DE HIPOTESIS SOBRE DIFERENCIA DE PROPORCIONES. Supongamos que tenemos la siguiente situación de prueba de hipótesis: en dos facultades se obtuvieron apoyo para una determinada iniciativa en las siguientes proporciones:

Facultad A 14 apoyaron en 25 encuestados
Facultad B 3 30

y queremos saber si la diferencia entre facultades es significativa al 0,05.

1. Definir las hipótesis H_0: $P_1 - P_2 = 0$
   H_A: $P_1 - P_2 \neq 0$
2. Elegir el nivel de significación, por ejemplo tomamos α=0,05
3. Elegir la variable, $$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sigma_{P1-P2}}$$

Definir el valor crítico y la región crítica para el nivel de significación 0,05 el valor es 1,96; por lo tanto la región crítica es el conjunto de valores mayores que 1,96 y menores de -1,96

5. Hacer los cálculos $$Z = \frac{(\frac{14}{25} - \frac{3}{30}) - 0}{\sqrt{\left(\frac{14+3}{25+30}\right)\left(1 - \frac{14+3}{25+30}\right)\left(\frac{1}{25} + \frac{1}{30}\right)}}$$
6. Tomar la decisión

16.8. INTERVALOS DE CONFIANZA PARA DIFERENCIA DE PROPORCIONES. La construcción de un intervalo de confianza para la diferencia de proporciones se hace como siempre, por ejemplo para el 95% de confianza:

$$p_1 - p_2 \pm 1{,}96 * \sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}, \text{ en el ejemplo } \left(\frac{14}{25} + \frac{11}{30}\right) \pm 1{,}96 * \sqrt{\frac{14 * 11}{25} + \frac{3 * 27}{30}}$$

16.9. TAMAÑO DE MUESTRA Para determinar el tamaño de muestra se sigue una metodología igual a la mostrada antes.

Ejemplo. Un fabricante desea estimar la diferencia en la defectuosa entre dos procesos de producción de fusibles una probabilidad de 0,95. Cuantos fusibles debe elegir de cada proceso? Usemos una precisión=0,06

$$1{,}96 = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

suponiendo igual tamaño de muestra n_1=n_2=n, suponiendo P=Q=0,50

$$1{,}96 = \frac{\frac{X}{n} - \frac{X}{n} - 0}{\sqrt{0{,}25\frac{2}{n}}} 3.$$

En la tesis de Joaquin Azanza se estudiaron vacas a la sombra y al sol en su efecto sobre la preñez. Se obtuvieron los siguientes datos:

	Preñada	Vacía
Sombra	8	6
Sol	5	10

Se desea saber si el sombreado afecta o no el porcentaje de preñez

Práctico 16. Proporciones.

1. Un vendedor de equipos de televisión afirma que por lo menos el 75% de los reclamos por televisores de color se deben al mal funcionamiento de un tubo en particular. Una muestra aleatoria de 150 reclamos mostraron que 102 se debieron a este tipo de tubo. Presenta esa muestra evidencia suficiente para rechazar la afirmación del vendedor? Use un α=0,01.
2. Una máquina falló 16 veces en los 400 primeros intentos. Construya un intervalo de confianza del 98% para la verdadera proporción de las veces que la máquina fallar.
3. Una muestra de 250 walkies-talkies contiene 31 que no funciona propiamente. Construya un intervalo de confianza del 98% para la verdadera proporción que no funciona propiamente.
4. Una propaganda usada para los neumáticos radiales Sears dice que aumentan el kilometraje en un 7.4% sobre los neumáticos más usados. ¿Qué razón puede existir para utilizar el porcentaje y no el promedio en la propaganda?

CONTESTE SI ES CIERTO O FALSO Y SI ES FALSO DIGA COMO CAMBIA LAS PALABRAS SUBRAYADAS PARA HACER VERDADERA LA FRASE

5. La distribución chi-cuadrado es asimétrica y su media es siempre 2.
6. $\sqrt{npq}$ es el error estándar de una proporción.
7. La distribución de p en el muestreo es aproximadamente chi- cuadrado.

