No.Lista: 07

Universidad Nacional Autonoma de México
Facultad de Ingeniería

Cuaderno

Celaya González David Alejandro
Grupo: 02
Estadistica
22/Enero/2021

ESTADISTICA:

EVALUACIÓN

Asistencia 80%

EXAMENES	60%	3 EXAMENES DURANTE EL SEMESTRE
TAREAS	30%	TAREAS SEMANALES
PARTICIPACIÓN	10%	TAREAS A MANO
CUADERNO	5%	

DISTRIBUCIONES F-FISHER, NORMAL, T-STUDENT, JI-CUADRADA.
PARAMETROS MEDIA, VARIANZA, PROPORCIÓN
ESTADISTICA DESCRIPTIVA.
LA FORMA CIENTIFICA DE VALIDAR LA INVESTIGACIÓN ES A TRAVÉS DE LA PROBABILIDAD Y LA ESTADISTICA.
INVESTIGAR VIENE DEL VOCABLO INVESTIGARE → HACER DILIGENCIAS PARA DESCUBRIR ALGO.

INVESTIGACION BASICA: TIENE POR OBJETO MISMO LA INVESTIGACIÓN, ES DECIR, SÓLO PERSIGUE EL CONOCIMIENTO DE LAS COSAS.
METODO CIENTIFICO: LOGRAR UN PROPOSITO, DEBEN SEGUIR UNA METODOLOGÍA.
ESTADISTICA: SIGNIFICA CIENCIA DEL ESTADO.
LA ESTADISTICA NO SOLO PROPORCIONA INFORMACIÓN O DATOS; SINO QUE LOS AGRUPA, ANALIZA, INTERPRETA Y PERMITE GENERAR INFERENCIAS O CONCLUSIONES DE UNA POBLACIÓN A PARTIR DE DATOS DE UNA MUESTRA.
ESTADISTICA COMO DEFINICIÓN: RAMA DE LAS MATEMÁTICAS QUE SE ENCARGA DE LA SELECCIÓN DE DATOS, SU ORGANIZACIÓN, PRECENTACIÓN Y REALIZACIÓN DE LAS CONCLUSIONES QUE SE PUEDEN OBTENER DE DICHOS DATOS.
SE CLASIFICA.
- UNIVARIABLE Y MULTIVARIABLE (NÚMERO DE VARIABLES)
- DESCRIPTIVA INFERENCIAL (APLICACIÓN)
- PARAMETRICA Y NO PARAMETRICA (BASADA EN LA INFORMACIÓN QUE POSEE)

POBLACIÓN: CONJUNTO DE INDIVIDUOS QUE SE ESTUDIAN CON BASE EN UNA CARACTERISTICA
MUESTRA: ES UN SUBCONJUNTO DE LA POBLACION.
ETAPAS DE LA INFORMACIÓN QUE SE DEBE REALIZAR EN LA ESTADÍSTICA.
- 01- REUNIR
- 02- ORGANIZAR
- 03- ANALIZAR

PARA REUNIR LA INFORMACIÓN SE REQUIERE DEFINIR AL CARACTER QUE VAMOS A ESTADIAR.
ORDENAR → GERARQUIZAR. ORGANIZAR → CUALITATIVOS
(NO PUEDO CAMBIAR EL ORDEN)

CELAYA GONZÁLEZ DAVID A.

28-SEP-20

ESCALAS DE MEDICIÓN

CUANDO LA VARIABLE SOLO SE LE PUEDE CLASIFICAR POR UN ATRIBUTO.

USAREMOS

- NOMINAL → NO JERARQUICO
- ORDINAL. → SI SE PUEDE HACER COMPARACION DE UN ATRIBUTO

} CUAL.

- INTERVALOS → EL CERO NO ES ABSOLUTO (TEMPERATURA)
- RAZÓN O PROPORCIÓN → CERO ABSOLUTO.

} CUAN.

MUESTREO: SE TOMA UNA MUESTRA PARA ESTUDIAR LA POBLACION.

TECNICAS DE MUESTREO.

ALEATORIO SIMPLE: ES CUANDO LOS ELEMENTOS DE UNA POBLACIÓN TIENEN LA MISMA PROBABILIDAD DE SER SELECCIONADOS. SE REQUIEREN 4 COSAS.

- NUMERAR A CADA INDIVIDUO DE LA POBLACIÓN
- CUANTIFICAR EL TAMAÑO DE LA MUESTRA (n = MUESTRA) (N = POBLACIÓN)
- ELEGIR CON UN GENERADOR DE NÚMEROS ALEATORIOS.
- LA MUESTRA SERÁN LAS 'n' NÚMEROS QUE SE SELECCIONARÓN EN EL PASO ANTERIOR

OJO: EN LA PRÁCTICA ES DIFICIL OCUPAR ESTA TECNICA.

ESTRATIFICADO: CUANDO A LA POBLACIÓN LA DIVIDES EN ESTRATOS Y SE REQUIERE QUE EN LA MUESTRA APARESCA UNO DE CADA ESTRATO

CONGLOMERADOS: SIMILAR AL ESTRATIFICADO PERO PARA POBLACIONES MUY GRANDES. NO SE ESTUDIAN TODOS LOS ESTRATOS.

SISTEMATICO: POR ALGÚN MÉTODO SISTEMATICO SE PEDE GENERAR UNA MUESTRA (TIEMPO O INDIVIDUOS).

ESTADISTICA DESCRIPTIVA

ES LA PARTE DE LA ESTADÍSTICA QUE TIENE COMO PROPOSITO ORGANIZAR Y PRESENTAR LOS DATOS DE UNA POBLACIÓN O DE UNA MUESTRA PARA SU ANÁLISIS E INTERPRETACIÓN.

TRES TECNICAS.

- DISTRIBUCION DE FRECUENCIAS: FORMA EN QUE SE AGRUPAN LOS DATOS (CUALITATIVOS Y CUANTITATIVOS)
- GRAFICAS: VIZUALIZAR LA FORMA EN QUE SE AGRUPAN LOS DATOS (CUALITATIVOS)
- MEDIDAS NÚMERICAS: RESUMEN DE LOS DATOS EN FORMA CUANTITATIVA. (CUANTITATIVOS).

EXISTE UNA GRAN VARIEDAD DE TABLAS DE DISTRIBUCION DE FRECUENCIAS; SIN EMBARGO, AQUÍ SE ESTUDIARÁ UNA TABLA TEORICA COMPLETA. PARA RESUMIR LOS DATOS, SE UTILIZAN INTERVALOS, CLASES O CATEGORIAS Y POSTERIORMENTE SE INDICAN LA FRECUENCIA DE C/U DE ELLOS.

LAS COLUMNAS QUE FORMAN UNA TABLA COMPLETA SON:

CLASES

- LIMITES DE CLASE (L_i)
- FRONTERAS DE CLASE: LAS FRONTERAS O LIMITES VERDADEROS DE UNA CLASE, SON LOS PUNTOS MEDIOS ENTRE LOS LÍMITES DE INTERVALOS CONSECUTIVOS. (L_{Ri})

LIMITES DE CLASE
1 - 3
4 - 6
7 - 9

FRONTERAS DE CLASE
0.5 (1-0.5) - 3.5 (3+0.5)
3.5 (4-0.5) - 6.5 (6+0.5)
6.5 (7-0.5) - 9.5 (9+0.5)

- MARCAS DE CLASE (x_i): ES EL PUNTO MEDIO DE UNA CLASE. SE CONSIDERA COMO EL VALOR REPRESENTATIVO DE UN INTERVALO. LAS MARCAS DE CLASE SE OBTIENEN PROMEDIANDO LOS LIMITES DE UN INTERVALO, O BIEN, LAS FRONTERAS.

MARCA DE CLASE
2
5
8

FRECUENCIA:

- FRECUENCIA: ES EL NÚMERO DE ELEMENTOS EN LA MUESTRA O EN LA POBLACIÓN QUE PERTENECE A LA CLASE EN CUESTIÓN.
  SI LOS DATOS DE UNA MUESTRA SON: 1, 9, 5, 8, 4, 1, 2, 7, 6, 3, 3, 2, 7, 9
  ¿ENTONCES COMO QUEDARÍAN AGRUPADOS EN CLASES? (f_i)

FRECUENCIA
6
3
5

CELAYA

○ FRECUENCIA ACUMULADO: ES EL NÚMERO DE DATOS EN LA MUESTRA O POBLACIÓN QUE SON MENORES O IGUALES QUE EL LIMITE SUPERIOR DEL INTERVALO EN CUESTIÓN. SE OBTIENE SUMANDO LA FRECUENCIA DEL INTERVALO ACTUAL Y DE LOS ANTERIORES INTERVALOS. (F_i)

FRECUENCIA ACUMULADA
6
6 + 3 = 9
6 + 3 + 5 = 14

TIENE QUE SER IGUAL AL TOTAL EN FRECUENCIAS.

○ FRECUENCIA RELATIVA: ES LA PROPORCIÓN DE DATOS QUE PERTENECEN A LA CLASE EN CUESTIÓN. Es el NÚMERO O COCIENTE DE LA FRECUENCIA ENTRE EL NÚMERO TOTAL DE DATOS (f_i^*)

$$\boxed{f_i^* = \frac{f_i}{n}}$$

PARA LA TABLA DEL EJEMPLO, SI EL TOTAL DE DATOS ES $n = 100$, ENTONCES:

FRECUENCIA RELATIVA
6/100 = 0.06
3/100 = 0.03
5/100 = 0.05

○ FRECUENCIA ACUMULADA RELATIVA:

02/10/2020

GRAFICAS:

CUANDO SE DESEA UN MAYOR IMPACTO DE LA FORMA EN LA QUE SE DISTRIBUYEN LOS DATOS, ÉSTOS SE PRESTAN EN UNA O VARIAS GRÁFICAS.

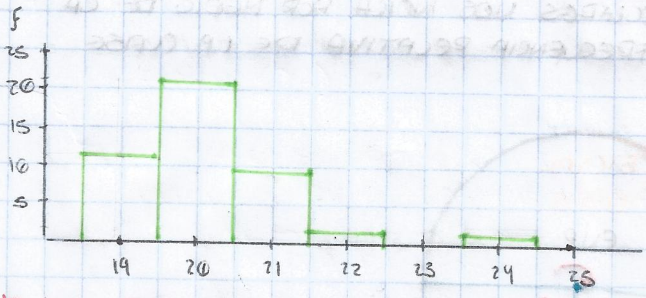
SON MUCHAS LAS GRAFICAS QUE SE PUEDEN UTILIZAR EN LA ESTADISTICA DESCRIPTIVA, DESTACANDO:

- EL HISTOGRAMA
- EL POLÍGONO
- LA OJIVA
- SECTORES CIRCULARES
- LA DE TALLOS Y HOJAS
- EL DIAGRAMA DE CAJA.

HISTOGRAMA: CUANTI - CUALI

EL HISTOGRAMA ES UNA GRÁFICA DE BARRAS RECTANGULARES CUYAS BASES ESTÁN CENTRADAS EN LA MARCA DE CLASE DEL INTERVALO, Y SUS ÁREAS PROPORCIONALES A LA FRECUENCIA DEL INTERVALO.

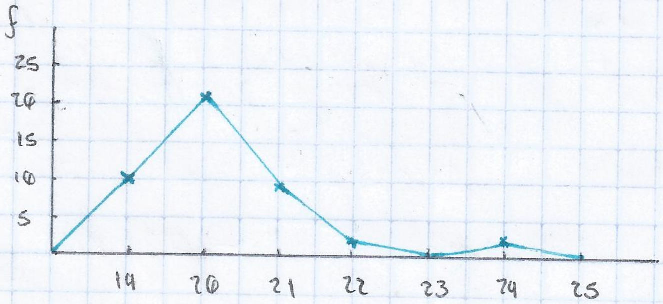
- GRAFICA DE BARRAS CONVENCIONAL ES PARA DATOS CUALITATIVOS.
- HISTOGRAMA ES PARA DATOS CUANTITATIVOS.



DEJAR ESPACIO EN EL EJE 'Y'

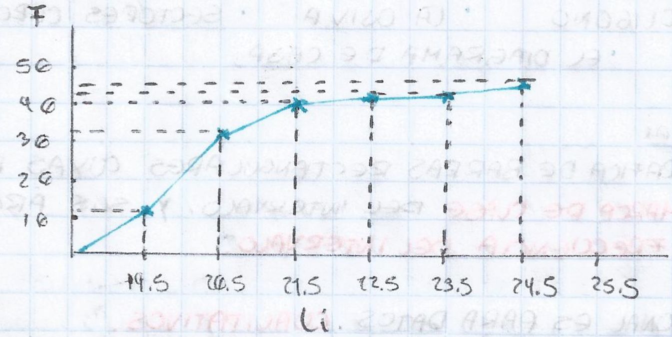
POLIGONO DE FRECUENCIA CUANTI - CUALI

EL POLIGONO DE FRECUENCIAS ES UNA GRÁFICA POLIGONAL O DE LINEAS RECTAS QUE INDICA PARA CADA MARCA DE CLASE LA FRECUENCIA. SE OBTIENE UNIENDO LOS PUNTOS MEDIOS DE LAS PARTES SUPERIORES DE LAS BARRAS DEL HISTOGRAMA.



CECAYA

OJIVA: (NO SE OCUPA PARA DATOS CUALITATIVOS)
LA OJIVA ES TAMBIÉN UNA GRÁFICA POLIGONAL, PERO SE DIBUJA UTILIZANDO LAS FRONTERAS CONTRA LAS FRECUENCIAS ACUMULADAS (O ACUMULADAS RELATIVAS). LA OJIVA INDICA, PARA CADA FRONTERA, LOS ELEMENTOS (O PROPORCIÓN DE ELEMENTOS), QUE SON MENORES O IGUALES QUE DICHA FRONTERA.



SECTORES CIRCULARES (PAY):
LA GRAFICA DE SECTORES CIRCULARES NOS INDICA POR MEDIO DE LA FRACCIÓN DE UN CIRCULO LA FRECUENCIA RELATIVA DE LA CLASE EN CUESTIÓN.

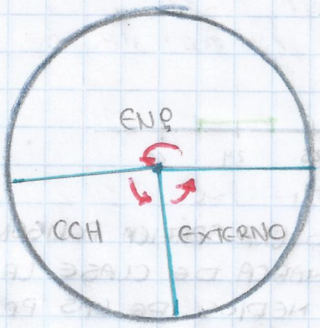


DIAGRAMA DE TALLO Y HOJAS: ES UNA REPRESENTACIÓN SEMI-GRÁFICA, UTILIZADA PARA DATOS CUANTITATIVOS, UTILIZADA AMPLIAMENTE EN LA DÉCADA DE LOS OCHENTA, CUANDO LAS COMPUTADORAS NO REALIZABAN GRÁFICAS MÁS SOFISTICADAS COMO LOS HISTOGRAMAS. CONSISTE EN SEPARAR LOS NÚMEROS EN DOS PARTES, POR EJEMPLO DECENAS Y UNIDADES.

DATOS: 20, 25, 40, 36, 72, 25, 25

D	U
2	0555
3	6
4	0
5	
6	
7	2
8	

DIAGRAMA DE CAJA: ES UNA FORMA DE PRESENTACIÓN ESTADÍSTICA DESTINADA FUNDAMENTALMENTE A RESALTAR ASPECTOS DE LA DISTRIBUCIÓN DE LAS OBSERVACIONES EN UNA O MÁS SERIES DE DATOS CUANTITATIVOS. REMPLAZA, EN CONSECUENCIA, AL HISTOGRAMA Y A LA CURVA DE DISTRIBUCIÓN DE FRECUENCIAS SOBRE LOS QUE TIENE VENTAJAS A LA INFORMACIÓN QUE BRINDA Y A LA APRECIACIÓN GLOBAL QUE SURGE DE LA LECTURA.

NOTA: PARA REALIZAR EL DIAGRAMA DE CAJA PRIMERO DEBEMOS CONOCER LAS MEDIDAS NUMÉRICAS.

CELAYA

DISTRIBUCIÓN DE FRECUENCIAS POR INTERVALOS: EXISTEN OTRAS TABLAS DE DISTRIBUCIÓN DE FRECUENCIAS, BASADAS EN INTERVALOS, EN DONDE SE UTILIZA LA NOTACIÓN DEL CÁLCULO PARA LOS INTERVALOS ABIERTOS Y CERRADOS POR EJEMPLO:

INTERVALO	MARCA DE CLASE x_i	FRECUENCIA f_i
[1, 4)	2.5	2
[4, 7)	5.5	5
[7, 10)	8.5	8
⋮	⋮	⋮

VALORES QUE AYUDAN A CONSTRUIR UNA TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

LONGITUD DE CLASE. ES LA DIFERENCIA ENTRE LA FRONTERA SUPERIOR Y LA INFERIOR DE UNA MISMA CLASE, SE DENOTA POR C

LA LONGITUD DE CLASE DE LA TABLA ANTERIOR ES: 3

RECOMENDACIONES PARA LA CONSTRUCCIÓN DE UNA TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

- EL NÚMERO DE CLASES ESTARÁ ENTRE 5 Y 20, INCLUSIVE. LA PRIMERA APROXIMACIÓN DEL NÚMERO DE CLASES SE OBTENDRÁ CON $\sqrt{n}$

- TODAS LAS CLASES SERÁN DE LA MISMA LONGITUD (C)

- LA LONGITUD DE CLASE SE APROXIMA MEDIANTE $C = \frac{\text{RANGO}}{\text{\# DE CLASES}}$, DONDE EL RANGO ES: $V_{MAXIMO} - V_{MINIMO}$.

