

UNIVERSIDAD DE PUERTO RICO
RECINTO DE RIO PIEDRAS
FACULTAD DE ADMINISTRACION DE EMPRESAS
Instituto de Estadística y Sistemas Computadorizados de Información



MANUAL DE LA ACADEMIA
Aplicación de Conceptos y Herramientas
Esenciales de Estadística
Marzo - 2009

Preparado por:
José Carlos Vega Vilca, Ph.D.
josevega02@yahoo.com



Instituto
de Estadísticas
de Puerto Rico
Estado Libre Asociado de Puerto Rico

PRESENTACIÓN

El Instituto de Estadística y Sistemas Computadorizados de Información de la Facultad de Administración de Empresas de la Universidad de Puerto Rico, Recinto de Río Piedras y el Instituto de Estadísticas de Puerto Rico, conscientes de la necesidad de capacitar a empleados de las diferentes agencias gubernamentales encargados de producir las estadísticas de Puerto Rico, en común esfuerzo hacen realidad esta capacitación mediante la “*Primera Academia, Aplicación de Conceptos y Herramientas Esenciales de Estadística*”. Este esfuerzo beneficia inicialmente a 25 empleados con miras de incrementar ese número mediante la realización de sucesivas academias y con el objetivo de mejorar la calidad del trabajo del empleado gubernamental.

El presente manual fue preparado con dos propósitos principales. El primer propósito es complementar las lecciones del profesor en clase, con la finalidad de agilizar y mejorar el proceso de enseñanza aprendizaje. El segundo propósito es proporcionar un documento de consulta inmediata, que facilite el poder incrementar las destrezas de aplicación de las técnicas estadísticas más usadas en la administración pública. En este manual se usó un conjunto de base de datos para ilustrar las aplicaciones de cada técnica estadística. Estas fueron procesadas con la herramienta más común al servidor público, Microsoft EXCEL.

CONTENIDO

página

1.- <u>Conceptos Básicos en Estadística:</u> Población, unidad elemental, muestra, variable, observación, parámetro, valor estadístico, definición de Estadística.	1
2.- <u>Organización de datos:</u> Tabla de frecuencias y gráficos de datos cuantitativos y categóricos	4
3.- <u>Medidas de Tendencia Central:</u> Media, mediana y moda. Medidas de posición: cuartiles, percentiles	14
4.- <u>Medidas de variabilidad:</u> Rango o amplitud, varianza, desviación estándar, coeficiente de variabilidad, desviación intercuartílica.	18
5.- <u>Probabilidades:</u> Experimento aleatorio, espacio muestral, evento, esquema de probabilidad. Definición clásica de probabilidad, definición frecuentista de probabilidad.	21
6.- <u>Variable aleatoria:</u> Definición. Distribución de variables aleatorias discretas: Binomial y Poisson. Distribución de variables aleatorias continuas: Normal, Normal estándar, T de Student, Ji-cuadrado, F de Snedecor.	23
7.- <u>Estadística Inferencial:</u> Estimación de parámetros: intervalo de confianza para una media y una proporción. Prueba de hipótesis: una media, una proporción, diferencia de medias dependientes e independientes, homogeneidad de varianzas. Prueba de hipótesis en tablas de contingencia.	35
8.- <u>Análisis de Regresión y Correlación:</u> Regresión lineal simple y múltiple: análisis de varianza, coeficiente de determinación. Coeficiente de correlación lineal, prueba de hipótesis.	45
9.- <u>Muestreo:</u> Tipos de muestreo probabilístico: Muestreo Aleatorio Simple, Muestreo Aleatorio Sistemático, Muestreo Aleatorio Estratificado y Muestreo Aleatorio por Conglomerados.	50
Bibliografía	61



CONCEPTOS BASICOS

Población: conjunto de unidades elementales con características similares. La población es elegida de acuerdo al interés de la investigación.

- Conjunto de familias que viven en el barrio Santa Fe de la ciudad Ponce
- Conjunto de hoteles del Municipio de San Juan
- Conjunto de artículos producidos por una máquina.

Unidad elemental: es un elemento de la población sujeto a estudio.

- Un familia que viven en el barrio Santa Fe de la ciudad Ponce
- Un hotel del Municipio de San Juan
- Un artículo producido por una máquina

Muestra: subconjunto de unidades elementales de una población

Variable: es la característica de interés

- Var1: peso del artículo
- Var2: nivel de instrucción del cliente
- Var3: tiempo de vida útil del artículo
- Var4: estado civil del cliente
- Var5: número de sucursales del banco

TIPOS DE VARIABLE

Variable cualitativa o categórica: aquellas que expresan cualidades o categorías.

- 1) Variable cualitativa ordinal: tienen un orden de importancia. Ejm: Var2
- 2) Variable cualitativa nominal: no tienen orden de importancia. Ejm: Var4

Variable cuantitativa: aquellas que resultan de hacer mediciones o conteos. Se expresan en forma numérica

- 1) Variable cuantitativa discreta: resultan de hacer conteos. Están asociadas a números enteros. Ejm: Var5
- 2) Variable cuantitativa continua: resultan de hacer mediciones. Están asociadas a números reales. Ejm: Var1, Var3

Observación: es el dato registrado.

Parámetro: Es una constante que caracteriza a una población.

	<u>Parámetro</u>	<u>Valor estadístico</u>
1. Media:	μ	\bar{x}
2. Mediana:	Me	me
3. Moda	Mo	mo
4. Varianza poblacional:	σ^2	s^2



5. Desviación Estándar:	σ	s
6. Proporción:	P	p
7. Diferencia de Medias:	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
8. Diferencia de proporciones:	$P_1 - P_2$	$p_1 - p_2$
9. Coeficiente de correlación:	ρ	r

Valor estadístico: Es un valor calculado usando las observaciones de la muestra

ESTADISTICA

“Ciencia que recoge, organiza, presenta, analiza e interpreta datos con el fin de propiciar la toma de decisiones más eficaz”.

“Es una rama de las Matemáticas Aplicadas que brinda métodos y procedimientos para organizar y evaluar una investigación científica con el fin de tomar decisiones más confiables, cuando prevalecen condiciones de incertidumbre”.

Estadística Descriptiva: Métodos para organizar, resumir y presentar datos de manera informativa.

Estadística Inferencial: Métodos para determinar una propiedad de una población con base en la información de un muestra.



Ejercicios: Detectar los conceptos básicos en cada uno de los siguientes casos:

1. El peso medio de mujeres de 30 a 40 años en una ciudad es de 143 lb. Un estudio realizado en 16 mujeres de esas edades consiste en la aplicación de una dieta vegetariana para disminuir su peso. Se encontró que $\bar{x} = 138$ lb y $s = 13$ lb. ¿La dieta cumple su objetivo?
2. El número de accidentes mortales en una ciudad es, en promedio, de 12 mensuales. Después de una campaña de señalización y mejoramiento de las vías urbanas se contabilizaron en 6 meses sucesivos: 8, 11, 9, 7, 10, 9 accidentes mortales. ¿Fue efectiva la campaña?
3. Por muchos años un quinto de las casas en cierta ciudad se calientan con petróleo. Actualmente, una compañía petrolera afirma que esta proporción ha disminuido. ¿Tenemos razón en dudar de esta afirmación?. Para probar esta afirmación se hizo un estudio que consistió de una muestra aleatoria de 1000 casas de esa ciudad y se encontró que 136 casa se calientan con petróleo.
4. En una muestra de 200 piezas inspeccionadas, se encontró 10 piezas defectuosas. Se puede afirmar que la proporción de piezas defectuosas producidas por la fábrica es mayor del 8%?



ORGANIZACIÓN DE DATOS

Las herramientas usadas en la organización de datos son aplicadas según el tipo de variable en estudio, de la siguiente manera:

Variable cuantitativa continua

Una variable

- Tabla de frecuencias
- Histograma
- Polígono de frecuencias
- Ojiva

Dos variables

- Diagrama de dispersión

Variable cuantitativa discreta

- Diagrama de líneas

Variable categórica

- Diagrama de barras
- Diagrama circular (*pie chart*)
- Diagrama de Pareto



Ejemplo:

Como parte de un estudio para conocer la aceptación de la nueva mega tienda “Vendo” ubicada en la ciudad de Mayagüez, se eligió una muestra de 35 clientes para conocer sus impresiones. Los resultados son los siguientes:

cliente	Razón de visita	Gasto semanal	Ingreso Mensual	Número de hijos	Forma de Pago
1	Oferta	66.0	1200	2	Efectivo
2	Guardería	72.5	1500	1	Crédito
3	Crédito	79.1	2100	3	Crédito
4	Oferta	82.7	2000	3	Efectivo
5	Guardería	55.3	1500	1	Efectivo
6	Parking	100.1	2200	2	Crédito
7	Aire	35.3	1450	3	Efectivo
8	Crédito	60.4	1310	1	Crédito
9	Aire	57.2	1150	2	Efectivo
10	Parking	140.0	2320	0	Crédito
11	Crédito	69.1	1350	2	Efectivo
12	Parking	73.1	1640	1	Crédito
13	Guardería	75.3	1680	3	Crédito
14	Aire	30.0	1100	0	Efectivo
15	Parking	95.2	1850	2	Efectivo
16	Guardería	65.3	1410	1	Efectivo
17	Crédito	68.0	1580	3	Crédito
18	Parking	115.3	2110	0	Efectivo
19	Parking	130.2	2180	2	Crédito
20	Aire	48.4	1640	3	Crédito
21	Guardería	86.0	1840	2	Crédito
22	Parking	102.2	1950	3	Efectivo
23	Oferta	50.1	1230	2	Efectivo
24	Crédito	101.2	2000	2	Crédito
25	Parking	102.2	2810	3	Crédito
26	Oferta	58.1	1530	4	Efectivo
27	Crédito	90.3	1980	2	Crédito
28	Parking	119.1	2900	4	Crédito
29	Oferta	125.1	2680	3	Efectivo
30	Crédito	70.2	1970	2	Crédito
31	Parking	118.4	2560	3	Crédito
32	Oferta	110.1	2180	4	Crédito
33	Crédito	84.3	1980	3	Efectivo
34	Oferta	77.2	2050	2	Crédito
35	Oferta	104.2	2500	4	Crédito

DISTRIBUCION DE FRECUENCIAS, de la variable cualitativa:
“Razón de visita”

Clases	Razón de visita	frecuencia absoluta	frecuencia porcentual (%)
1	Parking	10	28.57
2	Crédito	8	22.86
3	Oferta	8	22.86
4	Guardería	5	14.29
5	Aire	4	11.43
	TOTAL	35	100.0

TABLA DE FRECUENCIAS: EXCEL

- 1) Escribir en la columna G, las clases de respuesta: Parking, Crédito, Oferta, Guardería, Aire. Hacer uso de COUNTIF

1	Razón	Gastos	Ingresos	Hijos	Pago				
2	Oferta	66.0	1200	2	Efectivo	Parking	=COUNTIF(A\$2:A\$36,G2)		
3	Guardería	72.5	1500	1	Crédito	Crédito			
4	Crédito	79.1	2100	3	Crédito	Oferta			
5	Oferta	82.7	2000	3	Efectivo	Guardería			
6	Guardería	55.3	1500	1	Efectivo	Aire			

- 2) Hacer una copia en las demás respuestas: conteo o frecuencias absolutas
- 3) Hacer el cálculo de las frecuencias porcentuales

GRAFICO CIRCULAR: EXCEL

- 1) Seleccionar la frecuencia absoluta de la variable Razón
- 2) Seleccionar el icono de gráficos, seleccionar PIE, ...
Series/Category Labels: G2:G6, Next
Data Labels: Category name, Percentage