Comparación de proporciones.

8. Un vendedor de un nuevo fabricante de "walkie-takies" dice que el porcentaje de artefactos fallados en su producto es menor que el porcentaje de productos fallados de un competidor. Para probar eso se tomaron muestras al azar de cada fabricante. La muestra se resume abajo:

	Muestra	Defectuosos	No. estudiado
Vendedor	1	8	100
Competidor	2	2	100

Se puede rechazar la afirmación del vendedor con un nivel de significación del 5%?
9. Un político estudia su campaña y desea estimar la diferencia que ejerce entre los votantes masculinos y femeninos. Por lo tanto, solicita a su equipo de asesores que tomen dos muestras y encuentren un intervalo del 99% de confianza de la diferencia. Se toma una muestra de 1000 ciudadanos de cada sexo, y se encuentra que 388 hombres y 459 mujeres favorecen al señor. Realice el intervalo necesario.
10. Dos fabricantes que producen artefactos equivalentes dicen tener la misma proporción de fallas en sus productos. Una muestra aleatoria de cada uno muestra 14 de 300 y 25 de 400 defectuosos para cada uno (llamémosle A y B). Es eso evidencia para indicar diferencia en la proporción de artículos defectuosos a un nivel de significación de 0,05?
11. Se espera que un estudiante adivine en un examen de múltiple opción cuando no sabe la respuesta. Una pregunta particularmente difícil tiene una respuesta correcta entre 5 alternativas posibles. De los 50 estudiantes con las puntuaciones mas altas 15 contestaron esta pregunta correctamente, mientras que solo 7 de los 50 que sacaron las notas mas bajas lo hicieron. ¿Será verdad que los mejores contestaron significativamente (al 10%) mejor esta pregunta que los peores?
12. En un muestreo tomado para entender la personalidad, 22 de 71 personas de menos de 18 años expresaron un temor a conocer gente, mientras que 23 de 91 personas de 18 o mas años expresaron lo mismo. ¿Se puede rechazar la hipótesis nula de que no hay diferencia entre las dos proporciones verdaderas al α=0,02?
13. En una encuesta a 300 personas en una ciudad cerca de Salto, se encontró que 128 personas preferían Nueva Primavera sobre otras marcas de desodorante. En la ciudad B, se encontró 149 de 400 personas preferían Nueva Primavera. Encuentre el intervalo de confianza del 98% para la diferencia entre las dos proporciones.
14. Una prueba tomada a 100 jóvenes y 200 adultos mostró que 50 de los jóvenes y 60 de los adultos fueron conductores descuidados. Use los datos para estimar cuanto mayor es el porcentaje de jóvenes descuidados que el porcentaje de adultos descuidados para manejar al 90% de confianza.

CLASE 17
CUADROS DE CONTINGENCIA

17.1. DISTRIBUCION CHI-CUADRADO La distribución chi-cuadrado (o ji- cuadrado como me dicen que se escribe en español) proporciona otro modo de estudiar la diferencia entre proporciones.

Un grupo de 300 estudiantes de ambos sexos fueron consultados si preferían matemáticas, ciencias sociales o humanísticas. La tabla siguiente presenta los resultados:

Sexo	Matemáticas	C. Sociales	Humanidades	Total
Mujeres	35	72	71	178
Varones	37	41	44	122
Total	72	113	115	300

El enfoque consiste en calcular un valor esperado para cada celda del siguente modo: la probabilidad de que un encuestado sea mujer esta dada por 178/300 la probabilidad de que a una persona tenga preferencia por las matemáticas es 72/300. Estas probabilidades se conocen como marginales, ya que se calculan a partir de los márgenes de la tabla. Por lo tanto la probabilidad de que una persona al azar sea mujer y le guste las matemáticas es:

P[mujer y guste matemáticas] = P[mujer].P[guste matemáticas]

$$= \left(\frac{178}{300}\right).\left(\frac{72}{300}\right) =$$

Por lo tanto el numero esperado de mujeres es la probabilidad de que sea mujer por el numero de personas encuestadas (el tamaño de la muestra):
Número esperado de mujeres que gustan matemáticas (178*72)/300
Similarmente, numero esperado de hombres que gustan matemáticas =122*72/300
Si la diferencia entre lo observado y lo esperado es grande el supuesto de independencia no se cumple por lo que los dos sexos tienen diferente porcentaje de preferencia por las materias. La división entre los valores esperados se hace a efectos de relativizar los desvíos respecto al número de casos observados.

Luego se aplica el siguiente criterio (estadística):

$$\chi^2 = \frac{(o - e)^2}{e}$$

La distribución de $\sum(o-e)^2/e$ se considera aproximadamente χ^2
si todos los esperados son mayores de 5. Los grados de libertad de χ^2 son (f-1)(c-1) es decir el numero de filas menos 1 por el numero de columnas menos 1.

17.2. CUADROS DE CONTINGENCIA

Consideremos la siguiente situación, conocida como cuadro de contingencia:

Tamaño	Ganaderos-Lecheros		Agro-Ganaderos-Lecheros	
Predio	Baja	Alta	Baja	Alta
Chicos	16	16	89	193
Medianos	13	17	76	131
Grandes	-	-	-	31

La filosofía básica consiste en calcular los valores esperados si los dos criterios fueran independientes. En este caso los valores esperados son:

Tamaño	Ganaderos-Lecheros		Agro-Ganaderos-Lecheros	
Predio	Baja	Alta	Baja	Alta
Chicos	29*314/355	33*314/355	165*314/355	193*314/355
Medianos	29*237/355	33*237/355	165*237/355	131*237/355
Grandes	29*31/355	33*31/355	165*31/355	31*31/355
	29	33	165	355

Fórmula de cálculo abreviado. Existe una fórmula de cálculo abreviado presentada aquí para que se trabaje menos:

$$\chi^2 = \frac{\Sigma (o-e)^2}{e} = \frac{\Sigma o^2}{e} + \frac{\Sigma e^2 - 2\Sigma oe}{e} = \frac{\Sigma o^2}{e} + \frac{\Sigma e^2}{e} - 2 = \frac{\Sigma o^2}{e} - n$$

el último paso es debido a que se simplifican los valores del cuadrado en el numerador con el denominador y luego la suma de los esperados es igual a la suma de los observados con lo que se simplifica n - 2n. Esta fórmula de cálculo abreviado implica que no es necesario calcular los desvíos, lo que es un ahorro de tiempo considerable. Pero como suceda en el caso de los desvíos para la varianza se pierde la capacidad de detectar si la suma es efectivamente cero, como debe ser.

17.3. ESTUDIO DE LA DIFERENCIA ENTRE DOS PROPORCIONES

Se puede ver que la metodología presentada se puede aplicar a la prueba de hipótesis de diferencia entre dos o más de dos proporciones.

Equivalencia entre χ^2 y la prueba z. Ya habíamos dicho que la relación entre ambas era muy estrecha: una normal es una χ^2 con un solo grado de libertad o una χ^2 es una suma de normales elevadas al cuadrado. Por lo tanto no debe sorprender que ambas pruebas sean equivalentes en la presente situación.

Una observación adicional es que el método de χ^2 no proporciona lo que se conoce como capacidad de separar medias: es decir que si detectamos diferencias entre materias eso implica "no todas las materias tienen igual grado de preferencia" pero no sabemos si son todas diferente de todas o alguna en particular difiere de las demás. Supongamos que la preferencia expresada por una muestra de los estudiantes de 6o. año de secundaria acerca de la Facultad en la que estudian es así:

Derecho	Ingeniería	Agronomía
79	8%	13

si demostramos que "no todas las facultades tienen igual preferencia" no sabemos si la diferencia entre Ingeniería y Agronomía ellas dos.