- POSTERIORMENTE SE AJUSTA DE MANERA CONVENIENTE, DE FORMA QUE EL PRIMER LÍMITE INFERIOR SEA LIGERAMENTE MENOR O IGUAL QUE EL MENOR VALOR, Y EL ÚLTIMO LÍMITE SUPERIOR SEA LIGERAMENTE MAYOR O IGUAL QUE EL MAYOR DATO.

- TRATARÁ DE EVITARSE QUE HAYA CLASES CON FRECUENCIA CERO.

EJEMPLO.

LOS SIGUIENTES VALORES REPRESENTAN EL TIEMPO DIARIO DE TRASPORTE DE UNA MUESTRA DE 50 TRABAJADORES DE CIERTO HOSPITAL EN LA CDMX

69	56	73	66	64	44	36	69	76	53
79	72	82	77	71	48	49	49	60	67
73	70	64	56	31	62	56	55	51	45
30	40	80	49	59	60	76	67	30	72
45	43	77	49	46	42	63	41	64	79

ORDENAR LOS DATOS

DIAGRAMA DE TALLO Y HOJAS

1		
2		
3	0 0 1 6	4
4	0 1 2 3 4 5 5 6 8 9 9 9 9	13
5	1 3 5 6 6 6 9	7
6	0 0 2 3 4 4 4 6 7 7 9 9	12
7	0 1 2 2 3 3 6 6 7 7 7 9	12
8	0 2	2
9		

Suma 50 ✓

ORDENADOS

30	30	31	36	40	41	42	43	44	45
45	46	48	49	49	49	49	51	53	55
56	56	56	59	60	60	62	63	64	64
64	66	67	67	69	69	70	71	72	72
73	73	76	76	77	77	79	79	80	82

CELAYA

Limite de clase	FRONTERA DE CLASE	MARCAS DE CLASE	FRECUENCIA	F. ACUMULADA	F. RELATIVA
29 - 36	28.5 - 36.5	32.5	4	4	0.08
37 - 44	36.5 - 44.5	40.5	5	9	0.1
45 - 52	44.5 - 52.5	48.5	9	18	0.18
53 - 60	52.5 - 60.5	56.5	8	26	0.16
61 - 68	60.5 - 68.5	64.5	8	34	0.16
69 - 76	68.5 - 76.5	72.5	10	44	0.2
77 - 84	76.5 - 84.5	80.5	6	50	0.12
			TOTAL = 50		TOTAL 1

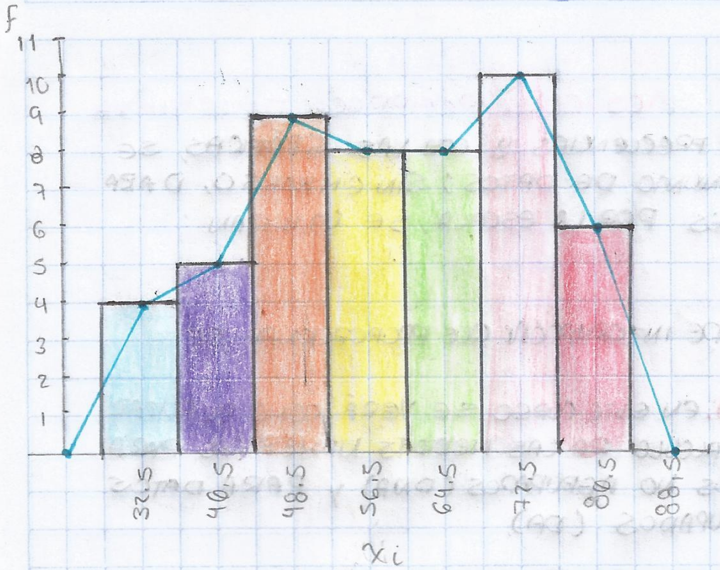
F. A. RELATIVA
0.08
0.18
0.36
0.52
0.68
0.88
1

- EN EL NÚMERO DE CLASES VALOR HACIA ABAJO $\sqrt{50} = 7.07 = 7$ CLASES.
- RANGO = 82-30 => $c = \frac{52}{7} = 7.42$ => 8

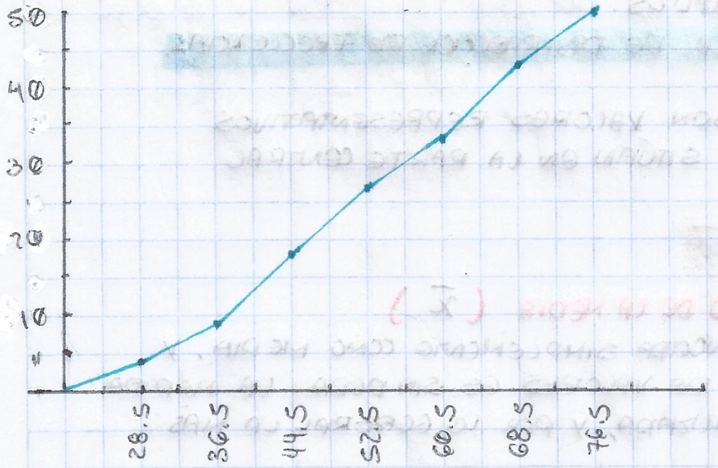
LA LONGITUD DE CLASE * # DE CLASES > RANGO.

OJO: LA LONGITUD DE CLASE SERÁ LA DISTANCIA ENTRE LAS FRONTERAS NO ENTRE LOS LIMITES

HISTOGRAMA
Y
POLIGONO DE FRECUENCIA.



OJIVA



CELAYA

MEDIDAS NÚMERICAS:

NOTA: CON LA TABLA DE DISTRIBUCION DE FRECUENCIAS Y CON LAS GRÁFICAS, SE DESCRIBE EL COMPORTAMIENTO DE UN CONJUNTO DE DATOS; SIN EMBARGO, PARA NO CAER EN SUBJETIVIDADES, O ERRORES POR LA ESCALA, SE UTILIZAN TAMBIEN LAS MEDIDAS NUMÉRICAS.

LAS MEDIDAS NÚMERICAS, POR EL TIPO DE INFORMACIÓN QUE PROPORCIONAN SE CLASIFICAN EN:

- MEDIDAS DE TENDENCIA CENTRAL.
- MEDIDAS DE DISPERSION
- MEDIDAS DE FORMA.

NOTA: EN ESTE CURSO SE VERÁ COMO REALIZAR EL CALCULO DE LAS MEDIDAS NUMÉRICAS PARA DATOS NO AGRUPADOS (DNA) Y PARA DATOS AGRUPADOS (DA).

- NO AGRUPADOS → PUEDEN ESTAR ORDENADOS.
- AGRUPADOS → YA ESTAN EN LA TABLA DE DISTRIBUCCION DE FRECUENCIAS.

MEDIDAS DE TENDENCIA CENTRAL: SON VALORES REPRESENTATIVOS DE UN CONJUNTO DE DATOS, QUE SE SITÚAN EN LA PARTE CENTRAL DE LOS MISMOS.

MEDIA

MEDIA ARITMETICA O ESTADISTICO DE LA MEDIA ($\bar{x}$)

LA MEDIA ARITMÉTICA ES MÁS CONOCIDA SIMPLEMENTE COMO MEDIA, Y ES EL PROMEDIO DE UN CONJUNTO DE VALORES. ES SIN DUDA LA MEDIDA DE TENDENCIA CENTRAL MÁS UTILIZADA, Y POR LO GENERAL LA MÁS REPRESENTATIVA.

$$\bar{x} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} x_i & \text{; Datos no agrupados} \\ \frac{1}{n}\sum_{i=1}^{m} x_i f_i & \text{; Datos agrupados} \end{cases}$$

x_i → Marca clase
f_i → frecuencia

m = clases o intervalos
n = elementos

EJEMPLO DATOS NO AGRUPADOS

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{50}(\underbrace{30+31+36+40}_{\text{EXPRESAR PRIMEROS CINCO}}+\ldots+77+\underbrace{77+79+79+80+82}_{\text{ULTIMOS CINCO}})$$

$\bar{X} = 58.7$

PARA DATOS AGRUPADOS:

$$\bar{X} = \frac{1}{50}[32.5(4)+40.5(5)+48.5(9)+56.5(8)+64.5(8)+72.5(10)+80.5(6)]$$

$\bar{X} = 58.9$

MEDIA GEOMETRICA. (G) (LO USAN LOS PSICOLOGOS)

LA MEDIA GEOMÉTRICA DE UN CONJUNTO DE VALORES POSITIVOS SE CALCULA CON LA RAIZ n-ÉSIMA DEL PRODUCTO DE LAS n OBSERVACIONES.

$$G = \begin{cases} \sqrt[n]{X_1 X_2 \ldots X_n} & \text{; DATOS NO AGRUPADOS} \\ \sqrt[n]{X_1^{f_1}\ X_2^{f_2} \ldots X_m^{f_m}} & \text{; DATOS AGRUPADOS.} \end{cases}$$

DNA $G = \sqrt[50]{30 \cdot 30 \cdot 31 \cdot 36 \cdot 41 \ldots 77 \cdot 79 \cdot 79 \cdot 80 \cdot 82}$

$G = \sqrt[50]{30.5^4 \cdot 40.5^5 \cdot 48.5^9 \cdot 56.5^8 \cdot 64.5^8 \cdot 72.5^{10} \cdot 80.5^6}$

(INCISO 1)

CECAYA

MEDIA ARMONICA. (H)

LA MEDIA ARMONICA DE UN CONJUNTO DE DATOS, ES EL RECIPROCO DE LA MEDIA ARITMETICA DE LOS RECIPROCOS DE CADA UNO DE LOS VALORES.

$$H = \begin{cases} \dfrac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \dfrac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \text{ ; DATOS NO AGRUPADOS} \\ \dfrac{1}{\frac{1}{n}\sum_{i=1}^{m}\frac{f_i}{x_i}} = \dfrac{n}{\sum_{i=1}^{m}\frac{f_i}{x_i}} \text{ ; DATOS AGRUPADOS} \end{cases}$$

MEDIA PONDERADA (MP)

ES UNA MEDIDA DE TENDENCIA CENTRAL, QUE ES APROPIADA CUANDO EN UN CONJUNTO DE DATOS CADA UNO DE ELLOS TIENE UNA IMPORTANCIA RELATIVA (O PESO) RESPECTO DE LOS DEMÁS DATOS

$$MP = \frac{\sum_{i=1}^{n} x_i W_i}{\sum_{i=1}^{n} W_i}$$

EJEMPLO OBTENER EL PROMEDIO DE JUAN PEREZ EN LA ASIGNATURA DE ESTADISTICA SI OBTUVO 6 EN EL EXAMEN, 7 EN TAREAS Y 8 EN PARTICIPACIÓN.

60% ; 30% ; 10%

$$MP = \frac{6(60) + 7(30) + 5(10)}{60 + 30 + 10} = \frac{620}{100} = 6.2$$

MEDIANA ($\tilde{x}$): LA MEDIANA DE UN CONJUNTO DE DATOS ORDENADOS, ES EL VALOR QUE DIVIDE AL CONJUNTO DE DATOS EN DOS CONJUNTOS DE IGUAL TAMAÑO, O BIEN ES EL PROMEDIO DE LOS DOS VALORES CENTRALES.
CUANDO LOS DATOS NO ESTAN AGRUPADOS, SE DEBE ORDENAR EN FORMA ASENDENTE O DESCENDENTE Y SELECCIONAR EL VALOR CENTRAL. SI LOS DATOS SON PARES, ENTONCES SE TOMA EL PROMEDIO DE LOS VALORES CENTRALES; SI LOS DATOS SON IMPARES ENTONCES SE TOMA EL VALOR CENTRAL.
TOMANDO VALORES DE TABLAS.

$$\frac{50}{2} = 25$$

$\tilde{x}_1$ = POSICIÓN
$\tilde{x}$ = MEDIANA

$$\tilde{x} = (60+60)/2 = 60$$

30, 31, 32, 36, 40, 41, 42, 43, 44, 45
45, 46, 48, 49, 49, 49, 49, 51, 53, 55
56, 56, 56, 59, 60, 60, 62, 63, 64, 64
64, 66, 67, 67, 69, 69, 70, 71, 72, 72
73, 73, 76, 76, 77, 77, 79, 79, 80, 82

CUANDO LOS DATOS ESTAN AGRUPADOS, ENTONCES SE REALIZA UNA INTERPOLACIÓN LINEAL UTILIZANDO LAS FRONTERAS Y LA FRECUENCIA ACUMULADA (ES DECIR LOS DATOS DE LA OJIVA), PARA ENCONTRAR EL VALOR EN EL CUAL LA FRECUENCIA ACUMULADA ES DE $n/2$

$$y = \frac{(x-x_1)}{(x_2-x_1)}(y_2-y_1) + y_1$$

	FRONTERA		FRECUENCIA A.
y_1	52.5	x_1	18
y	$\tilde{x}$	x	25 → $n/2 = 50/2$
y_2	60.5	x_2	26

$$y = 7 + 52.5$$
$$y = 59.5$$

MODA (x_{mo}): LA MODA DE UN CONJUNTO DE DATOS ES EL VALOR QUE SE REPITE CON MAYOR FRECUENCIA.

PARA DATOS SIN AGRUPAR, SE DEBE CONTAR LAS REPETICIONES QUE PUEDAN EXISTIR Y EL QUE SE REPITA MAYOR NÚMERO DE VECES SERÁ LA MODA. SI TODOS LOS DATOS APARECEN EL MISMO NÚMERO DE VECES, ENTONCES SE DICE QUE NO EXISTE MODA. (AMODAL)

$$x_o = 49$$

PARA DATOS AGRUPADOS, LA MODA SE APROXIMA CON LA MARCA DE CLASE DEL INTERVALO CON MAYOR FRECUENCIA, O BIEN, UTILIZANDO LA FORMULA

$$x_{mo} = \text{FRONTERA INFERIOR} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) c$$

CELAYA

DONDE:
FRONTERA INFERIOR Δ ES LA FRONTERA INFERIOR DEL INTERVALO CON MAYOR FRECUENCIA.

Δ_1 = ES EL EXCESO DE LA FRECUENCIA MODAL SOBRE LA FRECUENCIA DE LA CLASE INMEDIATA ANTERIOR.

Δ_2 = ES EL EXCESO DE LA FRECUENCIA MODAL SOBRE LA FRECUENCIA DE LA CLASE INMEDIATA POSTERIOR

C = LONGITUD DE CLASE

$$\bar{x}_{mo} = 68.5 + \left(\frac{(10-8)}{(10-8)+(10-6)}\right)(8) = 71.16$$

L_{Ri}	f
	4
	5
	9
	8
	8
68.5	10
	6

- NOS IMPORTA LAS FRECUENCIAS DE LAS CLASES.
- SI HAY INTERVALOS CON FRECUENCIA IGUAL ENTONCES TOMAMOS LA PRIMERA.

¿ES LA MEDIA ARITMETICA LA MEDIDA QUE MEJOR REPRESENTA A LOS DATOS SIEMPRE? R. POR LO GENERAL, PERO NO SIEMPRE

EMPLEADO	SALARIO MENSUAL	
1	$ 8,000	2
2	$ 12,700	4
3	$ 6,500	1
4	$ 92,000	5 (DATO ATIPICO)
5	$ 9,500	3

$\bar{X}$ = 25,740

X_{mo} = NO HAY

$\tilde{X}$ = 9,500

CUANTILES: ASÍ COMO LA MEDIANA ES EL VALOR QUE DIVIDE A UN CONJUNTO DE DATOS ORDENADOS DE IGUAL TAMAÑO EN DOS CONJUNTOS, LOS DATOS PUEDEN DIVIDIRSE EN 4 CONJUNTOS DE IGUAL TAMAÑO (CUARTILES), EN 10 CONJUNTOS DE IGUAL TAMAÑO (DECILES) Y EN 100 CONJUNTOS DE IGUAL TAMAÑO (PERCILES)

SIEMPRE ORDENAR DE MENOR A MAYOR LOS DATOS.

NOS SERVIRA PARA DISCRIMINAR DATOS O NO.

	n número par	n número impar	
CUARTILES	$Q_K = \frac{kn}{4}$	$Q_K = \frac{k(n+1)}{4}$	$K = 1, 2, 3$
DECILES	$D_K = \frac{kn}{10}$	$D_K = \frac{k(n+1)}{10}$	$k = 1, \ldots, 9$
PERCENTILES	$P_K = \frac{kn}{100}$	$P_K = \frac{k(n+1)}{100}$	$K = 1, \ldots, 99$

OJO: EL VALOR CALCULADO CON ESTA FORMULA NOS INDICA LA POSICIÓN DEL CUANTIL, NO EL VALOR DEL MISMO.

OBTENER LA POSICIÓN DEL VALOR ENTERO MÁS LA POSICIÓN DEL VALOR ENTERO MAS UNO, MENOS LA POSICIÓN ENTERA.

$$\text{CUANTIL} = \underset{\text{dato}}{\tilde{x}_{\text{ENTERA}}} + \underset{\text{POSICION}}{(\tilde{x}_{\text{DECIMAL}})}\,(\underset{\text{DATO}}{x_{\text{ENTERO}+1}} - \underset{\text{DATO}}{x_{\text{ENTERO}}})$$

$$P_{97} = 50.44 = 82 + (0.44)(84 - 82)$$

$$P_{97} = 82.88$$

MEDIDAS DE DISPERSIÓN

CELAYA

LAS MEDIDAS DE DISPERSIÓN PROPORSIONAN UN INDICADOR DE ALEJAMIENTO DE LOS DATOS. TAMBIEN SE LES LLAMA MEDIDAS DE VARIACIÓN.