Gráfico Circular: Razón de Preferencia

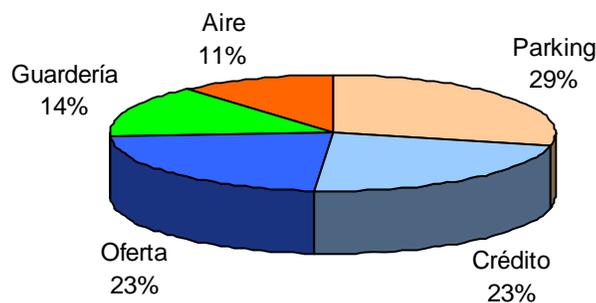
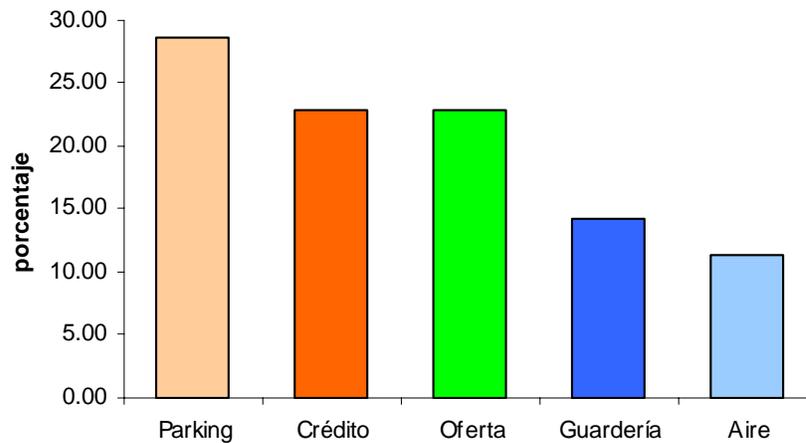


Gráfico de Barras: Razón de Preferencias



DISTRIBUCION DE FRECUENCIAS, de la variable cuantitativa discreta
“Número de hijos”

Número de hijos en la familia	frecuencia absoluta	frecuencia porcentual (%)
0	3	8.57
1	5	14.29
2	12	34.29
3	11	31.43
4	4	11.43
TOTAL	35	100

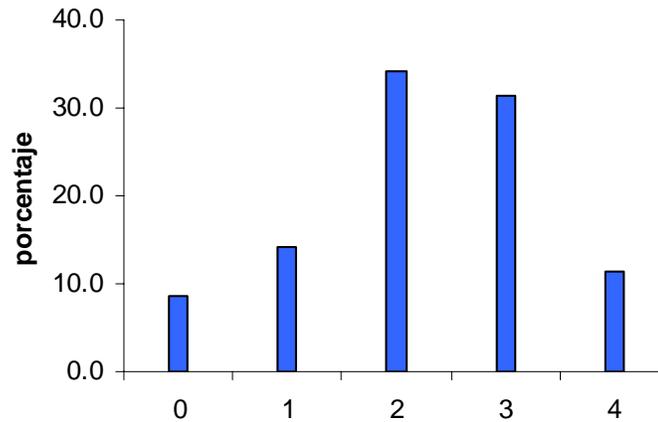
TABLA DE FRECUENCIAS: EXCEL

Es el mismo procedimiento para hacer la tabla de frecuencias de la variable: Razón de Preferencias. Hacer uso de COUNTIF

GRAFICO DE LINEAS: EXCEL

Es el gráfico de barras, pero con un ajuste del ancho de las barras

Gráfico de Líneas: Número de Hijos



DISTRIBUCION DE FRECUENCIAS, de la variable cuantitativa continua
“Gasto semanal en la tienda VENDO”

Paso 1.- Hallar el rango o amplitud de los datos

$$\text{Rango} = \text{obs. mayor} - \text{obs. menor}$$

$$\text{Rango} = 140.0 - 30.0 = 110.0$$

Paso 2.- Hallar el número de clases (k). Dos maneras:

a) Por la experiencia del investigador, usualmente

$$4 \leq k \leq 15$$

b) Por la fórmula de Sturges

$$k = 1 + 3.3 \log(n)$$

$$k = 1 + 3.3 \log(35) = 6.095 \cong 6$$

Paso 3.- Hallar el Tamaño del Intervalo de Clase (TIC)

$$TIC = \frac{\text{Rango}}{k}$$

- Igual # decimales que los datos
- Redondeo por exceso

$$TIC = \frac{110.0}{6} = 18.333 \cong 18.4 \text{ (redondeo por exceso)}$$

Paso 4.- Hallar los intervalos de clase y realizar conteo de datos

Gasto mensual (intervalos de clase)	Promedio de clase	frecuencia absoluta	Frecuencia porcentual	Frecuencia acumulada absoluta	Frecuencia acumulada relativa
[30.0 – 48.4>	39.2	2	5.71	2	5.71
[48.4 – 66.8>	57.6	8	22.86	10	28.57
[66.8 – 85.2>	76.0	10	28.57	20	57.14
[85.2 – 103.6>	94.4	7	20.00	27	77.14
[103.6 – 122.0>	112.8	5	14.29	32	91.43
[122.0 – 140.4]	131.2	3	8.57	35	100.00

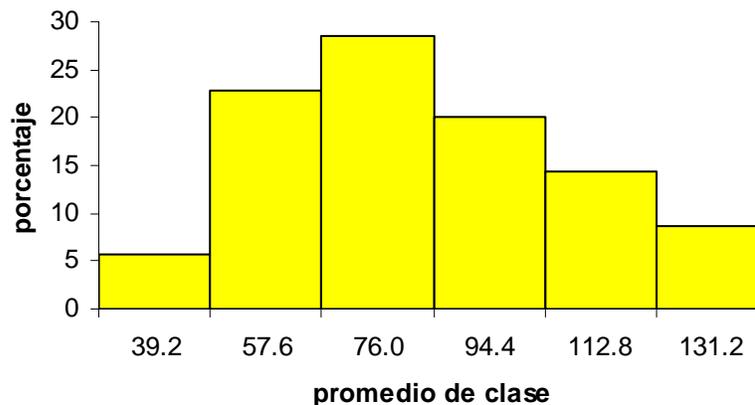
TABLA DE FRECUENCIAS: EXCEL

- 1) Para completar los pasos 1, 2 y 3, se usan los comandos: MAX, MIN, LOG10
- 2) Codificar los datos de la variable Gastos. Uso del comando IF

`=IF(B2<48.4,1,(IF(B2<66.8,2,(IF(B2<85.2,3,(IF(B2<103.6,4,(IF(B2<122,5,6))))))))))`

- 3) Copiar el procedimiento anterior en toda la columna
- 4) Calcular los límites de clase. Uso del comando SUM
- 5) Calcular el promedio de la clase, también llamado *marca*
- 6) Calcular las frecuencias absolutas y porcentuales. Uso del comando COUNTIF
- 7) Calcular las frecuencias acumuladas absolutas y porcentuales: Uso del comando SUM

Histograma: Gastos semanales

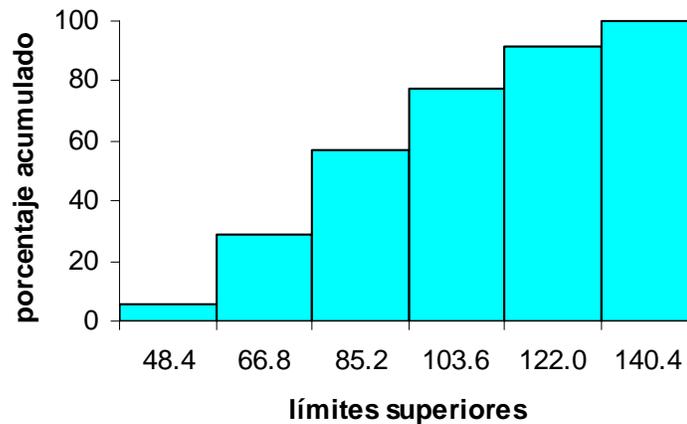


HISTOGRAMA: EXCEL

Es el gráfico de barras sobre las frecuencias porcentuales, pero con un ajuste del ancho de las barras

En el eje horizontal está representado el *promedio de clase o marca*

Ojiva: Frecuencias acumuladas



OJIVA: EXCEL

Es el gráfico de barras sobre las frecuencias acumuladas porcentuales, pero con un ajuste del ancho de las barras

En el eje horizontal está representado el límite superior de clase

Diagrama de dispersión: También llamado “Scatterplot”, muestra la dispersión de datos desde dos variables. Es usado para detectar la posible relación entre las dos variables.

Diagrama de dispersión

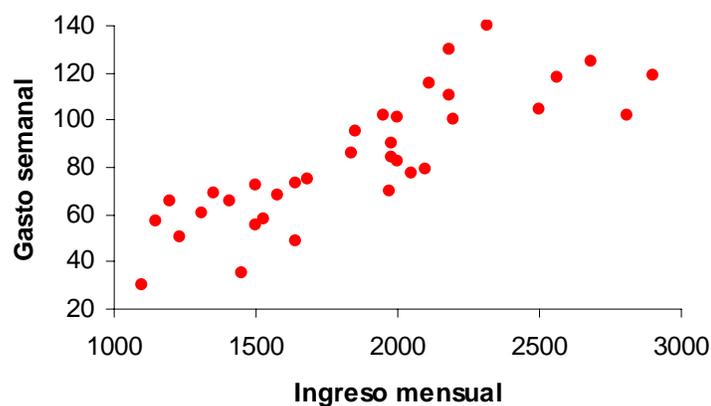




DIAGRAMA DE DISPERSION: EXCEL

Seleccionar el icono de gráficos y seleccionar XY(Scatter)
Representar la variable ingreso mensual en el eje horizontal
Representar la variable gasto semanal en el eje vertical

Tablas de contingencia: Muestra en forma simultánea la frecuencia de dos variables categóricas. En este caso: “Forma de Pago” y “Razón de visita”.

	Parking	Crédito	Oferta	Guardería	Aire	TOTAL
Crédito	7	6	3	3	1	20
Efectivo	3	2	5	2	3	15
TOTAL	10	8	8	5	4	35

TABLA DE CONTINGENCIA: EXCEL

- 1) Ordenar la variable Pago
- 2) Escribir las categorías de la variable, Forma de Pago: Crédito, Efectivo
- 3) Escribir las categorías de la variable, Razón de preferencia: Parking, Crédito, Oferta, Guardería, Aire.
- 4) Hacer el conteo en las celdas correspondientes. Usar COUNTIF

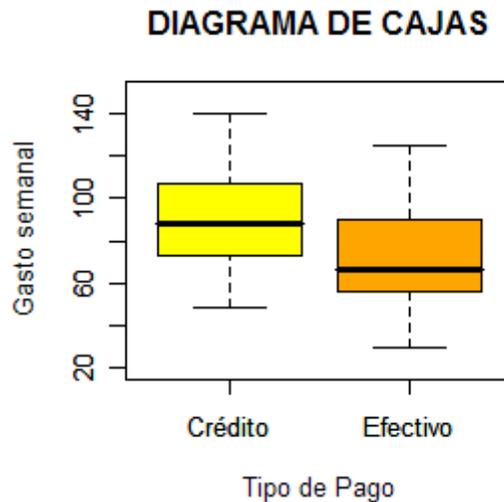
Diagrama de Tallos y Hojas

Una forma de representar las frecuencias de los datos

Diagrama de tallos y hojas para la variable Gastos

3		05
4		8
5		0578
6		05689
7		033579
8		346
9		05
10		01224
11		0589
12		5
13		0
14		0

Diagrama de Caja: También llamado “Boxplot”, muestra la dispersión de la variable en estudio. Es usando para comparar la variabilidad de dos o más conjuntos de datos.