Practico 17. Chi-Cuadrado.

1. La Regional Norte de la Universidad afirma que sus notas se distribuyen del siguiente modo: 10% saca 10-12, 20% saca 9-10, 40% saca 6-8, 20% saca 3-5 y 10% pierde los exámenes. En una encuesta de 200 estudiantes tomados al azar entre los que habían dado examen. Se encontró que 16 estuvieron en la primera dase, 43 en la segunda, 61 en la tercera, 48 en la cuarta y 32 perdieron. ¿Contradice este resultado la afirmación al nivel de significación del 0,05?

2. Una muestra al azar de 100 empleados en el registro de asistencia del año pasado mostró el siguiente grado de ausentismo en cada una de las siguientes categorías:

	Hombres		Mujeres	
	Casados	Solteros	Casadas	Solteras
Número de empleados	40	14	16	30
Días de ausencia	180	110	75	135

¿Proveen esos datos de evidencia suficiente para rechazar la hipótesis de que el grado de ausentismo es el mismo para todos los estados civiles? Use α=0,01 y un año de trabajo de 240 días.

3. Una muestra aleatoria de 60 adultos que trabajan para una empresa local fueron interrogados acerca del tiempo que pasaron mirando televisión la semana pasada.

	Varones	Mujeres
Mas de 15 horas	10	18
Menos de 15 horas	19	13

¿Muestra esta información al 0,05 de significación suficiente evidencia para rechazar la afirmación que el sexo y el tiempo que se mira televisión son independientes?

4. Se tomó una muestra aleatoria de 500 hombres casados de todo el país y cada persona se clasificó de acuerdo al tamaño de la comunidad en que reside y del tamaño de la comunidad en que se crió. Los resultados:

	Tamaño de la comunidad de residencia			
Tamaño de comunidad en que se crió	Menos de 10.000	10.000 a 49.999	50.000 o más	Total
Menos de 10.000	42	45	25	112
10.000 a 49.999	18	64	61	143
50.000 o mas	4	54	187	245
Total	64	163	273	500

Contradice esa información la idea de independencia al 0,01 nivel de significación.

5. Se probaron 100 granos de cada una de cuatro marcas de maíz para palomitas por su facilidad de "reventado". El número de granos que no explotaron se muestra abajo.

Marca	A	B	C	D
No. que no explotó	12	5	11	19

¿Se puede rechazar la hipótesis de que las 4 marcas explotan igualmente al α=0,05?

6. En un estudio de polución ambiental se tomo una muestra aleatoria de 100 hogares y se les preguntó si algún miembro de la familia estaba preocupado por la contaminación del aire. Un resumen de las respuestas se da en la siguiente tabla:

	¿Hay algún miembro de la familia preocupado por la contaminación del aire?	
Comunidad	Si	No
I	63	37
II	81	19
III	48	52

IV 54 46

¿Se puede concluir que las comunidades difieren en su captación del problema de la contaminación del aire al nivel de significación del 0,05?

7. En el año 83 se observaron los siguientes datos en un ensayo de cruzamientos (total de animales con terneros sobre el total de animales del ensayo):

Hereford 14/25

Cruza 3/30

¿Son iguales las proporciones? Use un α=0,05

Ejercicio 3. Los siguientes datos son de un estudio sobre enfermedades. Analice los datos, explicando todo lo que hace, en que supuestos se basa, si toma alguna decisión, etc. Puede haber mas de una forma de analizarlos.

		Numero de cigarrillos que fuma por día					
		1	5	15	25	50	Total
Hombres	Enfermos	33	250	196	136	32	
	Sanos	55	293	190	71	13	
Mujeres	Enfermos	7	19	9	6	0	
	Sanos	12	10	6	0	0	

Ejercicio 3. Los siguientes datos son de cuatro tipos de vacas y partos normales.

Tipos de vacas	Partos normales	Abortos	Totales
Charolais	447	68	
Indubrasil	492	14	
1/2 Charolais-Cebú	193	12	
3/4 Cebú-Charolais	254	10	

Estudie la situación explicando todo lo que considere conveniente.