RANGO, DESVIACIÓN MEDIA, DESVIACION MEDIANA, VARIANZA, DESVIACION ESTANDAR, RAGO INTERCUARTILICO.

CUANDO TENEMOS DOS MUESTRAS O POBLACIONES — COEFICIENTE DE VARIACION / COVARIANZA

RANGO: (R) ES LA DIFERENCIA ENTRE EL MAYOR VALOR MENOS EL VALOR MENOR. PARA DATOS AGRUPADOS SE UTILIZAN LOS LIMITES MAYOR Y MENOR. ES COMÚN. ES COMUN NO UTILIZAR LA OPERACIÓN DE RESTA Y SOLAMENTE INDICARLA.

$$R = \begin{cases} V_{MAYOR} - V_{MENOR} & DNA \\ L_{MAYOR} - L_{MENOR} & DA \end{cases}$$

DNA = 82 - 30 = 52

DA = 84 - 29 = 55

DESVIACIÓN MEDIA (DM): LA DESVIACIÓN MEDIA O DESVIACIÓN PROMEDIO DE UN CONJUNTO DE DATOS ES EL PROMEDIO DE LAS DISTANCIAS DE CADA VALOR CON RESPECTO A LA MEDIA.

$$DM = \begin{cases} 1/n \sum_{i=1}^{n} |x_i - \overline{x}| & DNA \\ 1/n \sum_{i=1}^{m} |x_i - \overline{x}| f_i & DA \end{cases}$$

Marca de clase

NÚMERO MAGICO DE ESTADISTICA = 30

TANTO LA DESVIACION MEDIA COMO LA DESVIACION MEDIANA SON POCO UTILIZADAS EN LA PRÁCTICA POR LO DIFICIL DE MANEJAR EL VALOR ABSOLUTO. PARA ELIMINAR EL SIGNO DE LAS DIFERENCIAS DE $(x_i - \overline{x})$ Y EVITAR EL CALCULO ABSOLUTO SE DEFINE LA VARIANZA O VARIANCIA UTILIZANDO EL CUADRADO DE LAS DIFERENCIAS.

VARIANZA S_n^2 o S_{n-1}^2 DEPENDIENDO DEL VALOR QUE SE USE PARA PROMEDIAR

SE DIVIDE ENTRE n CUANDO SE CONSIDERA QUE SE TIENE TODOS LOS DATOS POSIBLES (POBLACIÓN), Y SE DIVIDE ENTRE n-1 CUANDO SE TIENE SOLO UNA FRACCIÓN DE LOS DATOS (MUESTRA). LA FORMULA PARA LA VARIANZA ES:

$$S_n^2 = \begin{cases} 1/n \sum_{i=1}^{n} (x_i - \bar{x})^2 \ ; \text{DNA} \\ 1/n \sum_{i=1}^{m} (x_i - \bar{x})^2 f_i \ ; \text{DA} \end{cases}$$

VARIANZA POBLACIONAL

$$S_{n-1}^2 = \begin{cases} 1/(n-1) \sum_{i=1}^{n} (x_i - \bar{x})^2 ; \ \text{DNA} \\ 1/(n-1) \sum_{i=1}^{m} (x_i - \bar{x})^2 f_i ; \ \text{DA} \end{cases}$$

VARIANZA MUESTRAL.

OJO: SI LA MUESTRA ES MAYOR A 30 SE CONSIDERA GRANDE. ES DECIR QUE A PARTIR DE UN TAMAÑO DE MUESTRA IGUAL O MAYOR A 30 ($n \geq 30$) PODEMOS OCUPAR LA PRIMERA FORMULA, YA QUE LOS RESULTADOS SON MUY SEMEJANTES ENTRE AMBAS FORMULAS,

PARA DA.

$$S_{n-1}^2 = \frac{1}{49} \left[(32.5-58.9)^2(4) + (40.5-58.9)^2(5) + (48.5-58.9)^2(9) + (56.5-58.9)^2(8) + (64.5-58.9)^2(8) + (72.5-58.9)^2(10) + (80.5-58.9)^2(6) \right]$$

$$S_{n-1}^2 = 212.2449$$

CECAYA

DESVIACION ESTANDAR: (S_n) (SE ALEJAN LOS DATOS DE LA MEDIA)
LA DESVIACIÓN ESTANDAR DE UN CONJUNTO DE DATOS ES LA RAIZ CUADRADA DE LA VARIANZA. SE DENOTA POR S_n O POR S_{n-1}, DEPENDIENDO DE SI SE OBTIENE LA DESVIACIÓN ESTÁNDAR DE UNA MUESTRA O DE TODA LA POBLACIÓN.

ES CLARO QUE PARA CALCULAR LA DESVIACIÓN ESTÁNDAR DEBE CALCULARSE LA VARIANZA PRIMERO, DE FORMA QUE:

$$S_n = \sqrt{S_n^2} \qquad S_{n-1} = \sqrt{S^2_{n-1}} \qquad S^2_{n-1} = \sqrt{212.24} = 14.56$$

RANGO INTERCUARTILICO (R_Q) EL RANGO INTERCUARTÍLICO DE UN CONJUNTO DE DATOS ES LA DIFERENCIA ENTR EL TERCER Y EL PRIMER CUARTIL.

$$R_Q = Q_3 - Q_1$$

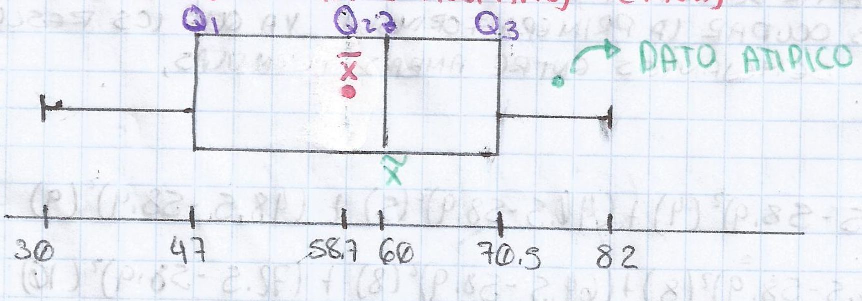
$R_Q = 70.5 - 43$ (SE TOMAN LOS VALORES DE LOS QUARTILES Y SE DEJA EXPRESADO)

(NOS DICE 50% DE LOS DATOS AL CENTRO)

NOS SIRVE PARA CONOCER LA MITAD DE LOS DATOS MÁS CERCA DE LA MEDIA.
EL RANGO SEMI-INTERCUARTILICO ES EL PROMEDIO DEL RANGO INTERCUARTILICO, ESTO ES.

$$Q = \frac{Q_3 - Q_1}{2}$$ (25% DE LOS DATOS AL CENTRO)

DIAGRAMA DE CAJA (DATOS HOSPITAL) (DNA)



RANGO
CUARTILES Q_1, Q_2, Q_3
MEDIA $\bar{x}$
MEDIANA ($\tilde{x}$)

BIGOTES.
1.5 x RANGO INTERCUARTILICO = 38.25

VALOR DONDE TERMINAN BIGOTES

$Q_1 - 1.5\, R_Q$

$Q_3 + 1.5\, R_Q$

CUANDO TENEMOS DOS MUESTRAS O POBLACIONES ES JUSTAMENTE PARA COMPARARLAS, BUSCAMOS SI VARIAN O DEPENDEN UNA DE LA OTRA.

COEFICIENTE DE VARIACION (CV) (¿CUANTO VARÍAN)

NOS AYUDA A COMPARAR LA VARIACIÓN DE 2 CARACTERES QUE POR SU NATURALEZA PUEDEN SER MUY DIFERENTES. LA ECUACION PARA CALCULARLO ES LA SIGUIENTE.

$CV = \frac{S_n}{\bar{X}}$ CV. POBLACIONAL $\qquad CV. = \frac{S_{n-1}}{\bar{X}}$ CV. MUESTRAL.

EJEMPLO: AÑOS DE ESTUDIO Y SUELDO DE 5 PERSONAS Y QUEREMOS SABER CUAL VARIA MÁS.

AÑOS DE ESTUDIO	SUELDOS
12	\$ 7,900
30	\$ 12,000
15	\$ 11,500
25	\$ 14,300
10	\$ 10,200

$\bar{X} = 18.4 \qquad \bar{y} = 11,180$

$S^2_{n-1} = 75.3 \qquad S^2_{n-1} = 5,557,000$

$S_{n-1} = 8.67 \qquad S_{n-1} = 2,357$

$CV = \frac{8.67}{18.4} \qquad CV = \frac{2,357}{11,180}$

$CV = 0.4712 \qquad CV = 0.2108$

CONCLUSIÓN: AÑOS DE ESTUDIO VARÍAN MÁS.

CRITERIO

$CV > 0.15 \longrightarrow$ DATOS HETEROGENEOS
$CV \leq 0.15 \longrightarrow$ DATOS HOMOGENEOS

CON ESTO CONCLUIMOS CUANDO SOLO TENEMOS UNA MUESTRA.

COVARIANZA (COV) (DEPENDENCIA QUE EXISTE ENTRE DOS VARIABLES)

$$Cov(x,y) = 1/n \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

$\bar{x}\bar{y} = 205,712$

$Cov = 11,648 \longrightarrow$ DATOS DE ARRIBA.

- SI LA COVARIANZA ES POSITIVA, ESTO INDICA QUE LA DEPENDENCIA ES DIRECTA, EN CASO DE SER NEGATIVA, LA DEPENDENCIA ES INVERSA.

MEDIDAS NUMERICAS. CELAYA

MEDIDAS DE FORMA (LAS MEDIDAS DE FORMA DE UN CONJUNTO DE DATOS SON EL SESGO Y LA CURTOSIS Y NOS MUESTRAN COMO SE DISTRIBUYEN LOS DATOS CON REFERENCIA A LA DISTRIBUCIÓN NORMAL.

PARA DEFINIR A LAS MEDIDAS, ES NESESARIO DEFINIR PRIMERO LOS MOMENTOS

MOMENTO RESPECTO AL ORIGEN

EL r-ÉSIMO MOMENTO CON RESPETO AL ORIGEN SE DEFINE MEDIANTE:

$$m'_r \begin{cases} 1/n \sum_{i=1}^{n} x_i^r \ ; \text{DNA} \\ 1/n \sum_{i=1}^{m} x_i^r f_i \ ; \text{DA} \end{cases}$$

MOMENTO CON RESPECTO A LA MEDIA

$$m_r \begin{cases} 1/n \sum_{i=1}^{n} (x_i - \bar{x})^r \ ; \text{DNA} \\ 1/n \sum_{i=1}^{m} (x_i - \bar{x})^r f_i \ ; \text{DA} \end{cases}$$

NOTA: DEBE OBSERVARSE QUE EL PRIMER MOMENTO CON RESPECTO AL ORIGEN m'_1 ES LA MEDIA, MIENTRAS QUE EL SEGUNDO MOMENTO CON RESPECTO A LA MEDIA m_2 ES LA VARIANZA S_n^2

SESGO (α_3)

EL SESGO DE UN CONJUNTO DE DATOS ES LA MEDIDA DEL GRADO DE SIMETRÍA (O ASIMETRÍA) DE LOS DATOS. SE DENOTA POR α_3 O POR a_3, Y SE DEFINE MEDIANTE

$$a_3 = m^3 / s^3$$

DONDE m_3 ES EL TERCER MOMENTO CON RESPETO A LA MEDIA Y s ES LA DESVIACIÓN ESTÁNDAR.

EJERCICIO PARA LOS DATOS QUE SE VIERÓN EN EL EJERCICIO ANTERIOR OBTENER EL SESGO PARA DATOS AGRUPADOS

$$m_3 = 1/50 \left[(32.5 - 58.9)^3 (4) + \ldots + (80.5 - 58.9)^3 (6) \right]$$

$$S^3 = (14.56)^3 \Rightarrow \quad \alpha_3 = \frac{-559.104}{3086.62} = -0.1811$$

EL SESGO SE COMPARA CON 0

- CUANDO EL COEFICIENTE DE SESGO ES MENOR QUE CERO SE DICE QUE LOS DATOS TIENEN UNA DISTRIBUCIÓN SESGADA A LA IZQUIERDA O CON SESGO NEGATIVO ($\alpha_3 < 0$)
- CUANDO EL COEFICIENTE DE SESGO ES POSITIVO SE DICE QUE LOS DATOS TIENEN UNA DISTRIBUCIÓN SESGADA A LA DERECHA O CON SESGO POSITIVO. ($\alpha_3 > 0$)
- SI EL COEFICIENTE DE SESGO ES CERO, ENTONCES LOS DATOS TIENEN UNA DISTRIBUCIÓN SIMÉTRICA O INSESGADA.

NOTA: PARA LAS MEDIDAS DE FORMA SE REQUIERE MENCIONAR QUE DISTRIBUCIÓN TIENEN LOS DATOS.

CURTOSIS (α_4)

EL COEFICIENTE DE CURTOSIS DE UN CONJUNTO DE DATOS MIDE EL GRADO DE APLANAMIENTO RELATIVO DE LA DISTRIBUCIÓN DE LOS DATOS. SE DENOTA MEDIANTE α_4 O POR a_4. SE DEFINE MEDIANTE LA EXPRESIÓN.

$$\alpha_4 = \frac{m_4}{S^4} = \frac{m_4}{(S^2)^2}$$

EJERCICIO: PARA LOS DATOS QUE SE VIERON EN EL EJERCICIO ANTERIOR OBTENER LA CURTOSIS PARA DATOS AGRUPADOS.

$$m_4 = \left[1/50 \left[(32.5 - 58.9)^4 (4) + \ldots + (80.5 - 58.9)^4 (6) \right] \right]$$

$$S^4 = (14.56)^4$$

$$\alpha_4 = \frac{85{,}554.78}{44{,}941.78}$$

$$\alpha_4 = 1.8992$$

CECAYA

CRITERIO

LA CURTOSIS SE COMPARA CONTRA TRES, PORQUE TRES ES LA CURTOSIS DE LA DISTRIBUCIÓN NORMAL.

- SI LOS DATOS TIENEN UNA DISTRIBUCIÓN MÁS PUNTIAGUDA QUE LA DISTRIBUCIÓN NORMAL $\alpha_4 > 3$ ENTONCES SE DICE QUE LOS DATOS TIENEN UNA DISTRIBUCIÓN LEPTOCÚRTICA.

- SI LOS DATOS TIENEN UNA DISTRIBUCIÓN COMO LA NORMAL, $\alpha_4 = 3$, ENTONCES SE DICE QUE LA DISTRIBUCIÓN ES MESOCÚRTICA.

- SI LOS DATOS TIENEN UNA DISTRIBUCIÓN APLANADA $\alpha_4 < 3$ ENTONCES RECIBE EL NOMBRE DE PLATICÚRTICOS.

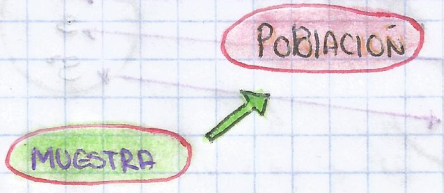
(SI LOS DATOS ESTÁN ALEJADOS O PEGADOS A LA MEDIA)
(LA CURTOSIS NO NOS PUEDE DAR UN VALOR NEGATIVO)

NOTA: PARA LAS MEDIDAS DE FORMA SE REQUIERE MENCIONAR QUE LA DISTRIBUCIÓN TIENEN LOS DATOS.

TEMA 2: CONCEPTOS BÁSICOS DE INFERENCIA ESTADÍSTICA.

INFERENCIA ESTADÍSTICA: LA INFERENCIA ESTADÍSTICA ES LA PARTE DE LA ESTADÍSTICA QUE TIENE POR OBJETO CONCLUSIONES ACERCA DE TODA UNA POBLACIÓN A PARTIR DE LA INFORMACIÓN CONTENIDA EN UNA MUESTRA, CUANTIFICANDO EN FORMA PROBABILÍSTICA EL GRADO DE CERTIDUMBRE DE DICHAS CONCLUSIONES.

Y PARA OBTENER LA INFORMACIÓN QUE PERMITE GENERAR LAS CONCLUSIONES UTILIZA EL MUESTREO ALEATORIO SIMPLE, DEL CUAL SE OBTIENEN MUESTRAS REPRESENTATIVAS.



PARÁMETRO: UN PARÁMETRO ESTADÍSTICO ES UN NÚMERO QUE RESUME EL COMPORTAMIENTO DE UNA VARIABLE ALEATORIA Y QUE DESCRIBE PARCIAL O COMPLETAMENTE SU DISTRIBUCIÓN DE PROBABILIDAD

LA MEDIA μ Y LA VARIANZA σ^2 SON PARÁMETROS DE CUALQUIER V.A.

DEFINICIÓN: LAS V.A. $X_1, X_2, \ldots, X_n$ FORMAN UNA MUESTRA ALEATORIA DEL TAMAÑO n, SI SON INDEPENDIENTES Y TIENEN LA MISMA DISTRIBUCIÓN DE PROBABILIDAD.