Utilizando R

Se debe instalar la librería *xlsReadWrite*, que lee archivo de datos EXCEL 2003

- 1) Seleccionar *Packages*
- 2) Seleccionar *Install package (s)...*
- 3) Seleccionar *Canada (BC)*
- 4) Seleccionar *xlsReadWrite*
- 5) Escribir: `library(xlsReadWrite)`
- 6) Escribir: `dat=read.xls("c:/folder/ejemplo1")`

Diagrama Boxplot: Gastos versus Pago

```
boxplot(Gastos~Pago,dat,  
        subset = Pago == "Crédito", col = "yellow",  
        main="DIAGRAMA DE CAJAS",  
        xlab="Tipo de Pago",  
        ylab="Gasto semanal",ylim = c(20,150))  
boxplot(Gastos~Pago,dat,add = TRUE,  
        subset = Pago == "Efectivo", col = "orange")
```

Diagrama Boxplot: Gastos versus Hijos

```
boxplot(Gastos~Hijos,dat, col = "yellow",  
        main="DIAGRAMA DE CAJAS",  
        xlab="Número de Hijos",  
        ylab="Gasto semanal",ylim = c(20,150))
```



DIAGRAMA DE CAJAS

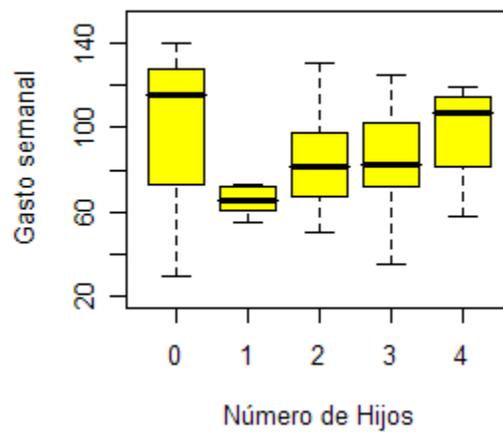


Diagrama de Tallos y Hojas: Gastos
`stem(dat[,2],scale=2)`

MEDIDAS DE TENDENCIA CENTRAL

Son medidas que se ubican aproximadamente al centro de un conjunto de datos, cuando estos están ordenados.

- Media aritmética (media o promedio muestral)
- Mediana
- Moda

MEDIA MUESTRAL (\bar{x})

Dado un conjunto de observaciones de una variable cuantitativa: x_1, x_2, \dots, x_n . La media muestral se define como:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

EXCEL: AVERAGE

Ejercicios

a) serie numérica: 1, 13, 7, 10, 4

media aritmética: $\bar{x} =$

b) serie numérica: 1, 1, 1, 1, 1, 1, 2, 31

media aritmética: $\bar{x} =$

Propiedades de la media aritmética:

- 1) Es una medida única.
- 2) Es afectada por todas las observaciones e influida por las magnitudes absolutas de los valores extremos de la serie.
- 3) $\sum_{i=1}^n (x_i - \bar{x}) = 0$

Observación: Si una muestra de tamaño n y promedio \bar{x} es dividida en dos grupos de n_1 y n_2 elementos con promedios grupales de \bar{x}_1 y \bar{x}_2 respectivamente; entonces se cumple lo siguiente:

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n} \quad n = n_1 + n_2$$



Ejercicio

El promedio de tiempo de vida útil de un grupo de 15 bombillas de la marca A es 180 horas. El promedio de tiempo de vida útil de 8 bombillas de la marca B es 195 horas. Cuál será el promedio de tiempo de vida útil de las 23 bombillas en estudio?.

MEDIANA MUESTRAL (*me*)

Es el valor central de un conjunto de observaciones cuando los datos están ordenados.

EXCEL: MEDIAN

Ejemplo:

El tiempo mensual, en horas, dedicados a estudiar son:

24.2, 15.8, 25.0, 18.3, 23.6, 14.3

Los datos ordenados son

14.3, 15.8, 18.3, 23.6, 24.2, 25.0

La mediana muestral calculada es:

$$me = \frac{18.3 + 23.6}{2} = 20.95$$

“El 50% de personas dedica mensualmente, menos de 20.95 horas en estudio”.

Propiedades de la mediana:

1. La mediana depende esencialmente del número *n* de observaciones y no de la magnitud de los valores involucrados.
2. Apropiaada para resumir series que contienen observaciones extremas.

Ejemplo

a) serie numérica: 1, 13, 7, 10, 4 ordenamiento: 1, 4, 7, 10, 13 → *me* =

b) serie numérica: 1, 1, 1, 1, 2, 2, 2, 3, 1 → *me* =

MODA MUESTRAL (*mo*)

Dado un conjunto de datos, la moda muestral es el valor que se repite con mayor frecuencia.

EXCEL: MODE

Ejemplo:

El tipo de apariencia registrada: (1) Normal, (2) Defectuoso, en un grupo de artículos es:

1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2



Se observa que la moda muestral, $mo = 1$
“Entre los artículos en estudio, el tipo de apariencia más frecuente es el Normal.”

Ejercicios

- a) serie numérica: 5, 0, 3, 1, 3, 1, 2, 2, 6, 4, 1 mo =
b) serie numérica: 2, 4, 1, 4, 3, 6, 1, 6, 3, 6, 4 mo =
c) serie numérica: 1, 2, 5, 3, 7, 8, 10 mo =

MEDIDAS DE POSICIÓN

Son medidas de tendencia central que dividen al conjunto de datos en dos, tres, cuatro, ..., etc. partes iguales, cuando los datos están ordenados.

CUARTILES

Son tres valores que dividen al conjunto de datos en cuatro partes iguales.

Q₁: Primer cuartil.- el 25% de los datos son menores que Q₁.

Q₂: Segundo cuartil.- coincide con la mediana de los datos.

Q₃: Tercer cuartil.- el 75% de los datos son menores que Q₃.

EXCEL: QUARTILE

PERCENTILES

Son noventa y nueve valores que dividen al conjunto de datos en 100 partes iguales.

Se denotan por: P₁, P₂, P₃,...,P₉₈, P₉₉. Así P₉₀, por ejemplo, indica que el 90% de los datos son menores que P₉₀.

EXCEL: PERCENTILE

Las equivalencias, son por ejemplo:

$$Q_1 = P_{25}$$

$$Q_2 = P_{50} = \text{mediana}$$

$$Q_3 = P_{75}$$

Ejemplo

Serie numérica: 17, 22, 17, 18, 17, 19, 26, 14, 33, 21



Ejercicios:

- 1) Cálculo de medidas de tendencia central y de posición para los datos de las variables Gastos e Ingresos.

medida	Gastos	Ingresos
media	83.349	1869.429
mediana	79.1	1950
Q1	65.65	1500
Q2	79.1	1950
Q3	102.2	2145.0
P25	65.65	1500
P75	102.2	2145.0
P80	105.4	2184.0

- 2) Calcular las medidas de tendencia central y de posición para los datos de la variable Gastos dentro de las categorías de la variable Pago.
- 3) Calcular las medidas de tendencia central y de posición para los datos de la variable Ingresos dentro de las categorías de la variable Pago.



MEDIDAS DE VARIABILIDAD

Son indicadores del grado de dispersión de los datos.

- Rango o Amplitud
- Varianza
- Desviación estándar
- Coeficiente de variabilidad
- Desviación intercuartílica

RANGO MUESTRAL

Rango = observación mayor – observación menor

EXCEL: MAX, MIN

VARIANZA MUESTRAL

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

EXCEL: VAR

DESVIACION ESTANDAR

$$S = \sqrt{\text{varianza}}$$

EXCEL: STDEV

COEFICIENTE DE VARIABILIDAD

$$CV = \frac{S}{\bar{x}} \times 100$$

EXCEL: STDEV / AVERAGE

DESVIACION INTERCUARTILICA

$$\text{Desviación Intercuartílica} = p_{75} - p_{25}$$

EXCEL: QUARTILE



Ejercicios:

- 1) Cálculo de medidas de variabilidad para los datos de las variables Gastos e Ingresos.

medida	Gastos	Ingresos
Rango	110.000	1800.000
Varianza	736.2	228029.1
Desv. Estándar	27.13292	477.5239
Coef. Variabilidad	32.55355	25.54384
Desv. Intercuartílica	36.6	645.0

- 2) Calcular las medidas de variabilidad para los datos de la variable Gastos dentro de las categorías de la variable Pago.
- 3) Calcular las medidas de variabilidad para los datos de la variable Ingresos dentro de las categorías de la variable Pago.

Propiedad

Si una variable Y está en función de otra variable X, en forma lineal

$$Y = a \pm b X \quad a, b \text{ son constantes}$$

Entonces se cumple lo siguiente:

- 1) $\bar{Y} = a \pm b \bar{X}$
- 2) $\text{Var}(Y) = b^2 \text{Var}(X)$, también: $S_Y^2 = b^2 S_X^2$
- 3) $S_Y = |b| S_X$
- 4) $CV(Y) = \frac{S_Y}{\bar{Y}} \times 100 = \frac{|b| S_X}{a \pm b \bar{X}} \times 100$

Ejemplo: En el mes de Julio el sueldo diario de trabajadores de construcción es en promedio \$83.36. En el mes de Agosto el sueldo diario incrementó en 15%, respecto al mes anterior



a) Cuál es el promedio del sueldo diario en Agosto.

Sean X : sueldo diario en Julio
 Y : sueldo diario en Agosto

$$Y = X + 0.15X \quad \rightarrow \quad Y = 1.15X$$

Por propiedad $\bar{Y} = 1.15\bar{X} \quad \rightarrow \quad \bar{Y} = 1.15 \times 83.36 = 95.86$

b) Cuál es la desviación estándar del sueldo diario en Agosto, si en Julio fue \$25.32

$$Y = 1.15X \quad \rightarrow \quad S_Y = 1.15 S_X$$
$$S_Y = 1.15 \times 25.32 = 29.12$$

Ejercicio: Si la transformación de datos hubiese sido $Y = 5 + 1.15X$

Calcular la media, varianza, desviación estándar y coeficiente de variabilidad de la nueva variable.



PROBABILIDADES

La teoría de probabilidades es la base de la teoría estadística que permite generar indicadores de confiabilidad o riesgo, usados para tomar decisiones. Se debe tener presente los siguientes conceptos.

Experimento aleatorio (ε)

Es una operación cuyo resultado no se puede predecir con certeza hasta que se realice.

ε_1 : Lanzar un dado

ε_2 : Elegir una persona de un grupo numerado del 1 al 10

ε_3 : Determinar el Exito o Fracaso, aleatoriamente, de dos proyectos presentados

Espacio muestral (Ω)

Es el conjunto de resultados básicos posibles de un experimento aleatorio.

$$\Omega_1 = \{1, 2, 3, 4, 5, 6\}$$

$$\Omega_2 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

$$\Omega_3 = \{(E, E), (E, F), (F, E), (F, F)\}$$

Evento (A, B, C, ...)

Es un subconjunto del espacio muestral

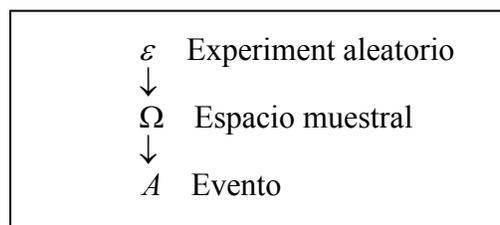
$$B = \{\text{resultado par}\} = \{2, 4, 6\}$$

$$C = \{\text{persona con número impar, mayor de 5}\} = \{7, 9\}$$

$$D = \{\text{que los resultados sean los mismos}\} = \{(E, E), (F, F)\}$$

Esquema de probabilidad

En todo estudio de aplicación de probabilidades se distingue el siguiente esquema



Ejercicios:

- 1) Al lanzar un dado, ¿cuál es probabilidad de obtener un número par?
- 2) De una caja de 10 artículos, donde 3 de ellos son defectuosos, se selecciona uno al azar, ¿cuál es probabilidad de que sea defectuoso?
- 3) Al jugar una “loto”, ¿cuál es probabilidad de ganar?