CONTESTE SI ES CIERTO O FALSO Y SI ES FALSO DIGA COMO CAMBIA LAS PALABRAS SUBRAYADAS PARA HACER VERDADERA LA FRASE

1. El número de grados de libertad para la prueba de un experimento multinomial es igual al numero de celdas en el cuadro de contingencia.
2. La frecuencia esperada en una prueba de chi-cuadrado se encuentra multiplicando la probabilidad hipotética de la celda por el numero de datos en la muestra.
3. La frecuencia observada de una celda no se permite que sea menor a 5 cuando se hace una prueba chi-cuadrado.
4. En un experimento multinomial tenemos (f-1) por (c-1) grados de libertad (f es el número de filas y c el numero de columnas del cuadro de contingencia).
5. Un experimento multinomial consiste de n pruebas idénticas e independientes.
6. Un experimento multinomial arregla los datos en una tabla de doble entrada tal que los totales en una dirección son predeterminados.
7. Los datos para los experimentos multinomiales y las tablas de contigencia son distribuidos de tal modo que caen necesariamente en una categoría.
8. La estadística $\Sigma(o - e)^2 / e$ tiene una distribución aproximadamente normal.
9. Los datos usados en una prueba multinomial por X son siempre enumerativos en su naturaleza.
10. La hipótesis nula que se prueba en un test de homogeneidad es que la distribución de proporciones es la misma para cada una de las subpoblaciones.
11. La prueba de χ^2 de este tipo pueden ser a una o dos colas

12. La distribución chi-cuadrado es asimétrica y su media es siempre 2.

CLASE 18
BONDAD DE AJUSTE A UN MODELO

18.1. INTRODUCCION

La distribución χ^2 tiene un uso muy difundido en los llamados problemas de bondad de ajuste a modelos. Los modelos mas comunes, pero no los únicos, son de ajuste a binomial, normal y Poisson, pero puede haber otros casos.

Notemos que (o-e) mide la discrepancia entre lo observado y lo postulado. Lo más cercano que son lo menor que resulta el χ^2 . El numerador es solo para relativizar el resultado, es decir que si dos estudios tienen distinto número de observaciones no influya ...Por lo tanto las pruebas de χ^2 de este tipo son siempre a una sola cola, ya que los valores pequeños de χ^2 indican buen ajuste a la teoría.

Estas hipótesis ("un dado es correcto", los factores son independientes", etc.) no afirman cosas acerca de un parámetro.

18.2. BINOMIAL

Ejemplo 18.1) Ajuste los siguientes datos a una binomial:

Número de Hijos	Número de familias
0	8
1	16
2	38
3	22
4	10
5	6

La tarea consiste en calcular los valores que se esperarían si la distribución fuera binomial exacta y compararlos con los observados. Para calcular los esperados

18.3. NORMAL

Ejemplo 18..2. Ajústense los siguientes datos a una distribución normal.

103 133 111 184 127 124 117 102 124 115 153 122 105 104 115 140
115 113 117 125 135 127 125 121 84 87 108 85 101 117 90 144
106 111 97 70 113 113 110 64 94 100 55 90 93 107 93 89
126 119 82 98 57 100 134 111 113 93

Una de las características de este tipo de problemas es la laboriosidad, el trabajo que dan y el tiempo que consumen.