A PARTIR DE ESTE MOMENTO, DEBEMOS TENER PRESENTE LAS CARACTERÍSTICAS DE INDEPENDENCIA E IDÉNTICA DISTRIBUCIÓN. (I.I.D)

RECORDATORIO EN PROBABILIDAD SE VIO QUE ERA UNA VARIABLE ALEATORIA (V.A) Y ES UNA FUNCIÓN CUYO DOMINIO ES EL ESPACIO MUESTRAL Y RANGO O CONTRADOMINIO UN SUBCONJUNTO DE LOS NÚMEROS REALES, SE SUELEN DENOTAR X, Y, Z

ESPACIO MUESTRAL

$$X: S \longrightarrow R \quad ; \; R \in \mathbb{R}$$

VARIABLE

$$x_i \in X_i$$

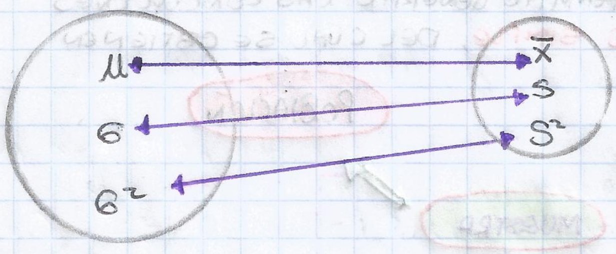
DATOS

CADA DATO ES UNA REALIZACIÓN DE LA VA.

CECAYA

ESTADÍSTICO: UN ESTADÍSTICO ES UNA FUNCIÓN DE LAS V.A. QUE SE PUEDEN OBSERVAR EN UNA MUESTRA Y QUE NO DEPENDE DE LOS PARÁMETROS DESCONOCIDOS.

LA MEDIA $\bar{X}$ Y LA VARIANZA S^2 SON LOS ESTADÍSTICOS QUE MÁS SE UTILIZARÁN



PARÁMETROS POBLACIONALES — ESTADÍSTICOS MUESTRALES

DISTRIBUCIÓN MUESTRAL DE LA MEDIA: DISTRIBUCIÓN NORMAL

LA DISTRIBUCIÓN NORMAL SERVIRÁ PARA CARACTERIZAR LA DISTRIBUCIÓN MUESTRAL DE LA MEDIA

TEOREMA 1:

SI $X_1, X_2, \ldots, X_n$ SON VARIABLES ALEATORIAS INDEPENDIENTES, Y TODOS CON DISTRIBUCIÓN NORMAL CON MEDIA μ_i Y VARIANZA σ_i^2, ENTONCES LA VARIABLE ALEATORIA Y DEFINIDA COMO

$$Y = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

TIENE DISTRIBUCIÓN NORMAL CON MEDIA $\sum_{i=1}^{n} \mu_i$ Y VARIANZA $\sum_{i=1}^{n} \sigma_i^2$

TEOREMA 2:

SI $X_1, X_2, \ldots X_n$ SON VARIABLES ALEATORIAS INDEPENDIENTES E IDÉNTICAMENTE DISTRIBUIDAS Y $X_i \sim N(\mu, \sigma^2)$ ENTONCES LA VARIABLE ALEATORIA

$$Y = \sum_{i=1}^{n} X_i$$

TIENE DISTRIBUCIÓN NORMAL $Y \sim N(n\mu, n\sigma^2)$ POR LO QUE $\dfrac{Y - n\mu}{\sqrt{n}\,\sigma} \sim N(0,1)$

TEOREMA 3:

SI $X_1, X_2, \ldots, X_n$ SON VARIABLES ALEATORIAS INDEPENDIENTES Y TODAS CON DISTRIBUCIÓN NORMAL CON MEDIA μ_i Y VARIANZA σ_i^2 ENTONCES LA VARIABLE ALEATORIA Y_c DEFINIDA COMO:

$$Y_c = c_1 X_1 + c_2 X_2 + \ldots + c_n X_n = \sum_{i=1}^{n} c_i X_i$$

TIENE DISTRIBUCION NORMAL CON MEDIA $\sum_{i=1}^{n} c_i \mu_i$ Y VARIANZA $\sum_{i=1}^{n} c_i^2 \sigma_i^2$

TEOREMA 4:

SEA $X_1, X_2, \ldots, X_n$ UNA MUESTRA ALEATORIA CON DISTRIBUCIÓN NORMAL Y MEDIA μ Y VARIANZA σ^2. ENTONCES LA VARIABLE ALEATORIA $\bar{X} = 1/n \sum_{i=1}^{n} X_i$ TIENE DISTRIBUCIÓN NORMAL CON MEDIA μ Y VARIANZA σ^2/n.

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

NOTA: LA PROPIEDAD DE ADITIVIDAD DE LA DISTRIBUCIÓN NORMAL PERMITE QUE LAS MEDIAS SE PUEDAN RESTAR, PERO LAS VARIANZAS SE SIGUEN SUMANDO, DEBIDO A LAS PROPIEDADES DE LA SUMA DE VARIANCIAS.

EJERCICIO 1

UN EJE DE DIAMETRO EXTERIOR DISTRIBUIDO NORMALMENTE CON MEDIA 1.2 cm Y VARIANCIA 0.0016 cm² SE INSERTA EN UN COJINETE QUE TIENE UN DIÁMETRO INTERIOR DISTRIBUIDO NORMALMENTE CON MEDIA 1.25 cm Y VARIANZA 0.0009 cm². DETERMINAR LA PROBABILIDAD DE QUE NO EMBONEN LAS PIEZAS.

$$X_1 \sim N(1.2, 0.0016) \qquad X_2 \sim N(1.25, 0.0009)$$

PARA QUE NO EMBONEN $X_1 > X_2$ O BIEN $Y = X_1 - X_2 > 0$

$$Y \sim N(-0.05, 0.0025) \qquad \frac{Y - n\mu}{\sqrt{n}\,\sigma} \sim N(0,1)$$

POR LO QUE

$$P(Y>0) = P\left(\frac{Y-(-0.05)}{\sqrt{0.0025}} > \frac{0-(-0.05)}{\sqrt{0.0025}}\right)$$

$$P(Z>1)$$

CUANDO NO NOS DIGAN TAMAÑO MUESTR $n=1$.

CELAYA

IMPORTANTE:
LOS TEOREMAS ANTERIORES SON ÚTILES SI SE DESEA OBTENER LA DISTRIBUCIÓN DE UNA SUMA DE VARIABLES ALEATORIAS DISTRIBUIDAS NORMALMENTE ; SIN EMBARGO, ESTO NO SIEMPRE OCURRE. EN GENERAL LAS VARIABLES ALEATORIAS PUEDEN TENER CUALQUIER DISTRIBUCIÓN Y NO SOLO LA NORMAL. CUANDO SE PRESENTAN ESTOS CASOS SE UTILIZA EL TEOREMA CENTRAL DE LÍMITE.

DISTRIBUCIÓN MUESTRAL DE LA MEDIA : TEOREMA CENTRAL DE LÍMITE.

SEAN $X_1, X_2, \ldots, X_n$ UN CONJUNTO DE VARIABLES ALEATORIAS INDEPENDIENTES E IDÉNTICAMENTE DISTRIBUIDAS CON PARAMETROS $E(X_i) = \mu_X$ Y $VAR(X_i) = \sigma_X^2$ PARA $i = 1, 2, \ldots, n$ ENTONCES LA VARIABLE ALEATORIA $\bar{X}$ DEFINIDA COMO $\bar{X} = 1/n \sum_{i=1}^{n} X_i$.

TIENE UNA DISTRIBUCIÓN QUE CONVERGE A LA NORMAL, CON PARAMETROS $\mu_{\bar{X}} = \mu_X$ Y $\sigma_X = \frac{\sigma_X}{\sqrt{n}}$ ESTÁNDAR CUANDO $n \to \infty$, ESTO ES:

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X,\ \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\right)$$

EN GENERAL, SI LAS n V.A. INDEPENDIENTES TIENEN PARÁMETROS $E(X_i) = \mu_i$ Y $Var(X_i) = \sigma_i^2$ Y SE DEFINE LA V.A. $Y = \sum_{i=1}^{n} X_i$ ENTONCES:

$$\frac{Y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$$

TIENE DISTRIBUCIÓN QUE CONVERGE A LA NORMAL ESTANDAR.
EN LA PRÁCTICA, EL TEOREMA CENTRAL DE LIMITE (TCL) PROPORCIONA UNA BUENA APROXIMACIÓN CUANDO ES MAYOR O IGUAL A 30, SIN IMPORTAR LA DISTRIBUCIÓN DE LAS VA. DE MUESTREO.

EL TEOREMA CENTRAL DE LIMITE SE AJUSTA SOLO A $\overline{X}$ O A LA SUMATORIA T:

$$Z = \frac{\overline{X} - \mu}{6/\sqrt{n}} \qquad Z = \frac{T - n\mu}{6/\sqrt{n}}$$

EJERCICIO 2.

LA RESISTENCIA A LA RUPTURA DE UN REMACHE ESPECIAL TIENE UN VALOR MEDIO DE 10,000 [kg] POR CENTIMETRO CUADRADO Y UNA DESVIACIÓN ESTÁNDAR DE 500

a) ¿CUÁL ES LA PROBABILIDAD DE QUE LA RESISTENCIA MEDIA A LA RUPTURA DE LA MUESTRA PARA UNA MUESTRA ALEATORIA DE 40 REMACHES, SEA ENTRE 9900 Y 10200?

b) SI EL TAMAÑO MUESTRAL HUBIERA SIDO 15 EN LUGAR DE 40, ¿PODRÍA CALCULARSE LA PROBABILIDAD PEDIDA EN EL INCISO (a)?

a) $= P(9900 \leq \overline{X} \leq 10200) = P\left(\frac{9900 - 10{,}000}{500/\sqrt{40}} \leq z \leq \frac{10200 - 10{,}000}{500/\sqrt{40}}\right)$

$Z = \frac{\overline{x} - \mu}{6/\sqrt{n}}$ $\quad = P(-1.26 \leq z \leq 2.53) = 0.8905$

b) COMO $n = 15$ SE REQUIERE CONOCER LA DISTRIBUCIÓN DE LA POBLACIÓN, YA QUE TCL SOLO SE PUEDE UTILIZAR CON UN $n >= 36$ NO SE PUEDE CALCULAR LA PROBABILIDAD.

CELAYA

EJERCICIO 3.

LA CANTIDAD PROMEDIO QUE GASTA UNA EMPRESA DURANTE UN AÑO EN SERVICIOS MÉDICOS POR CADA EMPLEADO FUE DE $ 2 575 PESOS CON UNA DESVIACIÓN ESTÁNDAR DE $325 PESOS.

a) ¿CUÁLES LA PROBABILIDAD QUE A PARTIR DE UNA MUESTRA ALEATORIA DE 70 EMPLEADOS SE OBSERVE UNA MEDIA MUESTRAL COMPRENDIDA ENTRE $2500 Y $2700? ¿QUÉ DIFERENCIAS ENCUENTRA CON EL PROBLEMA 4 DEL EJERCICIO 11.2?

b) ¿CUÁL ES LA PROBABILIDAD QUE A PARTIR DE UNA MUESTRA ALEATORIA DE 100 EMPLEADOS SE OBSERVE UNA MEDIA MUESTRAL MENOR A $2600?

c) EL GERENTE DE LA EMPRESA ASEGURA QUE GASTA UN PROMEDIO MÍNIMO ANUAL DE $2600 EN SERVICIOS MÉDICOS EN AL MENOS 90% DE SUS EMPLEADOS ¿QUÉ TAMAÑO MÍNIMO DE LA MUESTRA DEBE ELEGIR EL GERENTE PARA CONFIRMAR SU AFIRMACIÓN DE MANERA ESTADÍSTICA?

D) EL LIDER SINDICAL DE LA EMPRESA AFIRMA QUE LA EMPRESA GASTA UN PROMEDIO MÍNIMO ANUAL DE $2 550 EN SERVICIOS MÉDICOS EN A LO MÁS 25% DE SUS EMPLEADOS. ¿QUÉ TAMAÑO MÍNIMO DE LA MUESTRA DEBE ELEGIR EL LIDER SINDICAL PARA CONFIRMAR DE MANERA ESTADÍSTICA SU AFIRMACIÓN?

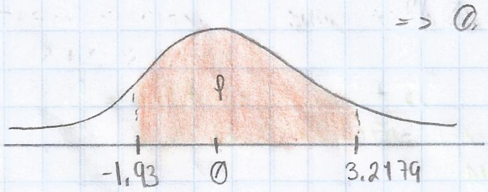
a) $\mu = 2{,}575$ $\quad \sigma_x = 325 \quad n = 70$

MEDIA MUESTRAL

$$P(2500 \le \bar{x} \le 2700) = P\left(\frac{2500 - 2575}{325/\sqrt{70}} \le z \le \frac{2700 - 2575}{325/\sqrt{70}}\right)$$

$$= P(-1{,}9307 \le z \le 3{,}2179)$$

$$\Rightarrow 0{,}9923 - 0{,}0268 = 0{,}9725$$



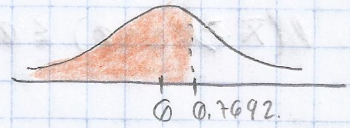
∴ LA PROBABILIDAD DE ESTAR ENTRE 2500 Y 2700 PESOS ES 97,25 %.

VALOR DIRETO EN TABLAS.

b) $n = 100$ $\quad P(\bar{x} < 2600) = ?$

z_0

$$P\left(z < \frac{2600 - 2575}{325/\sqrt{100}}\right) = P(z < 0{,}7692)$$



DE TABLAS. $\quad P(\bar{x} < 2600) = 0{,}7794$

LA PROBABILIDAD DE QUE LA MEDIA MUESTRAL CON $n = 100$ SEA MENOR A 2600 ES 77,94 %

c) $n = ?$ $\quad \bar{x} = 2600$ $\quad P(\bar{x} \geq 2600) \geq 0.9$

$$P\left(z \geq \frac{2600 - 2575}{325/\sqrt{n}}\right) = 0.9$$

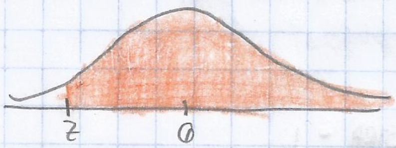
BUSCAMOS Z EN TABLAS. 0,1 PORQUE LAS TABLAS DAN <

z

PARA 10 % = -1,282

PARA 90 % = 1,282

SOLO CAMBIA EL SIGNO PORQUE ES SIMETRICA



POR QUE ES MAYOR

=> DESPEJANDO 'n'

$$\frac{-1{,}282 \cdot 325}{\sqrt{n}} = 2600 - 2575$$

$$\left(\frac{-1{,}282 \cdot 325}{2600 - 2575}\right)^2 = n$$

∴ EL TAMAÑO MINIMO DE LA MUESTRA ES DE 278 EMPLEADOS.

$n = 277{,}7555$

$n = 278$ personas REDONDEO SIEMPRE HACIA ARRIBA

NOTA: SIEMPRE QUE CALCULEMOS UN TAMAÑO DE MUESTRA ES $n \geq$

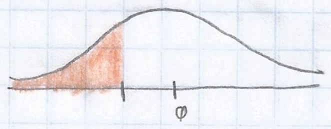
CELAYA

$$P(\bar{X} \geq 2550) \leq 0.25$$

$$P\left(Z \geq \frac{2550-2575}{325/\sqrt{n}}\right) \Rightarrow n \geq \left(\frac{-0.674 \cdot 325}{2550-2575}\right)^2$$

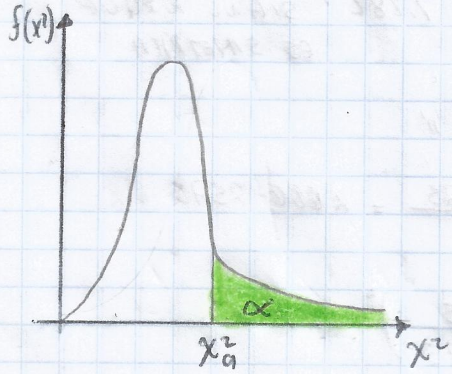
$n \geq 76.77$

$n \geq 77$ PERSONAS.



DISTRIBUCIÓN PARA CARACTERIZAR A LA VARIANCIA.

POR LAS CARACTERISTICAS ESPECIALES QUE PRESENTA LA DISTRIBUCIÓN NORMAL, SE HAN ESTUDIADO DISTRIBUCIONES QUE PUEDEN GENERARSE A PARTIR DE ELLA, TAL ES EL CASO DE LA DISTRIBUCIÓN JI CUADRADA.



GRADOS DE LIBERTAD

$v = (n-1)$ TAMAÑO MUESTRA $- 1$

LA DISTRIBUCIÓN JI CUADRADA PRESENTA UN SESGO POSITIVO.