Definición clásica de probabilidad (definición *a priori*)

La probabilidad de ocurrencia de un evento A , denotado por $P(A)$, se define como la relación del número de elementos posibles del evento, $\#(A)$ y el número de elementos totales del espacio muestral, $\#(\Omega)$

$$P(A) = \frac{\#(A)}{\#(\Omega)}$$

Definición frecuentista de probabilidad (definición *a posteriori*)

Si un experimento consiste de n ensayos, y un evento A ocurre m veces, la probabilidad o frecuencia relativa de A , denotada $P(A)$, esta dada por

$$P(A) = \frac{m}{n}$$

Ejercicio:

Una muestra de 270 artículos de la producción del fin de semana en una fábrica, se organizó en una tabla de contingencia según las variables: “Calidad de producto” y “Turno de fabricación”

		Turno de fabricación			TOTAL
		I	II	III	
Calidad del producto	Bueno	50	70	60	180
	Regular	30	20	25	75
	Malo	5	8	2	15
	TOTAL	85	98	87	270

- Cuál es la probabilidad de que la calidad del producto sea Bueno?
- Cuál es la probabilidad de que la calidad del producto sea Malo?
- Cuál es la probabilidad de que el artículo haya sido fabricado en el turno II?
- Cuál es la probabilidad de que la calidad de producto sea Bueno y haya sido fabricado en el turno III?
- Cuál es la probabilidad de que el artículo haya sido fabricado en el turno I y sea de calidad Regular?



VARIABLE ALEATORIA

Una variable aleatoria es una función que asigna a cada posible resultado del experimento aleatorio, un número real

Ejemplo

Considere un experimento que consiste en lanzar dos monedas.
Se define una variable aleatoria X : número de caras.

Resultados del experimento	variable aleatoria
(T, T)	→ $X = 0$
(T, C)	→ $X = 1$
(C, T)	→ $X = 1$
(C, C)	→ $X = 2$

Rango de la variable aleatoria (R_X), es el conjunto de valores posibles de la variable aleatoria

$$R_X = \{0, 1, 2\}$$

Distribución de probabilidad, $P(X = x)$, también llamada tabla de probabilidad

R_X	0	1	2
$P(X = x)$	1/4	2/4	1/4

DISTRIBUCIONES DE VARIABLES ALEATORIAS DISCRETAS

Distribución Binomial

Una variable aleatoria discreta X tiene una distribución binomial si su función de probabilidad es dada por:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, 2, \dots, n$$

p = probabilidad de éxito en cada prueba

n = número de pruebas

v.a. X = número de éxitos en los n ensayos

Media de la variable binomial: $n \times p$

Varianza de la variable binomial: $n \times p \times (1-p)$



Ejemplo:

En una agencia bancaria, el 40% de los clientes tienen certificado bancario. Si se eligen 8 clientes al azar, cuál es la probabilidad de encontrar:

a) Exactamente 6 clientes con certificados bancarios

v.a. $X = \#$ de clientes con certificado bancario; $p = 0.40$; $n = 8$

$$P(X = 6) = \binom{8}{6} 0.40^6 (1 - 0.40)^{8-6} = 0.0413$$

=BINOMDIST(6,8,0.4,FALSE)

b) Todos los clientes tienen certificado bancario: $P(X = 8)$

=BINOMDIST(8,8,0.4,FALSE)

c) Ningún cliente tenga certificado bancario: $P(X = 0)$

=BINOMDIST(0,8,0.4,FALSE)

d) Al menos un cliente tiene certificado bancario: $P(X \geq 1)$

=1-BINOMDIST(0,8,0.4,FALSE)

e) A lo más 6 clientes tienen certificado bancario: $P(X \leq 6)$

=BINOMDIST(6,8,0.4,TRUE)

Ejercicio

Suponga que en una comunidad el 30% está desempleado. Si una muestra aleatoria de 10 personas es escogida de ésta comunidad

a) ¿Cuál es la probabilidad de que 4 personas estén desempleados?

b) ¿Cuál es la probabilidad de que ninguno este desempleado?

c) ¿Cuál es la probabilidad de que por lo menos una persona esté desempleada?

d) ¿Cuál es la probabilidad de que entre 1 y 4 personas, inclusive, sean desempleados?

e) Más de 8 desempleados.

f) Menos de 3 desempleados.

g) Como máximo 5 desempleados.

h) Como mínimo 2 desempleados.



Distribución de Poisson

Una variable aleatoria discreta X tiene una distribución de Poisson si su función de probabilidad es dada por:

$$P(X = x) = \frac{e^{-\mu} \mu^x}{x!} \quad x = 0, 1, 2, 3, \dots$$

$$\mu = \lambda t$$

λ : razón de ocurrencia de éxitos en una unidad de intervalo

t : número de unidades de intervalo

v.a. X = número de éxitos en t unidades de intervalo

Media de la variable Poisson: μ

Varianza de la variable Poisson: μ

Ejemplo

En una inmobiliaria se ha determinado que el número promedio de casas vendidas en un día laborable es 1.6 casas/día. Si el número de casas vendidas es una variable Poisson, calcule la probabilidad de que en un día cualquiera:

a) Se vendan exactamente 4 casas: $P(X = 4)$

En este caso $t=1$ y $\lambda=1.6 \rightarrow \mu = \lambda t = 1.6$

$$P(X = 4) = \frac{e^{-1.6} 1.6^4}{4!} = 0.0551312$$

$$= \text{POISSON}(4, 1.6, \text{FALSE})$$

b) No se venda ninguna casa: $P(X = 0)$

$$= \text{POISSON}(0, 1.6, \text{FALSE})$$

c) Se venda por lo menos una casa: $P(X \geq 1) = 1 - P(X = 0)$

$$= 1 - \text{POISSON}(0, 1.6, \text{FALSE})$$

d) Se venda entre 2 y 5 casas, inclusive: $P(2 \leq X \leq 5)$

$$P(X=2) + P(X=3) + P(X=4) + P(X=5)$$



e)Cuál es la probabilidad de vender 4 casas en dos días?

En este caso $t=2$ y $\lambda=1.6 \rightarrow \mu = \lambda t = (2)(1.6) = 3.2$

$$P(X = 4) = \frac{e^{-3.2} 3.2^4}{4!} = 0.1780928$$

=POISSON(4,3.2,FALSE)

f)Cuál es la probabilidad de vender a lo mas 4 casas en dos días?

En este caso $t=2$ y $\lambda=1.6 \rightarrow \mu = \lambda t = (2)(1.6) = 3.2$

$$P(X \leq 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)$$

=POISSON(4,3.2,TRUE)

g)Cuál es la probabilidad de vender al menos 4 casas en dos días?

En este caso $t=2$ y $\lambda=1.6 \rightarrow \mu = \lambda t = (2)(1.6) = 3.2$

$$P(X \geq 4) = 1 - P(X \leq 3)$$

=1-POISSON(4,3.2,TRUE)

Ejercicios

- 1) En el laboratorio de una farmacéutica se determinó que en un dm^3 de agua se encuentran distribuidos al azar 500 particular infecciosas. En $6 cm^3$, cuál es la probabilidad de encontrar:
 - a) Exactamente 8 partículas infecciosas
 - b) Ninguna partícula infecciosa
 - c) Al menos una partícula infecciosa.

- 2) La experiencia de un gerente bancario ha determinado que en el horario de 8:30 y 9:30 de la mañana se atienden a 48 clientes. Si el número de clientes atendidos en una hora es una variable Poisson. Determinar la probabilidad de que en 8 minutos de ese horario se atiendan exactamente 10 clientes.

DISTRIBUCIONES DE VARIABLES ALEATORIAS CONTINUAS

Definición: La función de densidad de una variable aleatoria continua X denotada $f(x)$, es una gráfica tal que:

- El área debajo de la misma entre cualquiera de dos puntos a y b es igual a la probabilidad de que X asuma valores entre estos dos números
- El área total debajo de la gráfica es igual a 1.

Distribución Normal

Una variable aleatoria continua X tiene una distribución normal si su función de densidad es dada por:

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$

X = variable aleatoria continua

μ = valor medio de la variable aleatoria X ; $-\infty < \mu < \infty$

σ = desviación estándar de la variable aleatoria X ; $\sigma > 0$

$\pi = 3.1416$ $e = 2.718282$

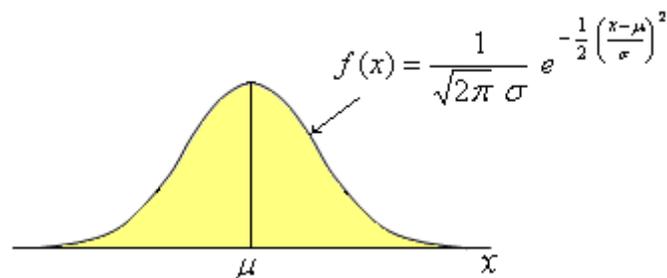
Notación: v.a. $X \sim N(\mu, \sigma^2)$

“la variable aleatoria X tiene distribución normal con media μ y varianza σ^2 ”

Características de la distribución normal:

1. Es simétrica con respecto a su media μ .
2. Es asintótica en eje horizontal
3. La media, mediana y moda son iguales.
4. El área total bajo la curva es 1.
5. Para cada valor de μ y σ^2 , existe una curva normal. No es única.

Gráfica de la distribución normal





Distribución Normal Estándar

Una variable aleatoria continua Z tiene una distribución normal estándar si su función de densidad es dada por:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad -\infty < z < \infty$$

$$\mu = 0$$

$$\sigma = 1$$

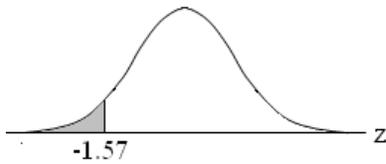
$$\pi = 3.1416 \quad e = 2.718282$$

Notación: v.a. $Z \sim N(0, 1)$

PROBABILIDADES EN LA DISTRIBUCION NORMAL ESTANDAR

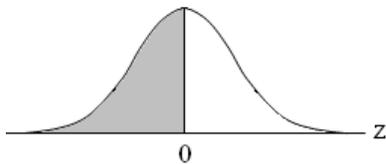
Calcular:

a) $P(Z < -1.57) =$



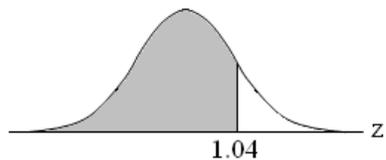
$$= \text{NORMDIST}(-1.57, 0, 1, \text{TRUE})$$

b) $P(Z < 0) =$



$$= \text{NORMDIST}(0, 0, 1, \text{TRUE})$$

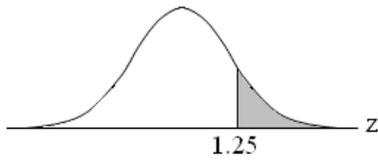
c) $P(Z \leq 1.04) =$



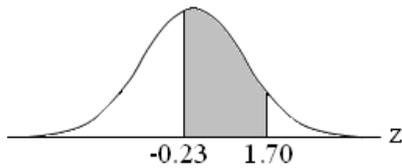
$$= \text{NORMDIST}(1.04, 0, 1, \text{TRUE})$$



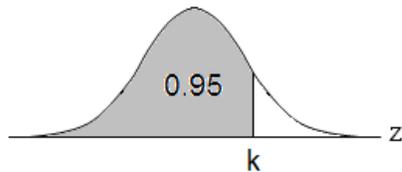
d) $P(Z \geq 1.25) = 1 - P(Z < 1.25)$



e) $P(-0.23 \leq Z \leq 1.70) =$



f) Hallar el valor “k”, tal que: $P(Z < k) = 0.95$



$=\text{NORMINV}(0.95,0,1)$

Ejercicios:

Calcular

- 1) $P(Z > 1.34)$
- 2) $P(Z > -2.1)$
- 3) $P(Z < -1.24)$
- 4) $P(1.1 < Z < 2.2)$
- 5) $P(-2 < Z < 1.85)$
- 6) $P(-2 < Z < -0.84)$

Hallar el valor k, en los siguientes casos

- 1) $P(Z < k) = 0.37$
- 2) $P(Z < k) = 0.90$
- 3) $P(Z > k) = 0.44$
- 4) $P(0.15 < Z < k) = 0.2$