LIMITES DE CLASE	MARCA DE CLASE	OBSERVADO	DIFERENCIA ESPERADO	O-E	2 (O-E) /E
50-60	5	2	0.7250	1.2750	2.2422
60-70	6	1	1.4036	-0.4036	0.1161
70-80	7	1	3.0972	-2.0972	1.4201
80-90	8	5	5.6028	-0.6028	0.0649
90-100	9	8	8.3114	-0.3114	0.0117
100-110	10	10	10.0920	-0.0920	0.0008
110-120	11	15	10.0456	4.9544	2.4435
120-130	12	9	8.2012	0.7988	0.0778
130-140	13	3	5.4810	-2.4810	1.1230
140-150	14	2	3.0044	-1.0044	0.3358
150-160	15	1	1.3456	-0.3456	0.0888
160-170	16	0	0.4930	-0.4930	0.4930
170-180	17	0	0.1508	-0.1508	0.1508
180-190	18	1	0.0464	0.9536	19.5981
TOTALES		58	58	0.0000	28.1665 = X

Media: 109.82 Desviación Estándar: 22.22025

Gran parte del trabajo en estudios de este tipo esta en el calculo de los valores esperados.

Límites de Clase	Valor Z	Area Corresp	Probabilidad de la clase
50	-2.69	0.4964	0.0125
60	-2.24	0.4875	0.0242
70	-1.79	0.4633	0.0534
80	-1.34	0.4099	0.0966
90	-0.89	0.3133	0.1433
100	-0.44	0.1700	0.1740
110	0.01	0.0040	0.1732
120	0.46	0.1772	0.1414
130	0.91	0.3186	0.0945
140	1.36	0.4131	0.0518
150	1.81	0.4649	0.0232
160	2.26	0.4881	0.0085
170	2.71	0.4966	0.0026
180	3.16	0.4992	0.0008
190	3.61	0.4998	
	1	58	

Practico 18. Ajuste a modelos.

1. En un cruzamiento de variedades de poroto se espera de acuerdo a la teoría genética que la mitad de las semillas producidas sean rugosas y la mitad lisa. Se tomó una muestra al azar de 40 semillas que consistía en 30 rugosas y 10 lisas. Pruebe la hipótesis mencionada con 10% de nivel de significación por medio de un prueba z.
2. El resultado de un cruzamiento de dos tipos de plantas de maíz da como resultado 3 genotipos diferentes A, B y C. Un modelo genético sugiere que la proporción de los tres genotipos es 1:2:1. Para la verificación experimental se tomaron 90 plantas resultado del cruzamiento anterior y se observó la frecuencia de los tres genotipos.

Genotipos	No de plantas
A	18
B	44
C	28

¿Permiten estos datos corroborar el modelo genético?

LA ESTADÍSTICA EN LA MEDICINA. Los ensayos clínicos utilizan la estadística para determinar los mejores procedimientos médicos.

Los médicos profesionales a menudo bromean sobre el hecho de que cuando una persona está enferma, uno de los lugares más peligrosos para estar es en un hospital. La broma se refiere a que, en ocasiones, los pacientes de un hospital se contagian de otro paciente. Una investigación ha demostrado que la mayoría de los contagios de enfermedades se da mediante las manos de los trabajadores de la salud, cuando atienden a un paciente y luego a otro. Así pues, un grupo de médicos y enfermeras del Colegio de Medicina de la Universidad de Iowa decidieron investigar la práctica del lavado de manos de los trabajadores de la salud en las tres unidades de cuidado intensivo (UCI), en un período de ocho meses.

El doctor Bradley Doebbling y sus colegas establecieron un estudio cruzado para comparar dos tipos de limpiadores. Cada mes cambiaban el tipo de limpiador disponible. Ambos limpiadores tienen algún agente antibacteriano: uno contiene el antibiótico gluconato clorhexidrina y el otro contiene una solución de 60% de alcohol.

El hospital tiene ya establecidos procedimientos para la detección de infecciones en los pacientes de las UCI y su rápido tratamiento. Los investigadores también han establecido períodos de observación aleatoria cada media hora para registrar el número de veces que los médicos y las enfermeras de las UCI se lavan las manos entre revisiones. Después compararon el número de infecciones reportadas bajo los dos sistemas de lavado de manos, utilizando el número de días paciente bajo los dos regímenes (un paciente que esté en el hospital durante un día cuenta como un "día paciente").