SEAN $Z_1, Z_2, \dots, Z_v$; v VARIABLES ALEATORIAS INDEPENDIENTES CON DISTRIBUCIÓN NORMAL ESTÁNDAR

$$X^2 = Z_1^2 + Z_2^2 + \dots + Z_v^2$$

ES UNA V.A. QUE RECIBE EL NOMBRE DE JI CUADRADA CON v GRADOS DE LIBERTAD Y SE DENOTA MEDIANTE EL SIMBOLO $\chi^2_{(v)}$

SU FUNCIÓN DE DENSIDAD ES:

$$f_{\chi^2}(x) = \begin{cases} \dfrac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)} x^{\left(\frac{v}{2}-1\right)} e^{-\frac{x}{2}} & x > 0 \\ 0 & \text{eoc.} \end{cases}$$

TEOREMA 6

LA DISTRIBUCIÓN JI CUADRADA, AL IGUAL QUE LA DISTRIBUCIÓN NORMAL, PRESENTAN LAS CARACTERÍSTICAS DE ADITIVIDAD

SEAN $X_1^2, X_2^2, \ldots X_n^2$ VARIABLES ALEATORIAS INDEPENDIENTES CON DISTRIBUCIÓN JI CUADRADA Y V_1, V_2, V_n GRADOS DE LIBERTAD, RESPECTIVAMENTE. ENTONES LA VARIABLE.

$$Y = X_1^2 + X_2^2 + \ldots + X_n^2 = \sum_{i=1}^{n} X_i^2$$ TIENE DISTRIBUCIÓN JI CUADRADA

CON $V = \sum_{i=1}^{n} V_i$

TEOREMA 7

SI X^2 ES UNA V.A QUE TIENE DISTRIBUCIÓN JI CUADRADA CON V GRADOS DE LIBERTAD, $X^2 \sim \chi^2_{(V)}$, ENTONCES.

$$E(X^2) = V \quad , \quad Var(X^2) = 2V$$

Y LA FUNCIÓN GENERADORA DE MOMENTOS DE X^2 ES.

$$M_{X^2}(\theta) = (1 - 2\theta)^{-\frac{V}{2}}$$

PARA CARACTERIZAR A LA VARIANZA DE UNA MUESTRA O POBLACIÓN SE UTILIZA LA DISTRIBUCIÓN JI CUADRADA DE LA SIGUIENTE FORMA

VARIANCIA

$$\frac{n S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

PARAMETRO

$n \geq 30$

ESTADÍSTICO

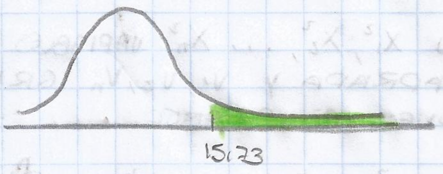
$$\frac{(n-1) S_{n-1}^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

$n < 30$

CELAYA

EJERCICIO 4 y 5.

a) Si $X^2 \sim \chi^2_{(15)}$ OBTENER $P(X^2 \geq 15{,}73)$

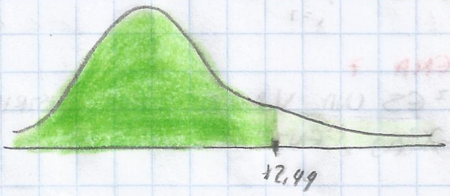


POR TABLAS $P(X^2 \geq 15{,}75) = 0{,}4$

b) Si $X^2 \sim X^2_{(20)}$ OBTENER $P(X^2 < 12{,}49)$

$1 - P(X^2 \geq 12{,}44)$

$P(X^2 < 12{,}44) = 1 - 0.9 = 0.1$



EJERCICIO 6

SEA UNA MUESTRA ALEATORIA DE TAMAÑO 20 TOMADA DE UNA POBLACIÓN CON MEDIA 8 Y VARIANCIA 4. OBTENER LA PROBABILIDAD DE QUE LA VARIANZA MUESTRAL S^2_{n-1} SEA MAYOR O IGUAL A 5,7

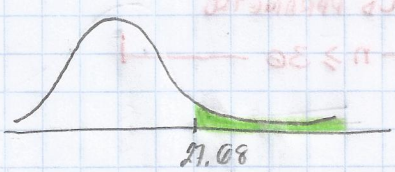
$n = 20 \quad v = 19 \quad \mu = 8 \quad \sigma^2 = 4$

$P(S^2_{n-1} \geq 5{,}7)$

$$P\left(\overset{X^2}{\frac{n-1}{\sigma^2} S^2_{n-1}} \geq \overset{X^2_0}{\frac{19}{4}(5{,}7)}\right) = P(X^2 \geq 27{,}08)$$

DIRECTO DE TABLAS

$P(X^2 \geq 27{,}08) = 0.1$



DISTRIBUCION PARA CARACTERIZAR A LA MEDIA MUESTRAL, MUESTRA PEQUEÑA, VARIANZA DESCONOCIDA.

TEOREMA 8. SEAN Z Y X^2 DOS VA. INDEPENDIENTES CON DISTRIBUCIÓN NORMAL ESTÁNDAR Y JI-CUADRADA RESPECTIVAMENTE, ES DECIR:

$$Z \sim (0,1) \quad , \quad X^2 \sim \chi^2(v)$$

ENTONCES LA VA. T DEFINIDA COMO $T = \dfrac{Z}{\sqrt{\dfrac{X^2}{V}}}$

TIENE UNA DISTRIBUCIÓN t de STUDENT CON V GRADOS DE LIBERTAD Y FUNCIÓN DE DENSIDAD DADA POR.

$$f_n(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}} \qquad -\infty < t < \infty$$

$$v > 0$$

TEOREMA 9:

SI T ES UNA VARIABLE ALEATORIA QUE TIENE DISTRIBUCIÓN t-STUDENT CON v GRADOS DE LIBERTAD.

$\Rightarrow$

$$E(T) = 0 \qquad VAR(T) = \frac{v}{v-2} \qquad v > 2$$

COMO PUEDE OBSERVARSE LA DISTRIBUCIÓN t DE STUDENT POSEE EL PARAMETRO v, QUE AL IGUAL QUE PARA LA DISTRIBUCIÓN JI-CUADRADA, RECIBE EL NOMBRE DE GRADOS DE LIBERTAD.

LA FUNSIÓN DE DENSIDAD t DE STUDENT ES SIMÉTRICA Y UNIMODAL AL IGUAL QUE LA NORMAL, PERO SIEMPRE ESTA CENTRADA EN CERO; ES MUY PARECIDA. A LA DISTRIBUCIÓN NORMAL ESTANDAR.

PARA CARACTERIZAR A LA MEDIA DE UNA MUESTRA PEQUEÑA CON VARIANZA DESCONOCIDA SE UTILIZA LA DISTRIBUCIÓN T-STUDENT DE LA SIGUIENTE FORMA.

$$\frac{\bar{X} - \mu_x}{\frac{S_{n-1}}{\sqrt{n}}} = T \sim t_{(n-1)}$$

ECLAYA

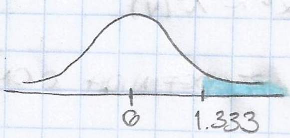
EJERCICIO 7 Y 8

SEA T UNA V.A. CON DISTRIBUCION t-student, CON 19 GRADOS DE LIBERTAD, OBTENER:

7) $P(T > 1.33)$

8) EL VALOR DE t, tal que $P(T < t) = 0.01$

7) $P(T > 1.33) = 0.10$ → POR TABLAS

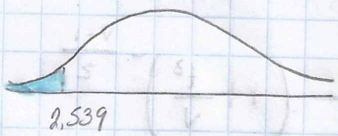


8) $P(T < t) = 0.01$

$P(T < t) = P(T > -t) = 0.01$

$-t = 2.539$

$t = -2.539$



EJERCICIO 9

LOS SIGUIENTES SEIS DATOS SON LOS TIEMPOS DE PERMANENCIA (ESPERA Y ATENCIÓN) EN UN BANCO:

15, 32, 18, 26, 27 Y 20

SI EL BANCO AFIRMA QUE EL TIEMPO PROMEDIO DE PERMANENCIA ES DE 20 MINUTOS O MENOS, DETERMINAR SI LA AFIRMACIÓN ES RAZONABLE.

$$\bar{x} = \frac{1}{6}(15+32+18+26+27+20) = 23$$

$$S_{n-1}^2 = \frac{1}{5}(15-23)^2+(32-23)^2+(18-23)^2+(26-23)^2+(27-23)^2+(20-23)^2$$

$$S_{n-1}^2 = (6.387)^2$$

$\nu = 5$

$$P(\bar{X} > 23) \Rightarrow P\left(T > \frac{23-20}{\frac{6.387}{\sqrt{6}}}\right) = P(T \geq 1.15) = 0.15 = 15\%$$

NECESITAMOS ALEJARNOS POR ESO (>)

DE QUE LA MEDIA SEA IGUAL O MAYOR A 23.

∴ LOS DATOS NO PROPORCIONAN UNA FUERTE EVIDENCIA DE QUE EL BANCO NO TENGA RAZÓN, ES DECIR, LA AFIRMACIÓN DEL BANCO PUEDE SER RAZONABLE.

DISTRIBUCIÓN MULTINOMIAL.

ESCALA DE LICKERT → ESCALA DE SATISFACCIÓN → CUALITATIVAS

TS	S	NS	I	TI
5	4	3	2	1

TS = TOTALMENTE SATISFECHO
TI = TOTALMENTE INSATISFECHO
PONDERAR # IMPAR PARA QUE ALLA PUNTO MEDIO.

SEAN $X_1, X_2, \ldots, X_n$ VARIABLES ALEATORIAS CONJUNTAS QUE REPRESENTAN EL NÚMERO DE ÉXITOS DEL PRIMER TIPO X_1, EL NÚMERO DE EXTOS DEL SEGUNDO TIPO X_2, ETC. CON $\sum_{i=1}^{n} x_i = n$ # ensayos QUE SE OBTIENE EN n ENSAYOS INDEPENDIENTES, CADA UNO DE LOS CUALES PERMITE K RESULTADOS MUTUAMENTE EXCLUYENTES, CUYAS PROBABILIDADES SON $P_1, P_2, \ldots, P_K$ CON $\sum_{i=1}^{n} = 1$, ENTONCES LAS V.A. TIENEN DISTRIBUCIÓN MULTINOMIAL CON PARAMETROS n y $P_1, P_2, \ldots, P_K$

$$\boxed{f_{X_1 X_2, \ldots, X_K}(X_1, X_2, \ldots, X_n) = \frac{n!}{X_1!\, X_2! \ldots X_n!} P_1^{X_1} P_2^{X_2} \ldots P_n^{X_n}}$$

PARA $X_i = 0, 1, \ldots, n$; $1 \leq i \leq K$ y $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{n} P_i = 1$

EJEMPLO

SUPONGASE QUE UN PROCESO DE PRODUCCIÓN SE SELECCIONAN, DE MANERA ALEATORIA, 25 ARTICULOS. ESTE PRODUCCIÓN POR LO GENERAL PRODUCE UN 90% DE ARTÍCULOS LISTOS PARA VENDERSE Y UN 7% REPROCESABLES ¿CUAL ES LA PROBABILIDAD DE QUE 22 DE 25 ARTICULOS ESTÉN LISTOS PARA VENDERSE Y QUE SEAN REPROCESABLES?

$n = 25$

ARTICULOS LISTOS PARA VENDERSE $= P_1 = 0.9$

ARTICULOS REPROCESABLES $= P_2 = 0.07$

ARTICULOS NO REPROCESABLES Y NO ESTEN LISTOS $= 0.03$

ENTONCES LAS VARIABLES TIENEN DISTRIBUCIÓN MULTINOMIAL POR LO QUE

$P(X_1 = 22, X_2 = 2, X_3 = 1)$

$$P(X_1 = 22, X_2 = 2, X_3 = 1) = \frac{25!}{22!\, 2!\, 1!} (0.9)^{22} (0.03)^{1} = 0.0998 = 9.98\%$$

REGRESIÓN LINEAL CELAYA.

LA REGRESIÓN PROPORSIONA LA POSIBLE RELACIÓN ENTRE LAS VARIABLES DE UNA ECUACIÓN CON EL OBJETIVO DE PREDECIR UNA DE ELLAS (VARIABLE DEPENDIENTE O VARIABLE DE SALIDA) EN FUNCIÓN DE LA OTRA U OTRAS (VARIABLES INDEPENDIENTES O VARIABLES DE ENTRADA) EXISTEN DOS TIPOS DE REGRESIÓN EN GENERAL.

1) REGRESIÓN SIMPLE 2) REGRESIÓN MULTIPLE

LA REGRESIÓN SIMPLE SE UTILIZA CUANDO SE RELACIONAN DOS VARIABLES MIENTRAS QUE LA MÚLTIPLE SE UTILIZA PARA MÁS DE DOS VARIABLES. EN ESTE CURSO SE ESTUDIARÁ LA REGRESIÓN SIMPLE EN LA CUAL EL TIPO DE CURVA PUEDE SER, LINEAL, POLINOMIAL, EXPONENCIAL Y ALGUNOS OTROS MODELOS. EN ESTE TEMA NOS CONCENTRAREMOS EN EL ESTUDIO DE REGRESIÓN LINEAL SIMPLE HACIENDO LO QUE SE LLAMA ANÁLISIS ESTADÍSTICO BIVARIADO DENOMINADO ASÍ POR EL MANEJO DE DOS CONJUNTOS DE DATOS EL CASO MÁS COMÚN DE ANÁLISIS ESTADÍSTICO BIVARIADO ES EL AJUSTE POR MÍNIMOS CUADRADOS.

AJUSTE POR MÍNIMOS CUADRADOS.

PARTIENDO DE QUE SE DESEA OBTENER UN MODELO LINEAL PARA LA VARIABLE DEPENDIENTE 'y' EN FUNCIÓN DE LA VARIABLE INDEPENDIENTE 'x', SE ESCRIBE.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

DONDE ε ES UN ERROR ALEATORIO QUE SE OBTIENE DEBIDO AL MODELO SIN CONSIDERAR EL ERROR EL MODELO SE PUEDE ESCRIBIR COMO

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \rightarrow \text{ESTIMADOR A PARTIR DE UNA MUESTRA.}$$

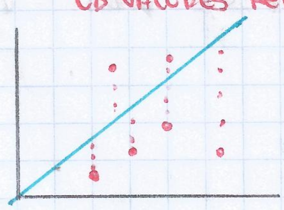
DONDE LA PENDIENTE Y LA ORDENADA AL ORIGEN TIENE UN ACENTO CIRCUNFLEJO PARA INDICAR QUE SE TRATA DE APROXIMACIONES DE LOS VERDADEROS PARAMETROS.

(x) (y)

CONSIDERANDO EL VALOR REAL Y EL APROXIMADO PARA CADA PUNTO OBTENER LA SUMA DE ERRORES CUADRADOS.

$$SEC = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 \rightarrow \text{DISTANCIAS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

↳ VALORES REALES



$$\frac{\partial SEC}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \qquad \frac{\partial SEC}{\partial \hat{\beta}_1} = -2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

DE DONDE

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - x)^2}$$

O BIEN

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\overbrace{\left(\sum_{i=1}^{n} y_i\right)}^{x} \overbrace{\left(\sum_{i=1}^{n} x_i\right)}^{y}}{n}}{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right.}$$

X	Y	X^2	Y^2	XY

Σ MULTIPLICACIÓN DE X Y Y = SS_{xy}

SUMA DE LA MULTIPLICACIÓN DE X CON X

$x^2 = SS_{xx}$

$B_1 = m$

SIMPLIFICANDO NOTACIÓN

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \qquad SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

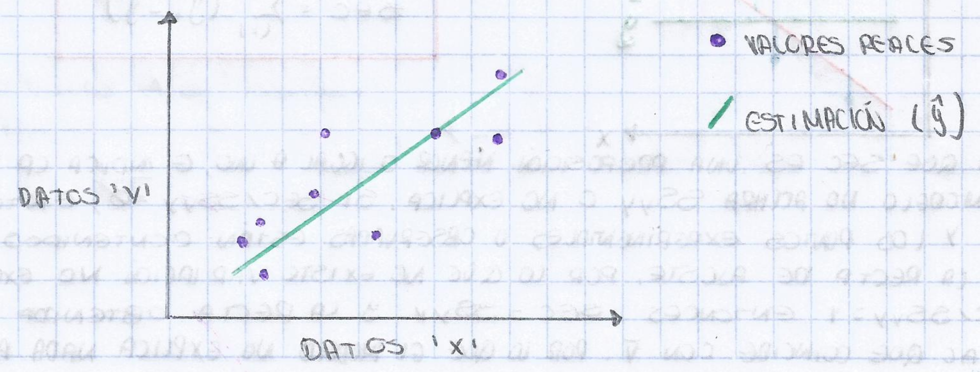
$$\Rightarrow \boxed{B_1 = \frac{SS_{xy}}{SS_{xx}}}$$

$\hat{\beta}_0$ y $\hat{\beta}_1$ SON ESTIMADORES (APROXIMADORES) INSESGADOS DE β_0 Y β_1, QUE SON LOS PARÁMETROS QUE SE DESEA OBTENER.

CABE ACLARAR QUE LA ECUACIÓN DE REGRESIÓN QUE SE OBTENGA ES VÁLIDA SOLO PARA PAREJAS COMPRENDIDAS EN EL RANGO DONDE SE HA EXPERIMENTADO.

DIAGRAMA DE DISPERSIÓN

UNA VEZ QUE SE HA DETERMINADO LA ECUACIÓN DE REGRESIÓN, ES ÚTIL LA REPRESENTACIÓN GRÁFICA DE LOS PUNTOS DE DATOS EN EL PLANO Y EN LO QUE SE DENOMINA DIAGRAMA DE DISPERSIÓN. CUANDO LA REGRESIÓN ES LINEAL, LOS PUNTOS DEBEN MOSTRAR ESA TENDENCIA, AUN QUE NO DEBE ESPERARSE QUE LOS PUNTOS SE UBIQUEN EXACTAMENTE EN UNA RECTA.



CECAVA

$\sum(y_i-\bar{y})^2$ = SEC = SUMA DE ERRORES AL CUADRADO

COVARIANZA:

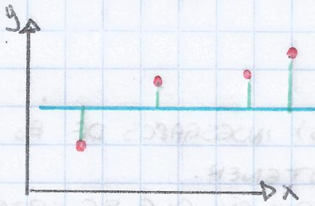
$$Cov = \frac{SS_{xy}}{n} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n}$$

COEFICIENTE DE DETERMINACIÓN

EL COEFICIENTE DE DETERMINACIÓN LINEAL r^2 DE LA MUESTRA ES:

$$r^2 = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}}$$

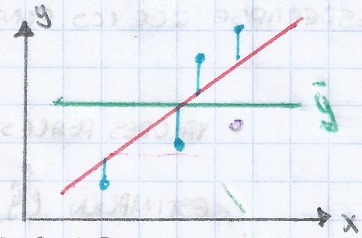
Y REPRESENTA LA PROPORCIÓN DE VARIACIÓN DE Y OBSERVADA QUE SE EXPLICA MEDIANTE EL MODELO DE REGRESIÓN.
SI EL VALOR DE Y ES INDEPENDIENTE DE X, ENTONCES EL VALOR MAS REPRESENTATIVO DE Y SERIA $\bar{y}$, Y PARA CADA VALOR REAL y_i SE OBTENDRIA UN ERROR CON RESPESTO DE $\bar{Y}$.