Estandarización de datos

Sea una v.a. $X \sim N(\mu, \sigma^2) \rightarrow$ la transformación $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$

Ejemplo

En una empresa los pagos mensuales de empleados por trabajar en sobretiempo están distribuidas en forma aproximadamente normal con una media de \$200 y una desviación estándar de \$20, entonces la probabilidad de que un empleado, seleccionado al azar en esta empresa, tenga un pago mensual por sobretiempo

a) Mayor de 240 dólares, es

$$\begin{aligned} P(X \geq 240) &= P\left(Z > \frac{240 - 200}{20}\right) \\ &= P(Z \geq 2.0) \\ &= 1 - P(Z < 2.0) \\ &= 1 - 0.9772 \\ &= 0.0228 \end{aligned}$$

b) Entre 150 y 250 dólares, es:

$$\begin{aligned} P(150 \leq X \leq 250) &= P\left(\frac{150 - 200}{20} \leq Z \leq \frac{250 - 200}{20}\right) \\ &= P(-2.5 \leq Z \leq 2.5) \\ &= 0.9938 - 0.0062 \\ &= 0.9876 \end{aligned}$$

Ejercicio

1) Una supervisor ha encontrado que los trabajadores del turno noche, en promedio tardan 10 minutos en realizar una tarea. Si los tiempos requeridos para concluir la tarea están distribuidos en forma aproximadamente normal con una desviación estándar de 3 minutos, encuentre:

- a) La proporción de trabajadores que concluyen la tarea en menos de cuatro minutos.
- b) La proporción de trabajadores que requieren más de cinco minutos para concluir la tarea.
- c) El supervisor ha determinado que en el turno de la noche el 33% de los trabajadores son los más lentos en completar la tarea. Hallar el tiempo mínimo necesario de un trabajador en completar la tarea para ser considerado dentro del grupo de los más lentos. Resp: 11.32 minutos

Distribución T de Student

La variable aleatoria continua X tiene una distribución T de Student, con m grados de libertad, denotada por: v.a. $X \sim t_{(m)gl}$,

El valor m es un número entero positivo que define a la distribución T

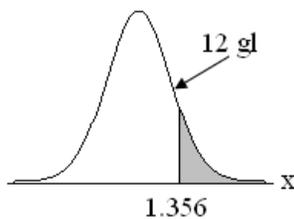
Características de la distribución T

1. Es simétrica respecto a la recta perpendicular al eje horizontal en el punto cero
2. El área total bajo la curva sobre el eje horizontal es una unidad de área.
3. Para cada valor de grado de libertad existe una curva de distribución t.

Ejemplo

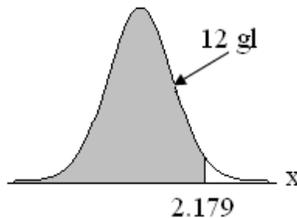
Si $X \sim t_{(12)gl}$, calcular:

1) $P(X > 1.356) = 0.1$



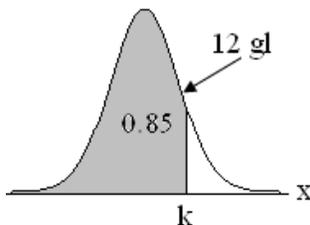
$=TDIST(1.356, 12, 1)$

2) $P(X < 2.179) = 0.975$



$=1-TDIST(2.179, 12, 1)$

3) determinar el k , tal que $P(X < k) = 0.85$



$=TINV(0.3, 12)$

$k = 1.0832$

Ejercicios:

Si $X \sim t_{(18)gl}$

Calcular la probabilidad:

- 1) $P(X > 1.842)$
- 2) $P(X < 1.231)$
- 3) $P(X < 0.824)$
- 4) $P(X > -1.24)$
- 5) $P(X < -2.18)$
- 6) $P(-1.23 < X < 1.23)$

Hallar el valor k en los siguientes casos

7) $P(-k < X < k) = 0.95$

Distribución Ji-Cuadrado

La variable aleatoria continua X tiene una distribución Ji-cuadrado, con m grados de libertad, denotado por: v.a. $X \sim \chi^2_{(m)gl}$

El valor m es un número entero positivo que define a la distribución Ji-cuadrado

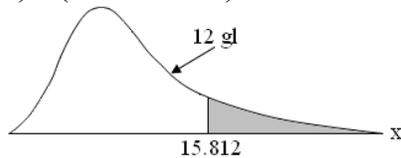
Características de la distribución Ji-cuadrado:

1. Es asimétrica hacia la derecha. La variable sólo toma valores positivos.
2. El área total bajo la curva sobre el eje horizontal es una unidad de área.
3. Para cada valor de grado de libertad existe una curva Ji-cuadrado.

Ejemplo

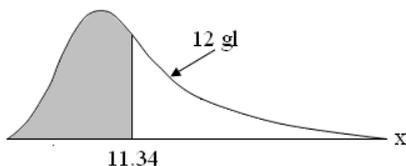
Si $X \sim \chi^2_{(12)gl}$, calcular:

1) $P(X > 15.812) = 0.199999$



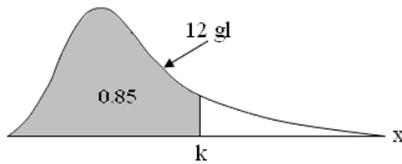
$=\text{CHIDIST}(15.812, 12)$

2) $P(X < 11.34) = 0.499973$



$=1-\text{CHIDIST}(11.34, 12)$

3) determinar el k , tal que $P(X < k) = 0.85$



$$=CHIINV(0.15, 12)$$

$$k = 16.98931$$

Ejercicios:

$$\text{Si } X \sim \chi_{(25)gl}^2$$

Calcular la probabilidad:

- 1) $P(X > 18.842)$
- 2) $P(X < 5.231)$
- 3) $P(X < 17.824)$
- 4) $P(15.23 < X < 31.23)$

Hallar el valor k en los siguientes casos

$$5) P(5.1 < X < k) = 0.95$$

Distribución F de Snedecor

La variable aleatoria continua X tiene una distribución F de Snedecor, con a y b grados de libertad, denotada por: $X \sim F_{(a,b)gl}$

Los valores a y b son enteros positivos que definen a la distribución F

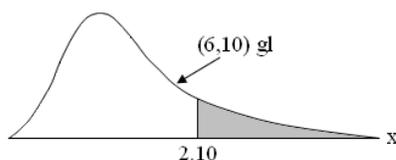
Características de la distribución F:

1. Es asimétrica hacia la derecha. La variable sólo toma valores positivos.
2. El área total bajo la curva sobre el eje horizontal es una unidad de área.
3. Para cada par de valores de grado de libertad existe una curva de distribución F.

Ejemplo:

$$\text{Si } X \sim F_{(6,10)gl}$$

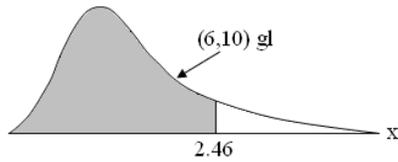
$$1) P(X > 2.10) = 0.1433238$$



$$=FDIST(2.1, 6, 10)$$

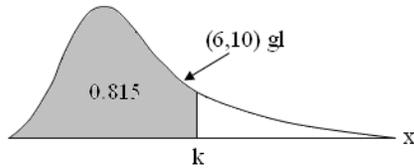


$$2) P(X < 2.46) = 0.90$$



$$=1-FDIST(2.46, 6, 10)$$

$$3) \text{ determinar el } k, \text{ tal que } P(X < k) = 0.815$$



$$=FINV(0.185, 6, 10)$$

$$k = 1.8553804$$

Ejercicios:

$$\text{Si } X \sim F_{(12,27)gl}$$

Calcular la probabilidad:

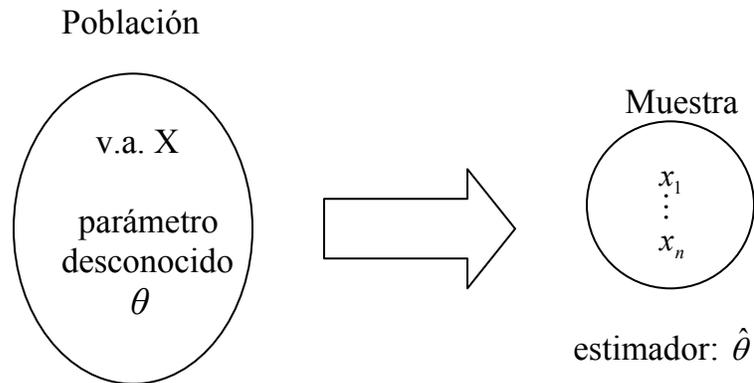
- 1) $P(X > 1.842)$
- 2) $P(X < 0.231)$
- 3) $P(X < 1.824)$
- 4) $P(1.23 < X < 2.23)$

Hallar el valor k en los siguientes casos

$$5) P(0.3 < X < k) = 0.95$$

ESTADISTICA INFERENCIAL

Se ocupa de los procedimientos que nos permiten analizar y extraer conclusiones de una población a partir de los datos de una muestra aleatoria mediante la teoría de probabilidades y de las distribuciones muestrales.



- 1) Estimación de Parámetros
 - Estimación puntual
 - Estimación por intervalo
- 2) Prueba de Hipótesis

ESTIMACION PUNTUAL

Es la estimación del parámetro por medio de un único valor.

Ejemplo: en una muestra se obtiene $\bar{x} = 6.4$ y $s^2 = 30.2$

$\bar{x} = 6.4$, es una estimación puntual del parámetro μ

$s^2 = 30.2$, es una estimación puntual del parámetro σ^2

ESTIMACION POR INTERVALO

El parámetro es $\theta \rightarrow$ El estimador es: $\hat{\theta} = [a, b]$

$\gamma = 1 - \alpha$: Coeficiente de Confianza

$\gamma = 90\%, 95\%$

El estimador $\hat{\theta} = [a, b]$ es un intervalo que brinda un γ % de confianza de contener al verdadero valor de parámetro θ



Intervalo de confianza para la media de una población

a) Si la varianza σ^2 es conocida (*distribución Z*)

$$\text{Intervalo de Confianza: } IC(\mu) = \bar{x} \pm Z_0 \frac{\sigma}{\sqrt{n}}$$

Ejemplo

Un investigador, interesado en obtener una estimación del nivel promedio diario (μ) de óxido de sulfuro que emite una planta industrial, toma una muestra de 10 días, y calcula la media muestral $\bar{x} = 22$. Suponga que se sabe que la variable de interés presenta una distribución aproximadamente normal con una varianza de 45. Construya un intervalo de confianza del 95% para μ .

Solución:

$$\begin{aligned} & \bar{x} \pm 1.96 \sigma / \sqrt{n} \\ & 22 \pm 1.96 \sqrt{\frac{45}{10}} \\ & 22 \pm 1.96(2.12) \\ & 22 \pm 4.16 \\ & (17.84, 26.16) \end{aligned}$$

Interpretación: El intervalo (17.84, 26.16) brinda un 95% de confianza en contener el verdadero valor de μ

b) Si la varianza σ^2 No es conocida (*distribución t*)

$$\text{Intervalo de Confianza: } IC(\mu) = \bar{x} \pm t_0 \frac{S}{\sqrt{n}}$$

Ejemplo

Una muestra de 30 niños de diez años de edad proporcionó un peso medio y una desviación estándar de 36.5 kg. y 5 kg, respectivamente. Suponiendo una población con distribución normal, encuentre los intervalos de confianza de 90% para la media de la población a partir de la cual se obtuvo la muestra.