Demostración de la existencia de un a diferencia entre limpiadores. El resultado que se obtuvo fue que durante los meses en que se utilizó la clorhexidrina en el hospital, las infecciones fueron 27% menos frecuentes que cuando se utilizó la solución de alcohol, y un análisis estadístico de la tasa de infecciones de los dos grupos indicó que esta diferencia es significativa.

La aparente efectividad superior de la clorhexidrina depende, en parte, de la disposición de los médicos y de las enfermeras en seguir las indicaciones sobre el lavado de manos. En general, los investigadores encontraron que los miembros del personal clínico se lavaban las manos en sólo aproximadamente 40% de las ocasiones, en los casos en que podría haber ayudado al control de infecciones, pero estaban más dispuestos a lavarse cuando tenían a disposición la clorhexidrina.

En una de las tres unidades de cuidado intensivo, la diferencia en la disposición a lavarse las manos fue de 48% de preferencia de la clorhexidrina contra 30% por el uso de la solución de alcohol ($P=0.002$ con una prueba t).

El resultado es sustancial. Los estudios han estimado el costo anual total del tratamiento de infecciones en hospitales, en Estados Unidos, entre cinco y diez mil millones de dólares. El animar al personal médico de un hospital a lavarse las manos entre una revisión y otra parece que disminuye en mucho las tasas de infecciones, y es más probable que médicos y enfermeras se laven las manos cuando tienen al alcance un limpiador que contenga un fuerte agente antibacteriano.

El juicio clínico. Los métodos estadísticos son empleados a menudo en la investigación del origen, tratamiento y control de diversa enfermedades. Debido a que gran parte de la investigación médica no se ajusta a la distribución normal, los métodos no paramétricos son particularmente útiles. Los doctores Charles H. Kirpatrick y David W. Alling, aplicaron la prueba de Mann-Whitney de una manera inteligente para evaluar los resultados de un juicio clínico aleatorio que involucraba el tratamiento de candidiasis oral crónica, una enfermedad caracterizada por infecciones recurrentes de la piel, uñas y membranas mucosas. Los resultados de sus pruebas indicaron que el clotrimazol, que había sido empleado con éxito en desórdenes similares, era un tratamiento altamente efectivo para dichas candidiasis.

Veinte pacientes que sufrían de candidiasis oral persistente fueron admitidos al estudio y se les asignaron mediante distribución aleatoria tratamientos con pastillas de clotrimazol o placebo. Las respuestas de los sujetos al tratamiento fueron evaluadas de dos a siete días después del tratamiento, como se muestra en la tabla RW14-1. Este formato captura dos tipos de resultados y los combina de forma tal que el mayor de cualesquier dos clasificaciones connota el resultado menos favorable: estas clasificaciones definen una clasificación ordenada. Los resultados de los tratamientos con clotrimazol y placebo se resumen en la tabla RW14-2. Los diez pacientes de las pastillas de clotrimazol no presentaron síntomas hacia el quinto día de tratamiento. Esta observación visual fue confirmada por una prueba de Mann-Whitey, que ofreció un fuerte respaldo estadístico.

La línea de fondo. Los tratamientos exitosos para enfermedades se encuentran sólo por medio de la investigación. En este caso, aunque se sabe que el clotrimazol ocasiona efectos colaterales adversos cuando se administra durante un período prolongado, los estudios preliminares que emplean clotrimazol oral en una programación intermitente han mostrado beneficios clínicos. El uso de métodos estadísticos permite de tratamientos médicos, lo que les presta credibilidad a sus hallazgos.

Tabla RW14-1

Sistema de clasif. para resultados de trat. de candidiasis crónica.

Clasificación	Hallazgos clínicos	Hallazgos de laborat.
1	Ausente	Negativo
2	Mejorado	Negativo
3	Mejorado	Positivo
4	No mejorado	Positivo

Tabla RW14-2

Resultados después de dos a siete días de trat. en 20 pacientes.

	Clasific. de resultados				
Grupos de tratamiento	1	2	3	4	Total de pacientes
Clotrimaz.	6	3	1	0	10
Placebo	1	0	0	9	10