$\sum(y_i-\bar{y})^2 = SS_{yy}$

=> LA SUMA DE ESTOS ERRORES ESTA DADO POR $\boxed{SS_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2}$

MIENTRAS QUE AL REALIZAR EL AJUSTE A UNA RECTA POR MINIMOS CUADRADOS, EL ERROR SE OBTIENE CON LA RECTA DE AJUSTE $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ Y LA SUMA DE LOS ERRORES AL CUADRADO ES



$$\boxed{SEC = \sum_{i=1}^{n}(y_i-\hat{y})^2}$$

ES CLARO QUE SEC ES UNA PROPOSICIÓN MENOR O IGUAL A UNO, E INDICA LA VARIACIÓN QUE EL MODELO NO ACLARA SS_{yy} O NO EXPLICA. SI SEC/SS_{yy} = 0, ENTONCES SEC = 0 Y LOS PUNTOS EXPERIMENTALES U OBSERVADAS ESTÁN CONTENIDOS TODOS SOBRE LA RECTA DE AJUSTE, POR LO QUE NO EXISTE VARIACIÓN NO EXPLICA.
SI SEC/SS_{yy} = 1 ENTONCES SEC = SS_{yy} Y LA RECTA OBTENIDA ES HORIZONTAL QUE COINCIDE CON $\bar{Y}$. POR LO QUE EL MODELO NO EXPLICA NADA ADICIONAL AL PROMEDIO.

COEFICIENTE DE CORRELACIÓN

EL COEFICIENTE DE CORRELACIÓN PROPORCIONA EL GRADO DE ASOCIACIÓN LINEAL DE LAS VARIABLES X y Y, EN OTRAS PALABRAS. EL COEFICIENTE DE CORRELACIÓN QUE SE ESTUDIA EN ESTE CURSO ES DE TIPO SIMPLE, ES DECIR, CONSIDERA SOLO DOS VARIABLES EN FORMA LINEAL.

EL COEFICIENTE DE CORRELACIÓN DE LA MUESTRA ES.

$$\boxed{r = \frac{SS_{xy}}{\sqrt{SS_{xx}\,SS_{yy}}}}$$

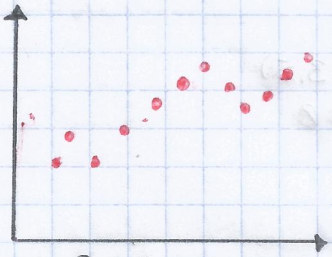
y PROPORCIONA EL GRADO DE ASOCIACIÓN LINEAL DE LAS VARIABLES X y Y

$$\boxed{SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}} \qquad \boxed{SS_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$$

TIPOS DE CORRELACIÓN:

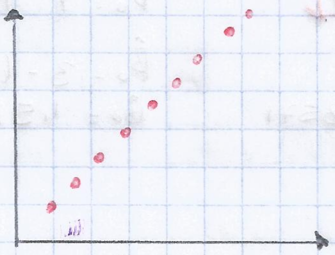
EXISTEN TRES TIPOS DE CORRELACIÓN

CORRELACIÓN DIRECTA O POSITIVA: SE OBTIENE CUANDO AL AUMENTAR (DISMINUIR) EL VALOR DE LA VARIABLE INDEPENDIENTE AUMENTA (DISMINUYE) TAMBIEN EL VALOR DE LA VARIABLE DEPENDIENTE. SI LA CORRELACIÓN TOMA EL VALOR DE UNO SE TIENE CORRELACIÓN POSITIVA PERFECTA.



$0 < r < 1$

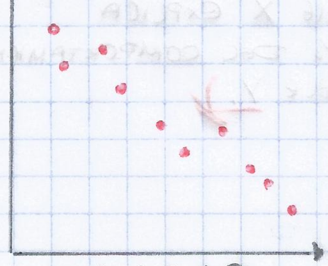
CORRELACIÓN POSITIVA



$r = 1$

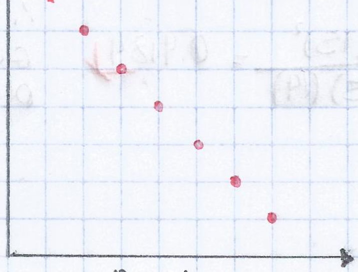
CORRELACIÓN POSITIVA PERFECTA.

CORRELACIÓN INVERSA O NEGATIVA, SE OBTIENE CUANDO AL AUMENTAR (DISMINUIR), EL VALOR DE LA VARIABLE INDEPENDIENTE DISMINUYE (AUMENTA) EL VALOR DE LA VARIABLE DEPENDIENTE. SI LA CORRELACION TOMA EL VALOR DE -1 SE TIENE CORRELACIÓN NEGATIVA PERFECTA.



$-1 < r < 0$

CORRELACIÓN NEGATIVA

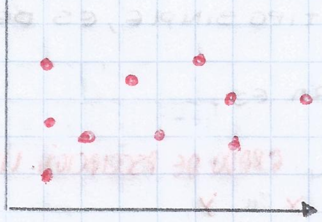


$r = -1$

CORRELACION NEGATIVA PERFECTA.

CELAYA

CORRELACIÓN NULA: SE DA CUANDO NO EXISTE RELACIÓN LINEAL ENTRE VARIABLES.



CORRELACIÓN NULA

$r = 0$

EJEMPLO

EMPLEA EL METODO DE MINIMOS CUADRADOS PARA AJUSTAR LOS SIGUIENTES PUNTOS A UNA RECTA.

X	1	2	3	4	5	6	21
Y	1	2	2	3	5	5	18

XY	1	4	6	12	25	30	78
X^2	1	4	9	16	25	36	91

y^2	1	4	4	9	25	25	68

$n = 6$ $\quad \bar{X} = 3.5$ $\quad \bar{Y} = 3$

a) ¿CUÁLES SON LAS ESTIMACIONES DE $\hat{\beta}_0$ y $\hat{\beta}_1$ DE MINIMOS CUADRADOS?

b) OBTENER EL COEFICIENTE DE CORRELACIÓN E INTERPRETARLO.

$$SS_{xy} = 78 - \frac{(21)(18)}{6} = 15 \qquad SS_{xx} = 91 - \frac{(21)^2}{6} = 17.5$$

$$SS_{yy} = 68 - \frac{(18)^2}{6} = 14 \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_0 = 3 - (0.8571)(3.5)$$

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{15}{17.5} = 0.8571 \qquad \hat{\beta}_0 = 1.5 \times 10^{-4} \approx 0$$

a) $\hat{y} = 0.8571X$

b)

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} = \frac{15}{\sqrt{(14)(17.5)}} = 0.9583$$

∴ DEPENDENCIA FUERTE

c) OBTENER EL COEFICIENTE DE DETERMINACIÓN E INTERPRETARLO.

$$r^2 = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}} = \frac{(15)^2}{(17.5)(14)} = 0.9184$$

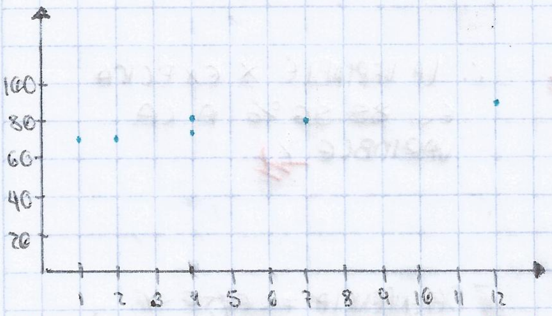
∴ LA VARIABLE X EXPLICA EL 91.84% DEL COMPORTAMIENTO DE LA VARIABLE Y.

LOS SIGUIENTES DATOS REPRESENTAN EL NÚMERO DE HORAS DE ESTUDIO Y LA CALIFICACIÓN OBTENIDA EN UN EXAMEN PARA UNA MUESTRA DE 6 ESTUDIANTES.

ESTUDIANTE	A	B	C	D	E	F	
HORAS	1	2	4	4	7	12	30
CALIFICACIÓN	71	71	74	80	80	86	462

XY	71	142	296	320	560	1032	2421
x^2	1	4	16	16	49	144	230
y^2	5041	5041	5476	6400	6400	7396	35754

a) REPRESENTA LOS DATOS EN UN DIAGRAMA DE DISPERSIÓN.



b) AJUSTAR A LOS DATOS A UN MODELO LINEAL DE REGRESIÓN EMPLEANDO EL CRITERIO DE MÍNIMOS CUADRADOS

$$SS_{xy} = 2421 - \frac{(30)(462)}{6} = 111 \qquad SS_{xx} = 230 - \frac{(30)^2}{6} = 80$$

$$SS_{yy} = 35754 - \frac{(462)^2}{6} = 180 \qquad \bar{X} = \frac{30}{6} \qquad \bar{Y} = \frac{462}{6}$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{111}{80} = 1{,}3875 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_0 = \frac{462}{6} - 1{,}3875\,(30/6) = 70{,}0625$$

$$y = 70.0625 + 1.3875X$$

c) SI ESTUDIA 5 HORAS ¿CUÁL CALIFICACIÓN ESPERARÍA?

$y(5) = 70{,}0625 + 1{,}3875(5)$

$y(5) = 77$

c.2) SI QUIERO OBTENER 85 DE CALIFICACIÓN ¿CUÁNTAS HORAS DEBO ESTUDIAR?

$$X = \frac{y - 70{,}0675}{1{,}3875} \Rightarrow X = \frac{85 - 70.0675}{1{,}3875} = 10{,}7658$$

NOTA: PARA GRAFICAR LA RECTA, USAR β_0 = ORDENADA AL ORIGEN Y EL VALOR MAX. DE X CON LA ECUACIÓN PARA DESPUÉS UNIR LOS PUNTOS.

CECAYA

d) CALCULAR LA COVARIANZA Y EL COEFICIENTE DE DETERMINACIÓN E INTERPRETAR LOS RESULTADOS DE LA RELACIÓN DE VARIABLES.

$$Cov = \frac{SS_{xy}}{n} = \frac{111}{6} = 18.5$$ ∴ DEPENDENCIA DIRECTA

$$r^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} = \frac{(111)^2}{(80)(186)} = 0.8556$$ ∴ LA VARIABLE X EXPLICA EL 85.56 % A LA VARIABLE Y

e) COEFICIENTE DE CORRELACIÓN

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{111}{\sqrt{(80)(186)}} = 0.925$$ ∴ DEPENDENCIA FUERTE Y POSITIVA

CONCEPTOS BASICOS SOBRE ESTIMADORES PUNTUALES.

LOS PRINCIPIOS BASICOS DE LA INFERENCIA ESTADISTICA CONSISTEN EN CREAR METODOS PARA REALIZAR CONCLUSIONES O INFERENCIAS ACERCA DE LA POBLACIÓN.

EXISTEN 2 MÉTODOS (CLASICO Y BOYSEANO)

METODO CLASICO: LA INFERENCIA SE REALIZA POR MÉTODO DE LOS RESULTADOS DE LA MUESTRA ALEATORIA

MÉTODO BOYSANO: LA INFERENCIA SE LLEVO A EFECTO CON BASE EN EL CONOCIMIENTO PREVIO SOBRE LA DISTRIBUCIÓN DE LOS PARAMETROS DESCONOCIDOS O EN LAS VARIABLES ALEATORIAS.

ESPACIO PARAMETRICO

DETERMINA EL CONJUNTO DE VALORES POSIBLES DE UN PÁRAMETRO. SEA $X_1, X_2, \ldots X_n$ UNA MUESTRA ALEATORIA CON FUNCIÓN DE DENSIDAD $f(x;\theta)$ DONDE LA FORMA DE LA FUNCIÓN ES CONOCIDA, POR EL PARAMETRO θ DESCONOCIDO. SOLO SABEMOS QUE PERTENECE AL ESPACIO PARAMÉTRICO DENOTADO POR Ω

EN LA INFERENCIA ESTADISTICA AL HABLAR DEL PÁRAMETRO HACEMOS REFERENCIA A UNA FAMILIA DE DENSIDADES. DEBEMOS ESTUDIAR LA FORMA DE OBTENER INFORMACIÓN DEL PARAMETRO θ DE LA FUNCIÓN DE DENSIDAD DADA.

ESTIMADOR PUNTUAL

CON BASE EN EL VALOR CALCULADO DE LA ESTADISTICA $\overline{X}$ SE PUEDE LLEVAR A EFECTO UNA INFERENCIA DEL PARAMETRO μ; ES DECIR SE PUEDE HACER UNA ESTIMACIÓN PUNTUAL DEL PARAMETRO.

CECAYA

SEA UNA POBLACION CON PARAMETRO θ ; $X_1, X_2, \ldots X_k$ UNA MUESTRA ALEATORIA DE LA POBLACION CON $\theta = u(X_1, X_2 \ldots X_n)$ LA ESTADISTICA CORRESPONDIENTE DE θ EN LA PARTE DE ESTIMACIÓN DE θ MIENTRAS QUE EL VALOR DE θ DE $\hat{\theta}$ OBTENIENDO DE LA REALIZACIÓN DE LA MUESTRA ALEATORIA SE LE LLAMO ESTIMADOR PUNTUAL DE θ.

CON ESTE SE PUEDE CONCLUIR QUE SI θ ES EL PARAMETRO CON ESPACIO PARAMETRICO R CUALQUIER ESTIMADOR PUNTUAL DE θ DEBE ESTAR CONTENIDO EN Ω.
ES DECIR SI $u(X_1, \ldots X_n)$ ES EL ESTIMADOR DEL PARAMETRO SU DISTRIBUCIÓN TENDRA COMO DOMINIO A Ω, LA MAYORIA DE LAS ESTIMACIONES ESTAN RELACIONADAS CON LA MEDIA EN LA VARIANZA μ CON $\bar{X}$ Y σ^2 CON S^2_{n-1}

ESTIMADORES INSESGADOS:

EN LA PROPIEDAD DESEABLE EL ESTIMADOR $\hat{\theta} = u(X_1, \ldots, X_n)$ SE REFIERE CUANDO EL VALOR ESPERADO DE LA DISTRIBUCIÓN DEL ESTIMADOR $\hat{\theta}$ ES IGUAL AL PARAMETRO O PARA PROBAR SI EL PARAMETRO ES INSESGADO UTILIZAMOS LA LINEALIDAD DEL VALOR ESPERADO.

$$E(a_1 X_1 + a_2 X_2 + \ldots + a_n X_n) = a_1 E(X_1) + a_2 E(X_2) + \ldots a_n E(X_n)$$

UN ESTIMADOR $\theta = (X_1, \ldots X_n)$ DE UNO.

FUNCIÓN $g(\hat{\theta})$ DEL PARAMETRO O SE LLAMA ESTIMADOR INSESGADO DE $g(\hat{\theta})$ SI $E(\hat{\theta}) = g(\hat{\theta})$ EN CASO CONTRARIO $E_\theta(\theta) \neq g(\theta)$ SE LLAMA ESTIMADOR SESGADO DE $g(\hat{\theta})$

ESTADISTICOS SUFICIENTES

UNA DE LAS PREOCUPACIONES AL USAR ESTADISTICA ES EN EL CUAL CONCENTRA TODA LA INFORMACIÓN CONTENIDA EN LA MUESTRA, ALEATORIA RESPECTO AL PARAMETRO DE ESTUDIO, ES DECIR SI TENEMOS UNA MUESTR ALEATORIA $X_1, \ldots, X_n$ CON FUNCIÓN DE DENSIDAD CONJUNTO:

$$f(X_1, \ldots, X_n \theta) = \prod_{i=1} f x_i ; \theta$$

LA ESTADISTICA QUE LO BUSCA DEBE BASARSE EN:

A) LA FORMA DE LA DENSIDAD δ_1

B) EL ESPACIO PARAMETRICO Ω

C) LAS OBSERVACIONES DE LA MUESTRA, SI CUMPLE CON ESTAS CARACTERISTICAS SE LE LLAMA ESTADISTICO SUFICIENTE, EL SIGUIENTE TEOREMA NOS AYUDA A DETERMINAR SI TIENE ESTADISTICA SUFICIENTE O NO.

TEOREMA:

CRITERIO DE FACTORIZACIÓN NEYMAN - FISHER

SEA $X_1, \ldots X_n$ UNA MUESTRA ALEATORIA DE UNA POBLACIÓN CON DENSIDAD $f(X; \theta)$ CON $\theta \in \Omega$ Y SEA $T = V(X_1, \ldots, X_n)$ UNA FUNCIÓN DE OBSERVACIONES, ENTONCES T ES UNA ESTADISTICA SUFICIENTE PARA θ BASADO EN LA MUESTRA ALEATORIA $\Longleftrightarrow$ LA DENSIDAD CONJUNTA

$$f(X_1, \ldots, X_n \theta) = f_T(V(X_1, \ldots X_n); \theta) \; H(X_1, \ldots, X_n)$$

DONDE $H(X_1, \ldots, X_n)$ ES UNA FUNCIÓN QUE NO DEPENDE DE θ

COROLARIO:

T ES UN ESTADISTICO SUFICIENTE PARA θ SI

$\Psi_1 : X \Omega \rightarrow t; \; \Psi_2(X_1, \ldots X_n)$ EN DONDE $\Psi_2(X_1, \ldots X_n)$ NO DEPENDE DE θ, ES CONVENIENTE USAR LA FUNCIÓN INDICADA.