Solución: coeficiente de confianza = 90%



$$\begin{aligned} & \bar{x} \pm 1.699 s / \sqrt{n} \\ & 36.5 \pm 1.699 \times 5 / \sqrt{30} \\ & 36.5 \pm 1.5509 \\ & (34.94, 38.05) \end{aligned}$$

Intervalo de confianza para una proporción: n grande

En este caso, la estimación por intervalo para la proporción p de éxitos en cierta población, se obtiene mediante los límites

$$\text{Intervalo de Confianza: } IC(p) = \hat{p} \pm z_0 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ejemplo

En una muestra aleatoria de 400 automóviles detenidos en un puesto de revisión, 152 de los conductores llevaban puesto el cinturón de seguridad. Construya el intervalo de confianza del 95% para la proporción real de conductores que llevan puesto el cinturón de seguridad.

$$\text{Ya que } \hat{p} = \frac{152}{400} = 0.38 \quad \implies \quad IC(p) = 0.38 \pm 1.96 \sqrt{\frac{0.38(1-0.38)}{400}}$$

$$IC(p) = (0.332, 0.428)$$



PRUEBA DE HIPOTESIS

Es un método estadístico de comprobación de una hipótesis y es realizado utilizando los valores observados que constituyen la muestra

HIPOTESIS DE INVESTIGACION: es una suposición o reclamo que motiva una investigación. El reclamo pretende describir una característica (parámetro) de la población

HIPOTESIS ESTADISTICA: es una reformulación estadística de una hipótesis de investigación, que refiere al valor de un parámetro.

Se hace uso de dos hipótesis estadísticas complementarias:

- **hipótesis nula**: lo establecido, lo aceptado
- **hipótesis alterna**: el reto, lo nuevo

Ejemplos

Caso 1: Un investigador de una empresa reclama que el tiempo promedio que un trabajador tarda en completar una tarea ha aumentado desde el último estudio efectuado hace dos años, cuyo resultado dio $\mu = 15$ minutos

Hipótesis de investigación: El tiempo promedio (μ) que un trabajador tarda en completar una tarea es mayor de 15 minutos.

Hipótesis estadísticas: $H_0: \mu \leq 15$
 $H_1: \mu > 15$

Caso 2: Un investigador bancario que estudia dos agencias de un banco reclama que los promedios de ahorros diarios en las dos agencias son diferentes

Hipótesis de investigación: Los promedios de ahorros diarios en dos sucursales de un banco son distintos.

Hipótesis estadísticas: $H_0: \mu_1 - \mu_2 = 0$
 $H_1: \mu_1 - \mu_2 \neq 0$

Caso 3: Un investigador industrial cree que con la nueva máquina para fabricar focos eléctricos no habrá grandes problemas en la producción, ya que la proporción de defectuosos es menor del 5%

Hipótesis de investigación: La proporción de defectuosos es menor del 5%.

Hipótesis estadísticas: $H_0: p \geq 0.05$
 $H_1: p < 0.05$



Note que la igualdad siempre se incluye en la hipótesis nula.

TIPOS DE ERROR

La hipótesis nula (H₀) es:

		Cierta	Falsa
Decisión	Rechazar H ₀	Error Tipo I	Decisión correcta
	No rechazar H ₀	Decisión correcta	Error Tipo II

- 1) ERROR TIPO I: cuando se rechaza H₀ siendo realmente verdadera
- 2) ERROR TIPO II: cuando se acepta H₀ siendo realmente falsa

$$P(\text{cometer error tipo I}) = \alpha$$

$$P(\text{cometer error tipo II}) = \beta$$

α : nivel de significación de la prueba

- 3) El “*poder*” de una prueba se define como $1-\beta$ y representa la probabilidad de rechazar H₀ cuando ésta es falsa (rechazar correctamente H₀).
- 4) La *zona de rechazo* es el intervalo o intervalos de valores por los cuales H₀ es rechazado.
- 5) La *zona de no rechazo* es el intervalo de los valores por los cuales H₀ no es rechazado.



Pasos necesarios para realizar una prueba de hipótesis

- 1) Formulación de hipótesis
- 2) Establecer el nivel de significación: α
Usualmente $\alpha = 0.01, 0.02, 0.05, 0.10$
- 3) Determinar la prueba estadística: t, Z, χ^2, F
Establecer las suposiciones de la prueba:
 - La muestra fue elegida al azar
 - La población de donde se extrae la muestra tiene distribución normal ó las muestras seleccionadas son suficientemente grandes
- 4) Determinar las regiones de aceptación y rechazo de H_0
Graficar la distribución correspondiente a la prueba elegida en el pto. 3 y representar el valor correspondiente a nivel de significación
- 5) Realizar el cálculo de la prueba estadística, elegida en el pto. 3
- 6) Establecer las conclusiones de la prueba

Definición

El *p-value*, es la probabilidad de observar un valor muestral tan extremo o más que el valor observado, si la H_0 es verdadera.

- Si el *p-value* < 0.01 , existe una evidencia fuerte en contra de H_0 .
- Si $0.01 < \textit{p-value} < 0.05$, existe evidencia moderada en contra de H_0 .
- Si el *p-value* > 0.05 , existe poca o ninguna evidencia en contra de H_0 .

Prueba de hipótesis acerca de la media

σ^2 es conocido	$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$		$Z_{\text{calculado}} = \frac{\bar{x} - k}{\sigma/\sqrt{n}}$
σ^2 no es conocido	$\frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t (n-1)g.l.$		$t_{\text{calculado}} = \frac{\bar{x} - k}{S/\sqrt{n}}$

Prueba de hipótesis acerca de una proporción

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1) \quad \Rightarrow \quad Z_{\text{calculado}} = \frac{\hat{p} - k}{\sqrt{\frac{k(1-k)}{n}}}$$

Ejercicios:

- 1) El fabricante de llantas radiales con cinturón de acero X-15 para camiones señala que el millaje medio que la llanta recorre antes de que se desgasten las cuerdas es de 60000 millas, con desviación estándar de 5000 millas. Una compañía compró 48 llantas y encontró que el millaje medio para sus camiones es de 59500 millas. ¿Se puede afirmar que el verdadero millaje medio de las llantas es menor de lo que afirma el fabricante?
- 2) Una compañía analiza una nueva técnica para armar un carro de golf; la técnica actual requiere 42.3 minutos, en promedio. El tiempo medio de montaje de una muestra aleatoria de 24 carros, con la nueva técnica, fue de 40.6 minutos y la desviación estándar de 2.7 minutos. ¿Se puede afirmar que el tiempo de montaje con la nueva técnica es más rápida?
- 3) Por mucho tiempo, se ha afirmado que el 60% de los jóvenes de una ciudad, son fumadores. Actualmente un investigador social dice que esta proporción ha disminuido, debido a una campaña de educación en salud. Para probar esta afirmación se hizo un estudio que consistió de una muestra aleatoria de 350 jóvenes de esa ciudad y se encontró que 210 fuman

Prueba de hipótesis acerca de diferencia de medias: muestras independientes

Varianzas poblacionales: son conocidas → Prueba Z

Se considera que los sueldos de trabajadores de la construcción en dos ciudades A y B, son variables con distribución normal, con desviaciones estándar de 4 y 6 dólares, respectivamente. ¿Se puede afirmar que el promedio de sueldos de los trabajadores de la ciudad B es mayor que el promedio de sueldos en la ciudad A?. Use los datos del archivo “hipótesis1.xls”.

z-Test: Two Sample for Means

	ciudadB	ciudadA
Mean	85.07519091	80.05214
Known Variance	36	16
Observations	55	40



Hypothesized Mean Difference	0
z	4.891417259
P(Z<=z) one-tail	5.00563E-07
z Critical one-tail	1.644853627
P(Z<=z) two-tail	1.00113E-06
z Critical two-tail	1.959963985

SI, se puede afirmar que el promedio de sueldos de los trabajadores de la ciudad B es mayor que el promedio de sueldos en la ciudad A

Varianzas poblacionales: no son conocidas → Prueba T

En un estudio reciente se comparó el tiempo (minutos) que pasan juntas las parejas: las parejas en que sólo trabaja uno de los cónyuges versus las parejas en que ambos trabajan. ¿Se puede concluir que en promedio las parejas en que sólo trabaja uno de los cónyuges pasan más tiempo, juntos viendo TV?. Use los datos del archivo “hipótesis2.xls”.

F-Test Two-Sample for Variances

	UnoTrabaja	DosTrabajan
Mean	59.18119286	50.64044
Variance	222.692634	313.6316649
Observations	42	35
df	41	34
F	0.710045123	
P(F<=f) one-tail	0.146817885	
F Critical one-tail	0.58361197	

Hay homogeneidad de varianzas

t-Test: Two-Sample Assuming Equal Variances

	UnoTrabaja	DosTrabajan
Mean	59.18119286	50.64044
Variance	222.692634	313.6316649
Observations	42	35
Pooled Variance	263.918328	
Hypothesized Mean Difference	0	
df	75	
t Stat	2.297072402	
P(T<=t) one-tail	0.012203249	
t Critical one-tail	1.665425374	
P(T<=t) two-tail	0.024406498	
t Critical two-tail	1.992102124	

SI, se puede concluir que en promedio las parejas en que sólo trabaja uno de los cónyuges pasan más tiempo, juntos viendo TV



Prueba de hipótesis de dos muestras: muestras dependientes

La gerencia de una cadena de mueblerías, diseñó un plan de incentivos para sus agentes de ventas. Para evaluar este plan innovador, se seleccionó a 30 vendedores, al azar, y se registraron sus ingresos “antes” y “después” de aplicar el plan. ¿Se puede afirmar que hubo un aumento significativo en el ingreso semanal del vendedor?. Usar los datos del archivo “*hipótesis3.xls*”.

t-Test: Paired Two Sample for Means

	<i>Después</i>	<i>Antes</i>
Mean	495.4862	438.5289
Variance	3635.399466	2895.180249
Observations	30	30
Pearson Correlation	0.120509339	
Hypothesized Mean Difference	0	
Df	29	
t Stat	4.114592909	
P(T<=t) one-tail	0.000146418	
t Critical one-tail	1.699126996	
P(T<=t) two-tail	0.000292837	
t Critical two-tail	2.045229611	

SI, se puede afirmar que hubo un aumento significativo en el ingreso semanal del vendedor

Prueba de hipótesis en tablas de contingencia

Prueba de diferencia de más de dos proporciones

En un estudio se obtuvo una muestra de tres grupos de personas: se preguntó a 100 hombres, 130 mujeres y 90 niños, si les agradaba o no el sabor de una nueva pasta dental. Los resultados fueron los siguientes:

Las hipótesis son:

H_0 : La proporción de “gusto por la nueva pasta dental” es la misma en los tres grupos de personas

H_1 : Al menos en uno de los grupos la proporción es diferente.

Valores observados

	Hombres	Mujeres	Niños	
Les gustó el sabor	60	67	49	176
No les gustó el sabor	40	63	41	144
Total	100	130	90	320

Se calcula la tabla de valores esperados

Valores esperados

	Hombres	Mujeres	Niños	
Le gustó el sabor	55.0	71.5	49.5	176
No les gustó el sabor	45.0	58.5	40.5	144
Total	100	130	90	320

$$\chi^2_{\text{calculado}} = 1.651$$

$$p\text{-valor} = 0.4381$$

=CHITEST(*observados,esperados*)

Prueba de homogeneidad de poblaciones

	Hombres	Mujeres	Niños	
Les gustó el sabor	52	56	45	153
Les resulta indiferente	15	23	11	49
No les gustó el sabor	33	51	34	118
Total	100	130	90	320

Prueba de independencia de variables

Se quiere investigar si existe en realidad una relación entre el “*desempeño en el programa de capacitación*” de la compañía y el “*éxito final en el trabajo*”.

Desde una muestra de 400 empleados sacados de los grandes archivos de una compañía, se obtuvo los siguientes resultados:

Desempeño en el programa de capacitación

Éxito en el trabajo (clasificación de la empresa)	Inferior a lo normal	En el nivel normal	Superior a lo normal	Total
Deficiente	23	60	29	112
Normal	28	79	60	167
Muy bueno	9	49	63	121
Total	60	188	152	400



ANÁLISIS DE REGRESION y CORRELACION

El estudio de las relaciones entre dos o más variables se puede llevar a cabo desde dos puntos de vista:

<u>Análisis de Regresión</u>	⇒	Estudio de la relación funcional existente entre las variables
<u>Análisis de Correlación</u>	⇒	Estudio del grado de asociación existente entre las variables

ANÁLISIS DE REGRESION LINEAL

El objetivo de este análisis es estimar y analizar una ecuación o modelo, que describa la relación funcional existente entre las variables:

$$Y = f(\underbrace{X_1, X_2, \dots, X_p}_{\text{variables independientes}})$$

↖
variable dependiente

Regresión Lineal Simple

$$y = b_0 + b_1 x$$

Regresión Lineal Múltiple

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

ANOVA: Análisis de varianza

Es la **descomposición** y evaluación de la variación total de la variable en estudio (Y), en componentes independientes atribuibles a fuentes de variación que brinden evidencias estadísticas acerca de la significación de X_1, X_2, \dots, X_k como variables explicativa del comportamiento de Y



$$\text{SC del TOTAL} = \text{SC de la Regresión} + \text{SC del Error}$$

SC: suma de cuadrados

COEFICIENTE DE DETERMINACION

Los resultados de un ANOVA pueden conducir a la conclusión que los datos observados se ajustan a la línea de regresión estimada; sin embargo se puede tener un modelo estimado que no proporciona una buena explicación del comportamiento de la variable en estudio (Y)

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{\text{SC(Regresión)}}{\text{SC(Total)}}$$

$$0 \leq r^2 \leq 1$$

Cuando

$r^2 \rightarrow 1$: Es un buen modelo: la variabilidad total es explicada por el modelo estimado

$r^2 \rightarrow 0$: No es un buen modelo: la variabilidad total no es explicada por el modelo estimado

r^2 : adecuado en regresión simple

r^2 (ajustado): adecuado en regresión múltiple

$$r^2_{\text{ajustado}} = 1 - \frac{CM(\text{error})}{SC(\text{total})/(n-1)}$$

COEFICIENTE DE CORRELACION LINEAL

Es una medida de asociación lineal entre dos variables aleatorias. Para una muestra de divariada de n -datos: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, el coeficiente de correlación muestral es definido por la siguiente fórmula:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{SP(x, y)}{\sqrt{SC(x) SC(y)}}$$

Propiedades de r

- 1) $-1 \leq r \leq 1$
- 2) No depende de las unidades de las variables en estudio.
- 3) El signo de r es el mismo que b_1

PRUEBA DE HIPOTESIS ACERCA DE ρ

Muchas veces se desea probar la hipótesis acerca de la existencia del coeficiente de correlación poblacional ρ

Hipótesis:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Prueba estadística:

$$T = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)gl$$

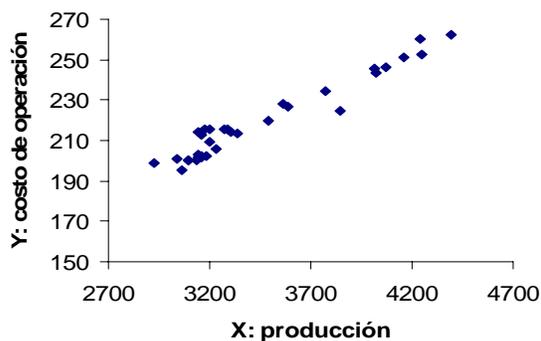
Ejemplo 1

Se consideran los datos mensuales de producción y costos de operación de una empresa británica de transporte de pasajeros por carretera durante los años 1949-52

X : producción, en miles de millas-vehículo recorridas por mes

Y : costo de operación, en miles de dólares por mes.

Usar los datos del archivo: “**regresión1.xls**”



Se puede visualizar la relación lineal entre las variables en estudio.



SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.97285
R Square	0.946436
Adjusted R Square	0.944708
Standard Error	4.625823
Observations	33

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	11720.88	11720.88	547.7496	2.89E-21
Residual	31	663.3453	21.39824		
Total	32	12384.22			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	64.96328	6.635974	9.789562	5.31E-11	51.42912	78.49743
X Variable 1	0.044673	0.001909	23.40405	2.89E-21	0.04078	0.048566

La línea de regresión estimada:

$$\text{COSTOS} = 64.963 + 0.04467 \text{ PRODUCCION}$$

$b_0 = 64.963$ Cuando NO hay producción en un mes determinado, el costo de operación en promedio es 64,963 dólares.

$b_1 = 0.04467$ Cuando la producción se incrementa en mil millas-vehículo recorrido por mes, el costo de operación en promedio se incrementa en 44.67 dólares.

Ejemplo 2

Se consideran los datos de 69 pacientes de los que se conoce su edad y una medición de su tensión sistólica. Si estamos interesados en estudiar la variación en la tensión sistólica en función de la edad del individuo, deberemos considerar como variable respuesta la tensión y como variable predictora la edad.

X: edad

Y: tensión sistólica

Usar los datos del archivo: “**regresión2.xls**”



Ejemplo 3

En 1962 el economista norteamericano Arthur Okun planteó un modelo macroeconómico para explicar las variaciones en la tasa de desempleo. Según este modelo, que se conoce hoy en día como la “ley de Okun,” existe una relación lineal entre el cambio en la tasa de desempleo y la tasa de crecimiento del Producto Interno Bruto (PIB) real. Se consideran los datos sobre desempleo y crecimiento económico en los Estados Unidos durante el período 1966-95.

Usar los datos del archivo: “**regresión3.xls**”

- a) Use estos datos para estimar el modelo de Okun, y explique el significado de los coeficientes obtenidos.
- b) En este problema, el punto donde la recta interseca al eje X tiene un significado económico interesante. Determine este punto para este caso, y explique su significado en términos del modelo de Okun.

Ejemplo 4

Se consideran los datos de un estudio estadístico de los costos administrativos en los bancos comerciales en Guatemala.

Y: Gastos Generales y de Administración, miles de dólares.

X1: Total de activos del banco, miles de dólares.

X2: Número de agencias del banco

Usar los datos del archivo: “**regresión4.xls**”



MUESTREO

Cuando se desea obtener información de los miembros de una población; es decir cuando se desea conocer los parámetros de una población, la primera alternativa es realizar un censo. Hay varias razones por las que a menudo se prefiere un muestreo a un censo.

VENTAJAS DEL METODO DE MUESTREO

Costo reducido.- Si los datos se obtienen únicamente de una pequeña fracción del total, los gastos son menores que los que se realizarían en un censo.

Mayor rapidez.- Los datos pueden ser recolectados y resumidos más rápidamente con una muestra que con un censo.

Mayor exactitud.- Si el volumen de trabajo es reducido se puede emplear personal capacitado al cual se le puede someter a entrenamiento intensivo

Cuidado de la población.- En estudios destructivos, conserva los elementos de la población; como por ejemplo, el estudio del tiempo de duración de baterías.

MUESTREO PROBABILISTICO

Todos los individuos tienen probabilidad conocida de ser elegidos.

Todas las posibles muestras de tamaño n tienen probabilidad conocida de ser elegidas.

Sólo estos métodos nos aseguran *representatividad* de la muestra.

Los tipos de muestreo probabilístico son:

1. Muestreo Aleatorio Simple
2. Muestreo Aleatorio Sistemático
3. Muestreo Aleatorio Estratificado
4. Muestreo Aleatorio por Conglomerados

MUESTREO NO PROBABILISTICO

Aplicado cuando el muestreo probabilístico resulta excesivamente costoso

Todos los individuos **no** tienen la misma probabilidad de ser elegidos.

No se tiene la certeza de que muestra extraída sea representativa

No se puede hacer generalizaciones.

SELECCIÓN ALEATORIA

Una muestra tiene *selección aleatoria* cuando el proceso de selección de unidades se hace por sorteo, ya que de esta manera todas las unidades tienen la misma probabilidad de ser seleccionadas.

Uso de función EXCEL: SAMPLING



MARCO DE MUESTREO

El marco muestral es una representación de todos los elementos de la población objetivo que consta de una lista de características que permitan identificar dicha población.

PARÁMETROS DE UNA POBLACIÓN

Total poblacional: X

$$X = \sum_{i=1}^N X_i$$

Media poblacional: μ

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

Proporción poblacional: P_x

$$P_x = \frac{Y}{N}$$

Donde: N = tamaño de la población, $Y = \sum_{i=1}^N X_i$, donde $X_i = \begin{cases} 1 & \text{presencia de éxito} \\ 0 & \text{fracaso} \end{cases}$

MUESTREO ALEATORIO SIMPLE

Si se tiene que seleccionar una muestra de n elementos de una población de tamaño N . El muestreo aleatorio simple es aquel en el que cada muestra posible de tamaño n tienen la misma probabilidad de ser seleccionada.

Estimación de la media poblacional: μ

Sean x_1, x_2, \dots, x_n los valores observados de una muestra de tamaño n , tomada de una población de tamaño N .

1) Estimación puntual de la media:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$



2) Estimación de la varianza de la media muestral: $var(\bar{x}) = \frac{s^2}{n} \left(\frac{N-n}{N} \right)$

3) Estimación del error estándar de la media muestral: $se(\bar{x}) = \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$

4) Estimación por intervalos de la media: $\bar{x} \pm z_0 \times se(\bar{x})$

Estimación del total de la poblacional: X

Sean x_1, x_2, \dots, x_n los valores observados de una muestra de tamaño n , tomada de una población de tamaño N .

1) Estimación puntual del total: $\hat{X} = N \bar{x}$

2) Estimación por intervalos del total: $N \bar{x} \pm z_0 \times N se(\bar{x})$

Estimación de la proporción poblacional: P

Sean x_1, x_2, \dots, x_n los valores observados (“1” y “0”) de una muestra de tamaño n , tomada de una población de tamaño N .

1) Estimación puntual de la proporción: $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$

2) Estimación de varianza de la proporción muestral: $var(\hat{p}) = \frac{\hat{p} \hat{q}}{n-1} \left(\frac{N-n}{N} \right)$

3) Estimación del error estándar de la proporción muestral: $se(\hat{p}) = \sqrt{var(\hat{p})}$

4) Estimación por intervalos de la media: $\hat{p} \pm z_0 \times se(\hat{p})$



Ejemplo1

Una empresa tiene 189 contables. En una muestra aleatoria de 50 de ellos, el número medio de horas trabajadas en sobretiempo en una semana fue de 9.7 horas con una desviación estándar de 6.2 horas. Halle un intervalo del 95% de confianza para el número medio de horas trabajadas en sobretiempo en una semana.

Ejemplo2

Un auditor, examinando un total de 840 facturas pendientes de cobro, de una empresa, tomó una muestra aleatoria de 120 facturas. Usando los datos del archivo “**muestreo1.xls**”, mediante muestreo aleatorio simple.

- Hallar un intervalo del 95% de confianza para estimar la cantidad total de cobros pendientes
- Hallar un intervalo del 95% de confianza para estimar la proporción de facturas por cobrar con menos de 100 dólares

MUESTREO SISTEMATICO de 1 en k

Si se tiene que seleccionar una muestra de n elementos de una población de tamaño N . El muestreo sistemático de 1 en k , donde $k = N/n$, se realiza de la siguiente manera:

- El primer elemento es seleccionado aleatoriamente entre los primeros k elementos
- Los próximos elementos son seleccionados cada k -elementos.

Ejemplo1

Desde una población de $N = 12$ hogares, se selecciona una muestra de 4 hogares para investigar acerca de la variable “número de personas que viven en el hogar”

hogares	1	2	3	4	5	6	7	8	9	10	11	12
#personas	4	3	5	6	3	4	3	4	7	5	2	1

- Usando el muestreo aleatorio simple, seleccionar los hogares
- Usando el muestreo sistemático de 1 en 3, seleccionar los hogares.