CECAYA

$$(ab) = \begin{cases} 1 & a \leq b \\ 0 & a > b \end{cases} \quad \text{ó} \quad I_{ab}(x) = \begin{cases} 1 & a \leq x \leq b \\ 0 & \text{eoc.} \end{cases}$$

PROPIEDAD DE INVARIANZA

SEA $X_1 \ldots X_n$ UNA MUESTRA ALEATORIA DE UNA POBLACIÓN CON DENSIDAD $f(X; \theta \in \Omega \in \mathbb{R})$ Y $T = u(X_1, \ldots X_n)$ UNA ESTADISTICA SUFICIENTE. PARA θ BASADA EN LA MUESTRA ALEATORIA Y SEA $g(\cdot)$ UNA FUNCIÓN INVARIANTE DE f ENTONCES:

A) $g(T)$ ES UNA ESTADISTICA SUFICIENTE PARA θ

B) T ES UNA ESTADISTICA SUFICIENTE PARA $g(\theta)$

PARA DECIR QUE ESTIMADOR ES MAS EFICIENTE SE TENDRA QUE CALCULAR LA VARIANZA POR TANTO

$$V(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1^2(X_1) + a_2^2(X_2) + \ldots + a_n^2(X_n)$$

DADO EL PARAMETRO θ Y UN CONJUNTO DE ESTIMADORES INSESGADOS DE ESTE $\hat{\theta}_1$, $\hat{\theta}_2$ ETC. SE LLAMA ESTIMADOR MAS EFICIENTE RECEPTIVO DE θ RESPECTO DEL CONJUNTO DE ESTIMADORES INSESGADOS AL ESTIMADOR CUYA VARIANZA SEA MENOR.

ERROR CUADRADO MEDIO

SEA $X_1 \ldots X_n$ UNA MUESTRA ALEATORIA DE DENSIDAD

$f(X, \theta; \theta \in \Omega \in \mathbb{R})$ Y $T = u(X_1, \ldots X_n)$ ESTIMADOR DE $g(\theta)$ SE DENOMINA ERROR CUADRADO MEDIO DE $g(\theta)$ RESPECTO DE $T = u(x)$

$E_\theta[T-g(\theta)^2] = \text{ecmo}(\pi)$,

OBSERVE QUE EL ERROR CUADRADO MEDIO SE PARECE AL VALOR ESPERADO DEL CUADRADO DEL SESGO $(E(T)-g(\theta))$ PERO EN LUGAR DEL VALOR ESPERADO TENEMOS AL ESTIMADOR $T-g(\theta)$

SEA $X_1,\dots,X_n$ UNA MUESTRA ALEATORIA DE UNA POBLACIÓN CON PARAMETRO θ, $\theta \in \mathbb{R}$ Y $T=(X_1,\dots X_n$ EL ESTIMADOR DE $g(\theta)$

$$\text{ecmo}(\pi) = V_\theta(T) + S_\theta^2 T$$

DONDE $V_\theta(T)$ Y $S_\theta(\pi)$ SON LA VARIANZA Y EL SESGO DE T CON PARAMETRO θ RESPECTIVAMENTE.

PROPIEDADES ASINTOTICAS DESEABLES DE LOS ESTIMADORES

1. ESTIMADORES CONSISTENTES:

SEA UNA MUESTRA ALEATORIA $X_1, X_2, \dots X_n$ TOMADA DE UNA POBLACIÓN CON PARAMETRO θ DENOTEMOS A UN ESTIMADOR DE $g(\theta)$ POR T_n BASADO EN LA MUESTRA DECIMOS QUE T_n ES UN ESTIMADOR CONSISTENTE PARA $g(\theta)$ SI A MEDIDA QUE AUMENTA 'n' AUMENTA LA PRESICIÓN DE T_n AUMENTA.

POR EJEMPLO:

$\bar{X} \xrightarrow[n\to\infty]{} \mu$ DE MANERA SIMILAR $S^2 \xrightarrow[n\to\infty]{} \sigma^2$

CELAYA

② ESTIMADORES CONSISTENTES EN PROBABILIDAD.

SEA UNA MUESTRA ALEATORIA $X_1, X_2, \ldots, X_n$ TOMADA DE UNA POBLACIÓN CON PARÁMETRO θ Y CONSIDERAMOS UNA SUCESIÓN DE ESTIMADORES

$\{T_n\}_{n \geq 1}$ de $g(\theta)$ BASADOS EN LA MUESTRA DECIMOS QUE LA SUCESIÓN ES CONSISTENTE EN PROBABILIDAD A $g(\theta)$ SI SE CUMPLE LA CONVERGENCIA EN PROBABILIDAD

PARA CUALQUIER $\epsilon > \theta$ SE CUMPLE QUE:

$$\lim_{n \to \infty} ([n - g(\theta)] < \varepsilon) = 1 \quad \forall \ \theta \in \Omega$$

ESTIMACIÓN DE PARAMETROS

PARAMETRO: ESTADISTICO ES UN NÚMERO QUE RESUME EL COMPORTAMIENTO DE UNA VARIABLE ALEATORIA Y QUE DESCRIBE PARCIAL O COMPLETAMENTE SU DISTRIBUCIÓN DE PROBABILIDAD. $(\mu, \sigma, \sigma^2, P)$ P = PROPORCIÓN (POBLACION)

ESTADISTICO: ES UNA FUNCIÓN DE LAS VARIABLES ALEATORIAS QUE SE PUEDEN OBSERVAR EN UNA MUESTRA Y QUE NO DEPENDEN DE PARAMETROS DESCONOCIDOS. TAMBIEN SE LLAMA ESTADÍSTICA O ESTADIGRAFO $(\bar{X}, S_{n-1}, S^2_{n-1}, \hat{p})$ (MUESTRA).

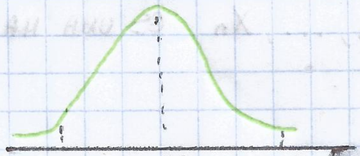
VARIABLES ALEATORIAS: FORMAN UNA MUESTRA ALEATORIA DE TAMAÑO n, SI SON INDEPENDIENTES Y TIENEN LA MISMA DISTRIBUCIÓN DE PROBABILIDAD.

METODOS PARA ESTIMAR PARÁMETROS

- LA ESTIMACIÓN PUNTUAL DE UN PARÁMETRO RELATIVO A UNA POBLACIÓN ES EL VALOR NUMERICO DE UN ESTADISTICO CORRESPONDIENTE A ESE PARÁMETRO.
  $\mu = 20.5$
- LA ESTIMACIÓN POR INTERVALOS ALEATORIOS. UN INTERVALO ALEATORIO ES UN INTERVALO EN DONDE AL MENOS UNO DE SUS LÍMITES ES UNA V.A.
  $\mu = (19.5, 21.5)$
- UNA HIPOTESIS ES UNA AFIRMACIÓN ACERCA DE LA DISTRIBUCIÓN DE PROBABILIDAD DE UNA V.A.
  $H_0 = \mu = 20.5$
  $H_1 = \mu \neq 20.5$

INSESGABILIDAD: CUANDO SE OBTIENE UNA ESTIMACIÓN PUNTUAL DE UN PARÁMETRO CUALQUIERA, ES DESEABLE QUE LA DISTRIBUCIÓN DE DICHA ESTIMACIÓN SE CENTRE EN EL PARÁMETRO REAL (AL CUAL SE LE LLAMARA PARAMETRO OBJETIVO), SI SE CUMPLE LA CONDICIÓN ANTERIOR ENTONCES EL ESTIMADOR SE LLAMA INSESGADO.

ESTIMADOR INSESGADO.



ESTIMADOR SESGADO.

SEA $\hat{\theta}$ UN ESTIMADOR PUNTUAL DEL PARÁMETRO θ. ENTONCES SI $E(\hat{\theta}) = \theta$ SE DICE QUE $\hat{\theta}$ ES UN ESTIMADOR INSESGADO DE θ, DE LO CONTRARIO SE DICE QUE ES SESGADO.

METODO DE MOMENTOS CELAYA

DEFINICIÓN: SEA X UNA VA. Y SEA $k \geq 1$ UN NUMERO ENTERO. EL K-ESIMO MOMENTO DE X, SI EXISTE, ES EL NÚMERO $E(X^k)$

A LOS NÚMEROS $E(X^1), E(X^2), \ldots$ SE LES LLAMA TAMBIEN MOMENTOS POBLACIONALES.

SEA $X_1, \ldots, X_n$ UN MA. DE LA DISTRIBUCIÓN $f(x; \theta)$

DEFINICIÓN: SEA $X_1, \ldots, X_n$ UNA MA. Y SEA $k \geq 1$ UN ENTERO. EL K-ESIMO MOMENTO MUESTRAL ES LA V.A.

$$\frac{1}{n} \sum_{i=1}^{n} X_i^k$$

ESTE METODO CONSISTE EN IGUALAR LOS MOMENTOS POBLACIONALES CON LOS CORRESPONDIENTES MOMENTOS MUESTRALES Y RESOLVER ESTA ECUACION (O SIST. DE ECUACIONES) PARA EL PARAMETRO O VECTOR DE PARAMETROS.

1^{er} MOMENTO POBLACIONAL $E(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$

2^{do} MOMENTO POBLACIONAL $E(x^2) = \frac{1}{n} \sum_{i=1}^{n} x_i^2$

EJEMPLO:

SEA X UNA VA. CON FUNCIÓN DE DENSIDAD EN DONDE $\theta > 0$

$$f(x, \theta) = \begin{cases} \theta \cdot x^{\theta - 1} & \text{si } 0 < x < 1 \\ 0 & \text{eoc.} \end{cases}$$

$\Rightarrow$

$E(x) = \dfrac{\theta}{1+\theta}$ SI $X_1, \ldots, X_n$ ES UNA MA. DE ESTA DISTRIBUCIÓN.

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ 1° MOMENTO MUESTRAL ENTONCES POR EL METODO DE MOMENTOS:

$$\frac{\hat{\theta}}{1+\hat{\theta}} = \bar{x} \Rightarrow \hat{\theta} = \bar{x} / 1 - \bar{x}$$

$\therefore$ ES EL ESTIMADOR PARA θ

METODO DE MAXIMA VEROSIMILITUD.

SEA $(X_1, \ldots, X_n)$ UN VECTOR ALEATORIO CUYA DISTRIBUCIÓN DEPENDE DE PARAMETRO θ

DEFINICIÓN:

LA FUNCIÓN DE VEROSIMILITUD DEL VECTOR $(X_1, \ldots, X_n)$ ES $L(\theta) = f_{X_1, \ldots, X_n}(X_1, \ldots X_n; \theta)$

LA LETRA "L" VIENE DE LIKELIHOOD QUE SE PUEDE TRADUCIR COMO VEROSIMILITUD

SI $X_1, \ldots, X_n$ SON INDEPENDIENTES

$L(\theta) = f_{X_1}(X_1; \theta) \ldots f_{X_n}(X_n; \theta)$

Y CUANDO SON IDENTICAMENTE DISTRIBUIDAS.

$L(\theta) = f(X_1; \theta) \ldots f(X_n; \theta)$

ESTE ES EL CASO DE UNA MUESTRA ALEATORIA. CONSISTE EN OBTENER EL VALOR DE θ QUE MAXIMIZA A LA FUNCIÓN DE VEROSIMILITUD $L(\theta) = f(X_1; \theta), \ldots, f(X_n; \theta)$ AL VALOR θ EN DONDE $L(\theta)$ ALCANZA SU MAXIMO SE LA LLAMA ESTIMACION DE MÁXIMA VEROSIMILITUD O ESTINACIÓN MÁXIMA VEROSÍMIL.

BASICAMENTE LA IDEA ES QUE θ DEBE SER TAL QUE EL VALOR NÚMERICO OSERVADO $(X_1, \ldots, X_n)$ DE LA MA. TENGA LA MÁXIMA PROBABILIDAD

EJEMPLO:

SEA $X_1, \ldots, X_n$ UNA MA. DE LA DISTRIBUCIÓN $EXP(\theta)$

LA FUNSIÓN DE VEROSIMILITUD ES

$$L(\theta) = f(X_1; \theta) \ldots f(X_n; \theta)$$
$$= \theta - e^{-\theta x} \ldots \theta - e^{-\theta X_n}$$
$$= \theta^n \cdot e^{-\theta \bar{x}}$$

$L(\theta)$ ES MÁXIMA EN EL MISMO PUNTO EN DONDE $\ln(L(\theta))$ LO ES.

$\Rightarrow$

$$\ln(L(\theta)) = n \cdot \ln(\theta) - \theta n \bar{x}$$

$$\therefore \frac{d}{d\theta} \ln(L(\theta)) = \frac{n}{\theta} - n\bar{x}$$

$\Rightarrow \theta = 1/x$ ADEMAS ES UN MÁXIMO DADO QUE $\dfrac{d^2}{d\theta^2} \ln(L(\theta)) = \dfrac{-n}{\theta^2} < 0$

$\hat{\theta} = 1/\bar{x}$ ES LA ESTIMACION PARA θ → NÚMERO

$\hat{\theta} = 1/\bar{x}$ ES EL ESTIMADOR PARA θ → ESTADISTICA.

CELAYA

INTERVALOS DE CONFIANZA PARA EL PARÁMETRO MEDIA EN UNA POBLACIÓN NORMAL. $N(\mu, \sigma^2)$

CASO 1: I.C. PARA LA MEDIA DE UNA POBLACIÓN NORMAL CUANDO SE CONOCE LA VARIANZA POBLACIONAL σ

$\bar{X} = 1/n \sum_{i=1}^{n} X_i$ $\quad \bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$ $\quad \bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$ $\quad Z \sim N(0,1)$

$$Z = \frac{\bar{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} \qquad P\left(-z_{\frac{\alpha}{2}} \le \frac{\bar{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P(L_1 \le \mu \le L_2) = 1 - \alpha$$

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \le \mu_x \le \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(-z_{\frac{\alpha}{2}} \le z \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\boxed{\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \le \mu_x \le \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}}}$$

NOTA:
ESTA FORMULA TAMBIEN SE UTILIZA CUANDO SE DESCONOCE LA VARIANZA, PERO EL TAMAÑO DE LA MUESTRA $n > 30$

EJEMPLO:
CONSTRUIR UN INTERVALO DEL 95% DE CONFIANZA PARA LA MEDIA DE UNA POBLACIÓN NORMAL CON VARIANZA 4, Y $n = 10$. ADEMAS, UTILIZANDO LOS DATOS: 4, 7, 5, 10, 23, 17, 9, 15, 7 Y 10 OBTENER UNA ESTIMACIÓN.

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}} \le \mu_x \le \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}}$$

$$\bar{X} = 1/n \sum_{i=1}^{n} X_i = 1/10\,(4+7+5+10+23+17+9+15+7+10) \qquad \bar{X} = 10.7$$

$$\sigma_x = \sqrt{\sigma_x^2} = 2 \qquad n = 10 \qquad 1 - \alpha = 0.95$$

$$P(Z \ge z_{\frac{\alpha}{2}}) = \frac{\alpha}{2} = 0.025$$

$$10.7 - 1.96\left(\frac{2}{\sqrt{10}}\right) \le \mu_x \le 10.7 + 1.96\left(\frac{2}{\sqrt{10}}\right) \Rightarrow 9.46 \le \mu_x \le 11.9$$

S_n = SESGADA.

S_{n-1} = INSESGADA

EN UNA ESCUELA DE INGENIERIA, SE SELECCIONARÓN 50 ALUMNOS Y SE DETERMINO EL PROMEDIO DE HORAS QUE ESTUDIAN A LA SEMANA, EL CUAL FUE DE 3.5 [h] CON UNA DESVIACIÓN ESTANDAR DE LOS DATOS IGUAL A $S_n = \sqrt{3,92}$ [h]. CON ESTOS DATOS, ESTIMAR LAS HORAS QUE ESTUDIARÓN EN PROMEDIO μ_x LOS ALUMNOS CON UN COEFICIENTE DE CONFIANZA IGUAL A 95%.

$n = 50 \qquad \bar{X} = 3.5 \qquad S_n = \sqrt{3,92}$

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}} \leq \mu_x \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}}$$

$\sigma \qquad \sigma$

$n \geq 30$

SOLO MUESTRAS GRANDES.

$1 - \alpha = 0.95 \qquad \alpha = 0.5$

$z_{\frac{\alpha}{2}} = 1.96 \qquad \frac{\alpha}{2} = 0.25$

RECORDANDO

$S_n = \frac{n-1}{n} \sigma^2 \qquad S_{n-1} = \frac{n-1}{n-1} \sigma^2 \Rightarrow$ PARA PASAR DE S_n a S_{n-1}

$S_{n-1} = \frac{n-1}{n} S_{n-1}$

$\Rightarrow S_{n-1} = \frac{n}{n-1} S_n = \frac{50}{49} (\sqrt{3,92}) = 2.02$ → DESVIACIÓN.