Ejemplo2

Un auditor, examinando un total de 840 facturas pendientes de cobro, de una empresa, tomó una muestra aleatoria de 120 facturas. Usando los datos del archivo “**muestreo1.xls**”, mediante muestreo sistemático de 1 en 7



- 1) Hallar un intervalo del 95% de confianza para estimar la cantidad total de cobros pendientes
- 2) Hallar un intervalo del 95% de confianza para estimar la proporción de facturas por cobrar con menos de 100 dólares

MUESTREO ESTRATIFICADO

Si se tiene que seleccionar una muestra de n elementos de una población de tamaño N , la cual está dividida en k estratos, mutuamente excluyentes de tamaños N_1, N_2, \dots, N_k , tal que:

$$N_1 + N_2 + \dots + N_k = N$$

El muestreo estratificado consiste en seleccionar una muestra desde cada estrato de tamaños n_1, n_2, \dots, n_k , tal que

$$n_1 + n_2 + \dots + n_k = n$$

Estimación de la media poblacional: μ

Sean $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ y $s_1^2, s_2^2, \dots, s_k^2$ las medias y las varianzas muestrales desde cada estrato

1) Estimación puntual de la media:
$$\bar{x}_{str} = \frac{1}{N} \sum_{i=1}^k N_i \bar{x}_i$$

2) Estimación de la varianza de la media muestral:

$$var(\bar{x}_{str}) = \frac{N_1^2 var(\bar{x}_1) + N_2^2 var(\bar{x}_2) + \dots + N_k^2 var(\bar{x}_k)}{N^2}$$

Donde:
$$var(\bar{x}_i) = \frac{s_i^2}{n_i} \left(\frac{N_i - n_i}{N_i} \right) \quad i = 1, 2, \dots, k$$

3) Estimación del error estándar de la media muestral: $se(\bar{x}_{str}) = \sqrt{var(\bar{x}_{str})}$

4) Estimación por intervalos de la media:
$$\bar{x}_{str} \pm z_0 \times se(\bar{x}_{str})$$

Estimación del total de la poblacional: X

Sean $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ y $s_1^2, s_2^2, \dots, s_k^2$ las medias y las varianzas muestrales desde cada estrato

1) Estimación puntual del total: $\hat{X} = N \bar{x}_{str}$

2) Estimación por intervalos del total: $N \bar{x}_{str} \pm z_0 \times N se(\bar{x}_{str})$

Estimación de la proporción poblacional: P

Sean $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_k$ las proporciones muestrales desde cada estrato

1) Estimación puntual de la proporción: $\hat{p}_{str} = \frac{1}{N} \sum_{i=1}^k N_i \hat{p}_i$

2) Estimación de varianza de la proporción muestral:

$$var(\hat{p}_{str}) = \frac{N_1^2 var(\hat{p}_1) + N_2^2 var(\hat{p}_2) + \dots + N_k^2 var(\hat{p}_k)}{N^2}$$

Donde: $var(\hat{p}_i) = \frac{\hat{p}_i \hat{q}_i}{n_i - 1} \left(\frac{N_i - n_i}{N_i} \right) \quad i = 1, 2, \dots, k$

3) Estimación del error estándar de la proporción muestral: $se(\hat{p}_{str}) = \sqrt{var(\hat{p}_{str})}$

4) Estimación por intervalos de la media: $\hat{p}_{str} \pm z_0 \times se(\hat{p}_{str})$

Ejemplo 1:

Una pequeña ciudad contiene un total de 1800 hogares. La ciudad está dividida en tres distritos que contienen 820, 540 y 440 hogares, respectivamente. Una muestra aleatoria estratificada de 310 hogares contiene 120, 100 y 90 hogares, respectivamente de estos tres distritos. Se pide a los miembros de la muestra que calculen su factura total de electricidad consumida en los meses de invierno. Las respectivas medias muestrales son \$290, \$352 y \$427, y las respectivas desviaciones típicas muestrales son \$47, \$61 y \$93.



Districtos	N_i	n_i	promedio	desviación típica
1	820	120	290	47
2	540	100	352	61
3	440	90	427	93

- 1) Hallar un intervalo del 95% de confianza para estimar la media de la factura total de electricidad consumida en los meses de invierno.
- 2) Hallar un intervalo del 95% de confianza para estimar la cantidad total de electricidad consumida en los meses de invierno.

Ejemplo2:

En una ciudad que tiene tres distritos se quiere conocer la proporción de hogares con alguna persona profesional. Se toman muestras aleatorias de esos hogares en cada uno de los tres distritos y se obtienen los resultados que muestra la tabla

Districtos	N_i	n_i	Profesionales	Proporción
1	1200	180	80	0.4444
2	1350	190	50	0.2632
3	1050	140	45	0.3214

Ejemplo3:

Una empresa tiene tres divisiones y los auditores están intentando estimar la cantidad total en facturas pendientes de cobro de la empresa. Hay un total de 870 facturas y en cada división hay 250, 300 y 320 facturas respectivamente. Una muestra aleatoria estratificada de 195 facturas contiene 60, 65 y 70 facturas tomadas desde las tres divisiones respectivamente. Usar los datos del archivo “**muestra2.xls**”

MUESTREO POR CONGLOMERADOS

La población $U = \{1, 2, \dots, N\}$ está dividida en conglomerados C_1, C_2, \dots, C_M los cuales forman las unidades primarias de muestreo, cada uno de estos conglomerados está constituido por elementos de la población, unidades finales.

El muestreo por conglomerados consiste en seleccionar aleatoriamente un cierto número de conglomerados (m), necesario para alcanzar el tamaño muestral establecido,

N = número de elementos en la población

M = número de conglomerados en la población

m = número de conglomerados en la muestra

Estimación del total de la poblacional: X

Sean x_1, x_2, \dots, x_m los totales de la variable en cada conglomerado de la muestra

1) Estimación puntual de la media de conglomerado:
$$\bar{x}_c = \frac{1}{m} \sum_{i=1}^m x_i$$

2) Estimación de la varianza de la media muestral:
$$var(\bar{x}_c) = \frac{s_x^2}{m} \left(\frac{M-m}{M} \right)$$

3) Estimación del error estándar de la media muestral:
$$se(\bar{x}_c) = \sqrt{var(\bar{x}_c)}$$

4) Estimación por intervalos del total poblacional:
$$M \times \bar{x}_c \pm z_0 \times M \times se(\bar{x}_c)$$

Estimación de razón poblacional: R

Sean x_1, x_2, \dots, x_m los totales de la variable x , en cada conglomerado de la muestra

Sean y_1, y_2, \dots, y_m los totales de la variable y , en cada conglomerado de la muestra

1) Estimación puntual de la razón de conglomerado:
$$r_c = \frac{\sum_{i=1}^m x_i}{\sum_{i=1}^m y_i}$$

2) Estimación de la varianza de la razón de conglomerado:

$$var(r_c) = r_c^2 \left\{ \frac{var(\bar{x}_c)}{\bar{x}_c^2} + \frac{var(\bar{y}_c)}{\bar{y}_c^2} - \frac{2}{m} \times \frac{M-m}{M} \times \frac{cov(x, y)}{\bar{x}_c \bar{y}_c} \right\}$$



$$\text{donde: } \text{var}(\bar{x}_c) = \frac{s_x^2}{m} \left(\frac{M-m}{M} \right), \quad \text{var}(\bar{y}_c) = \frac{s_y^2}{m} \left(\frac{M-m}{M} \right)$$

$$\text{cov}(x, y) = \frac{\sum_{i=1}^m (x_i - \bar{x}_c)(y_i - \bar{y}_c)}{m-1}$$

3) Estimación del error estándar de la razón de conglomerado: $se(r_c) = \sqrt{\text{var}(r_c)}$

4) Estimación por intervalos de la razón poblacional: $r_c \pm z_0 \times se(r_c)$

Ejemplo

Una ciudad está dividida en 30 distritos escolares con cuatro escuelas elementales en cada una. Mediante muestreo por conglomerados se seleccionaron al azar 3 distritos escolares. Construya un intervalo del 95% de confianza para el total de niños con daltonismo en la ciudad y para la proporción de niños con esta condición.

Distrito Escolar seleccionado	Escuela del distrito escolar	Total de niños en la escuela	Número de niños daltónicos por escuela
1	1	130	2
	2	150	3
	3	160	3
	4	120	5
2	1	110	2
	2	120	4
	3	100	0
	4	120	1
3	1	89	4
	2	130	2
	3	100	0
	4	150	2

Conglomerado: distrito escolar

Unidad elemental: escuela

Variables: total de niños en la escuela y total de niños daltónicos

Los datos de la tabla anterior se pueden resumir en la siguiente tabla

Distrito Escolar seleccionado	Total de niños en el distrito escolar (y)	Número de niños daltónicos por distrito escolar (x)
1	560	13
2	450	7
3	469	8
	1479	28

Estimación del total de niños con daltonismo

- 1) Cálculo del promedio muestral, en conglomerado: $\bar{x}_c = 28 / 3 = 9.3333$
- 2) Cálculo de la varianza de la media muestral: $var(\bar{x}_c) = \frac{10.3333}{3} \left(\frac{30-3}{30} \right) = 3.1$
- 3) Cálculo del error estándar de la media muestral: $se(\bar{x}_c) = 1.760682$
- 4) Estimación por intervalos del total poblacional: $30 \times 9.3333 \pm 1.96 \times 30 \times 1.760682$

$$LC(X) = (176.47, 383.53)$$

Estimación de la proporción de niños con daltonismo: Uso de estimador de razón

- 1) Cálculo de la razón de conglomerado: $r_c = \frac{\sum_{i=1}^m x_i}{\sum_{i=1}^m y_i} = \frac{28}{1479} = 0.01893$
- 2) Cálculo de la varianza de la razón de conglomerado:

$$var(\bar{x}_c) = \frac{s_x^2}{m} \left(\frac{M-m}{M} \right) = \frac{10.3333}{3} \left(\frac{30-3}{30} \right) = 3.1 \quad \bar{x}_c = 9.3333$$

$$var(\bar{y}_c) = \frac{s_y^2}{m} \left(\frac{M-m}{M} \right) = \frac{3457}{3} \left(\frac{30-3}{30} \right) = 1037.1 \quad \bar{y}_c = 493$$

$$cov(x, y) = 189$$



$$\text{var}(r_c) = r_c^2 \left\{ \frac{\text{var}(\bar{x}_c)}{\bar{x}_c^2} + \frac{\text{var}(\bar{y}_c)}{\bar{y}_c^2} - \frac{2}{m} \times \frac{M-m}{M} \times \frac{\text{cov}(x, y)}{\bar{x}_c \bar{y}_c} \right\}$$

$$\text{var}(r_c) = 0.01893^2 \left\{ \frac{3.1}{9.3333^2} + \frac{1037.1}{493^2} - \frac{2}{3} \times \frac{30-3}{30} \times \frac{189}{9.33 \times 493} \right\} = 5.450036e-06$$

3) Cálculo del error estándar de la razón de conglomerado: $se(r_c) = 0.002334531$

4) Estimación por intervalos de la razón poblacional: $r_c \pm z_0 \times se(r_c)$

$$LC(R) = (0.0023345, 0.023505)$$



BIBLIOGRAFIA

Berenson, M. L., Levine, D. M., Krehbiel, T. C. (2008) “Basic Business Statistics”, Eleventh Edition, Pearson Prentice Hall.

Black, K., (2008) “Business Statistics”, 5th Edition, Wiley.

Cochran, W. G., (1977) “Sampling Techniques”, Thirds Edition, Wiley, Ney York.

Levy P. S., Lemeshow S. (1999), “Sampling of Populations, Methods and Applications”, Thirds Edition, John Wiley & Sons, Inc.

Lind, D., Marchal, W. G., Wathen, S. A. (2008) “Estadística Aplicada a los negocios y a la Economía”, Decimotercera Edición, McGraw-Hill, Mexico D. F.

Newbold, P., Carlson, W., Thorne, B. (2008) “Estadística para Administración y Economía”, Sexta Edición, Pearson Educación, S. A. Madrid, España.