ERROR

ERROR

LONGITUD DEL INTERVALO = 2 VECES EL ERROR (ALGURA)

$$\underbrace{3.5 - 1.96 \frac{2}{\sqrt{50}}}_{\text{ERROR}} \leq \mu \leq \underbrace{3.5 + 1.96 \frac{2}{\sqrt{50}}}_{\text{ERROR}}$$

$2.9456 \leq \mu \leq 4.05$

CELAYA

CASO 2. I.C. PARA LA MEDIA DE 1 POBLACIÓN NORMAL CUANDO SE DESCONOCE LA VARIANZA POBLACIONAL σ

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right) \qquad \bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$$

$$Z \sim \cancel{N}(0,1) \qquad T = \frac{\bar{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}}$$

=>

$$\boxed{\bar{X} - t_{\frac{\alpha}{2},(n-1)} \frac{S_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2},(n-1)} \frac{S_{n-1}}{\sqrt{n}}}$$

$$P\left(T \geq t_{\frac{\alpha}{2},(n-1)}\right) = \frac{\alpha}{2}$$

NOTA: ESTA FORMULA SE UTILIZA SIEMPRE QUE SE DESCONOZCA LA VARIANZA POBLACIONAL, PERO A PARTIR DE QUE EL TAMAÑO DE LA MUESTRA $n \geq 30$ CONVERGE CON LA FORMULA DE LA DISTRIBUCIÓN NORMAL ESTANDAR.

EJEMPLO

DIEZ EJES DE PRECISION SON FABRICADOS EN CIERTO PROCESO TIENEN UN DIAMETRO PROMEDIO DE 0.908 cm, CON UNA DESVIACIÓN ESTÁNDAR DE 0.004 cm. CONSIDERANDO QUE LOS DATOS PROVIENEN DE UNA MUESTRA ALEATORIA CON DISTRIBUCIÓN NORMAL, CONSTRUIR UN INTERVALO DE CONFIANZA DEL 95% PARA EL DIAMETRO PROMEDIO REAL DE LOS EJES FABRICADOS.

$n=10 \quad \bar{X} = 0.908 \quad S_{n-1} = 0.004 \quad 1-\alpha = 0.95 \quad t_{\frac{\alpha}{2},(9)} = 2.262$

$\alpha = 0.05$

$\alpha = 0.025$

$$0.908 - 2.262 \frac{(0.004)}{\sqrt{10}} \leq \mu \leq 0.908 + 2.262 \frac{(0.004)}{\sqrt{10}}$$

$$0.905 \leq \mu \leq 0.911$$

PARA CONFIRMAR, SACAR PROMEDIO DE INTERVALO Y ESTE SERÁ $\bar{X}$

DETERMINACIÓN DEL TAMAÑO DE LA MUESTRA.

UNA DE LAS INTERROGANTES MAS COMUNES ES LA DETERMINACIÓN DEL TAMAÑO DE LA MUESTRA CUANDO SE VA A HACER UN MUESTREO, PARA LA ESTIMACIÓN DE UN PARAMETRO. PARA DETERMINAR EL TAMAÑO DE LA MUESTRA SE UTILIZAN LOS CONCEPTOS DE LA ESTIMACIÓN POR INTERVALOS, EN DONDE SE ACEPTA UN "ERROR" SOBRE LA ESTIMACIÓN PUNTUAL, POR EJEMPLO PARA LA MEDIA SE TIENE EL INTERVALO.

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

EN DONDE EL ERROR ES

$$E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

QUE SE SUMA Y SE RESTA AL ESTIMADOR PUNTUAL PARA GENERAR EL INTERVALO DE CONFIANZA. AL FIJAR UN VALOR PARA EL ERROR, Y CONOCIENDO O APROXIMANDO LA DESVIACIÓN ESTANDAR, SE PUEDE DESPEJAR EL TAMAÑO DE LA MUESTRA.

$$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E} \right)^2$$

EL REDONDEO SE HACE SIEMPRE HACIA ARRIBA, PARA ASEGURAR EL NIVEL DE SIGNIFICANCIA.

CASO 1 IC. PARA LA DIFERENCIA DE MEDIAS CUANDO SE CONOCEN LAS VARIANZAS POBLACIONALES σ_1^2 σ_2^2 (2 POBLACIONES).

CUANDO LA MUESTRA ES GRANDE EL ESTADISTICO PARA LA DIFERENCIA DE MEDIAS ESTA DADO POR

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n}}}$$

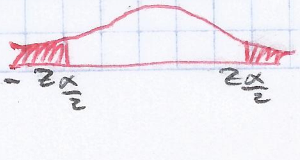
POR LO QUE EL INTERVALO DE CONFIANZA DE DOS LADOS PARA LA DIFERENCIA DE MEDIAS CON UN NIVEL DE CONFIANZA $(1-\alpha) = 100\%$ CENTRADO EN $\mu_1 - \mu_2$ ES.

$$\boxed{\bar{X}_1 - \bar{X}_2 - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

SIEMPRE ES $X_1 - X_2$ O RESPETAR EL ORDEN.

DONDE $z_{\frac{\alpha}{2}}$ SE OBTIENE DE TABLAS CON DISTRIBUCIÓN NORMAL ESTANDAR DE FORMA QUE.

$$P(Z \geq z) = \frac{\alpha}{2}$$



CELAYA

NOTA: SI SE DESCONOCEN LAS VARIANZAS DE LAS POBLACIONES σ_1^2 Y σ_2^2 Y LA MUESTRA ES GRANDE $n \geq 30$ ENTONCES SE PUEDE SUSTITUIR POR SUS ESTIMADORES PUNTUALES $S_{n_1-1}^2$ Y $S_{n_2-1}^2$

EJEMPLO

LA RESISTENCIA DEL CAUCHO A LA ABRASION AUMENTA SI SE AGREGA UNA CARGA LIBRE DE SILICE Y UN AGENTE DE ACOPLAMIENTO PARA ENLAZAR QUIMICAMENTE A LA CARGA CON LAS CADENAS DE POLIMERO DE CAUCHO CON EL AGENTE DE ACOPLAMIENTO TIPO I DIERON UNA RESISTENCIA PROMEDIO DE 92 Y LA VARIANZA DE LAS MEDICIONES FUE DE 20. CUARENTA MUESTRAS DE CAUCHO CON EL AGENTE DE ACOPLAMIENTO DEL TIPO II DIERON UN PROMEDIO DE 98 Y UNA VARIANZA DE 30 EN SUS MEDICIONES. ESTIMAR LA DIFERENCIA VERDADERA ENTRE LAS RESISTENCIAS PROMEDIO A LA ABRASIÓN EN UN INTERVALO DE CONFIANZA DEL 95%.

POBLACION I $n_1 = 50$
$\bar{X}_1 = 92$
$S_{n-1}^2 = 20$

POBLACION II $n_2 = 40$
$\bar{X}_2 = 98$
$S_{n-1}^2 = 30$

$1-\alpha = 0.95$
$\alpha = 0.05$
$\frac{\alpha}{2} = 0.025$

$$\bar{X}_1 - \bar{X}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$92 - 98 - 1.96\sqrt{\frac{20}{50} + \frac{30}{50}} \leq \mu_1 - \mu_2 \leq 92 - 98 + 1.96\sqrt{\frac{20}{50} + \frac{30}{50}}$$

$$(-8.1019 \leq \mu_1 - \mu_2 \leq -3.8981)$$

∴ Se concluye que las medias no son iguales y los signos negativos indican que μ_2 ES MÁS GRANDE QUE μ_1

∴ EL INTERVALO OBTENIDO, $-8.1019 \leq \mu_1 - \mu_2 \leq -3.8981$, TIENE LA INTERPRETACIÓN ADICIONAL DE QUE, AL NO CONTENER EL 0 UNA DE LAS MEDIAS ES MAYOR QUE LAS OTRAS, EN ESTE CASO EL INTERVALO INDICA QUE LA MEDIA μ_2 ES MAYOR QUE LA MEDIA μ_1. POR LO QUE

$$\mu_2 > \mu_1$$

CASO 2 I.C. PARA LA DIFERENCIA DE 2 POBLACIONES NORMALES CUANDO SE DESCONOCE LA VARIANZA POBLACIONAL σ_1^2, σ_2^2 PERO SABE QUE SON IGUALES $\sigma_1^2 = \sigma_2^2$

CUANDO LAS MUESTRAS SON PEQUEÑAS Y PROVIENEN DE POBLACIONES NORMALES CON VARIANZA DESCONOCIDA PERO IGUALES, ENTONCES EL ESTADISTICO PARA LA DIFERENCIA ENTRE DOS MEDIAS ESTA DADO POR:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

DONDE S_p^2 = ESTIMADOR CONVINADO DE LA VARIANZA.

DONDE

$$S_p^2 = \frac{(n_1-1)(S_1^2) + (n_2-1)(S_2^2)}{n_1 + n_2 - 2} \qquad T \sim t_{(n_1+n_2-2)}$$

POR LO QUE EL INTERVALO DE CONFIANZA DE DOS LADOS PARA LA DIFERENCIA DE MEDIAS, CON UN NIVEL DE CONFIANZA $(1-\alpha)100\%$ CENTRADO EN $\mu_1 - \mu_2$ ES

$$\boxed{\bar{X}_1 - \bar{X}_2 - t_{(n_1+n_2-2)} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{(n_1+n_2-2)} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

EJEMPLO

PARA DOS MUESTRAS EXTRAIDAS DE DISTRIBUCIONES NORMALES SE OBTUVIERON LOS SIG. RESULTADOS.

$n_1 = 9$	$n_2 = 7$
$\bar{y}_1 = 43.71$	$\bar{y}_2 = 39.63$
$S_1 = 5.88$	$S_2 = 7.68$

OBTENER EL INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS CON UN COEFICIENTE DE CONFIANZA IGUAL A 0.95. CONSIDERANDO $\sigma_1^2 = \sigma_2^2$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(9-1)(5.88)^2 + (7-1)(7.68)^2}{9+7-2} = 45.035$$

$$S_p = \sqrt{S_p^2} = 6.71 \Rightarrow 43.71 - 39.63 \pm 2.145(6.71)\sqrt{\frac{1}{9} + \frac{1}{7}}$$

$$-3.17 \leq \mu_1 - \mu_2 \leq 11.33$$

$$t_{0.025,(14)} = 2.145$$

CECAVA

1. EL INTERVALO OBTENIDO $-3.14 \le \mu_1 - \mu_2 \le 11.33$, TIENE INTERPRETACIÓN ADICIONAL DE QUE, AL CONTENER EL CERO, SE PUEDE ESTIMAR QUE LAS MEDIAS POBLACIONALES SON IGUALES. EN ESTE CASO EL INTERVALO INDICA QUE LA MEDIA μ_1 ES IGUAL QUE LA MEDIA μ_2 POR LO QUE SE CONCLUYE: $\mu_1 = \mu_2$

CASO 3 SI $\bar{X}_1$ Y $\bar{X}_2$ SON LAS MEDIAS ARITMÉTICAS DE LA REALIZACIÓN DE DOS MUESTRAS ALEATORIAS INDEPENDIENTES DE TAMAÑOS n_1 Y n_2, RESPECTIVAMENTE, DE POBLACIONES APROXIMADAMENTE NORMALES DE LAS QUE SE DESCONOCE σ_1^2, σ_2^2 PERO SE SABE QUE $\sigma_1^2 \neq \sigma_2^2$ EL I.C. DE DOS LADOS PARA LA DIFERENCIA DE MEDIAS CON UN NIVEL DE CONFIANZA $(1-\alpha)100\%$ CENTRADO EN $\mu_1 - \mu_2$ ES:

$$\bar{X}_1 - \bar{X}_2 - t_{(\frac{\alpha}{2}, v)}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{(\frac{\alpha}{2}, v)}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}{\left(\frac{S_1^2}{n_1}\right)^2\left(\frac{1}{n_1 - 1}\right) + \left(\frac{S_2^2}{n_2}\right)^2\left(\frac{1}{n_2 - 1}\right)}$$

NOTA: EN LA FÓRMULA DE GRADOS DE LIBERTAD, SE DEBE REDONDEAR AL ENTERO MÁS PRÓXIMO. POR EJEMPLO 14.5 = 15, 14.3 = 14

EJEMPLO

PARA DOS MUESTRAS EXTRAÍDAS DE DISTRIBUCIONES NORMALES SE OBTUVIERON LOS SIGUIENTES RESULTADOS.

$n_1 = 9$	$n_2 = 7$
$\bar{y}_1 = 43.71$	$\bar{y}_2 = 39.63$
$S_1 = 5.88$	$S_2 = 7.68$

OBTENER EL INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS CON UN COEFICIENTE DE CONFIANZA IGUAL A 0.95 CONSIDERANDO $\sigma_1^2 \neq \sigma_2^2$

$1 - \alpha = 0.95 \quad \frac{\alpha}{2} = 0.025$

$v = 11$

!.... ERROR ----!

$t_{(0.025, 11)} = 2.201 \quad 43.71 - 39.63 \pm 2.201\sqrt{\frac{5.88^2}{9} + \frac{7.68^2}{7}}$

$-3.629 \le \mu_1 - \mu_2 \le 11.789$

∴ EL INTERVALO OBTENIDO, $-3,62 \leq \mu_1 - \mu_2 \leq 11,78$, TIENE LA INTERPRETACIÓN ADICIONAL DE QUE, AL CONTENER EL CERO, SE PUEDE ESTIMAR QUE LAS MEDIAS POBLACIONALES SON IGUALES. EN ESTE CASO EL INTERVALO INDICA QUE LA MEDIA μ_1 ES IGUAL QUE LA MEDIA μ_2 POR LO QUE CONCLUYE LO SIGUIENTE

$$\mu_1 = \mu_2$$

SI X ES UNA VARIABLE ALEATORIA CON MEDIA μ_X Y VARIANZA σ_x^2 ENTONCES EL ESTADÍSTICO EMPLEADO ES

$$\frac{(n-1)S_{n-1}^2}{\chi^2_{\frac{\alpha}{2},(n-1)}} \leq \sigma^2 \leq \frac{(n-1)S_{n-1}^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}} \qquad \chi_2 = \frac{n-1}{\sigma_x^2} S_{n-1}^2$$

$$\chi^2 \sim \chi^2(n-1)$$

PARA CREAR UN INTERVALO DE CONFIANZA PARA UNA PROPORCIÓN SE HACE EL USO DE LA DISTRIBUCIÓN NORMAL ESTANDAR Y QUEDA:

$$\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

MIENTRAS QUE PARA LA DIFERENCIA DE PROPORCIONES.

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} + \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

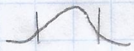
CELAYA

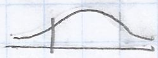
EJEMPLO

Considere los siguientes datos

8.2	8.28	8.24	8.23	8.21	8.25	8.24	8.23
8.24	8.25	8.2	8.26	8.19	8.23	8.26	

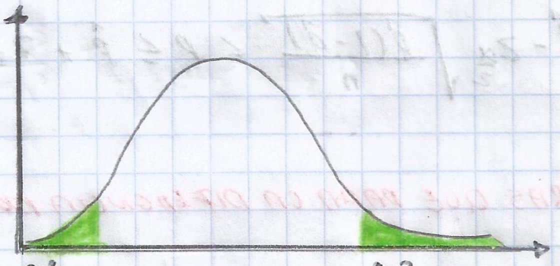
a) Obtener
Un intervalo de confianza de dos lados del 95% para σ_x^2
b) Un intervalo de confianza inferior del 95% para σ_x^2
c) Un intervalo de confianza superior del 95% para σ_x^2

a)  De donde a donde entra la población

b)  (Six Flags Superman altura)

c)  (Mc Donals y juegos infantiles)

$\bar{X} = 8.23$ $\quad S_{n-1} = 0.0253$



$\chi^2_{1-\frac{\alpha}{2}, n-1}$ $\quad \chi^2_{\frac{\alpha}{2}, n-1}$

a)

Usando

$$\frac{(n-1)S_{n-1}^2}{\chi^2_{\frac{\alpha}{2},(n-1)}} \le \sigma_x^2 \le \frac{(n-1)S_{n-1}^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}}$$

$\frac{\alpha}{2} = 0.025$ $\quad 1-\frac{\alpha}{2} = 0.975$ $\quad V = n-1 = 14$

A mayor probabilidad
X será más pequeña

$$\frac{(14)(0.0253)^2}{26.1189} \le \sigma_x^2 \le \frac{(14)(0.0253)^2}{5.63}$$

$\chi^2_{0.975, 14} = 5.63$

$\chi^2_{0.025, 14} = 26.12$

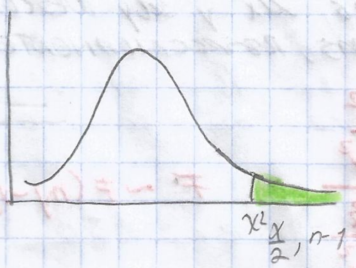
$$0.0003438 \le \sigma_x^2 \le 0.0015916$$

$$0.01854 \le \sigma_x \le 0.039$$

b)

$\alpha = 0,05$
$V = 14$

$\chi^2_{0.05,14} = 23.68$

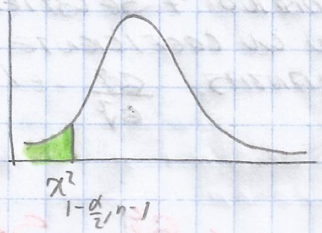


$$\frac{(n-1)S_{n-1}^2}{\chi^2_{\alpha,n-1}} \le \sigma^2_x$$

$$\frac{(14)(0,0253)^2}{23,68} \le \sigma^2$$

$$0,0003784 \le \sigma_x^2$$

c) $1-\alpha = 0.95$ $\quad V = 14$



$$\sigma_x^2 \le \frac{(n-1)S_{n-1}^2}{\chi^2_{1-\alpha,(n-1)}}$$

$$\sigma_x^2 \le \frac{(14)(0,0253)^2}{6,57}$$

$$\sigma_x^2 \le 0,0013639$$

CELAYA

INTERVALO DE CONFIANZA PARA LA RAZÓN DE VARIANZAS.

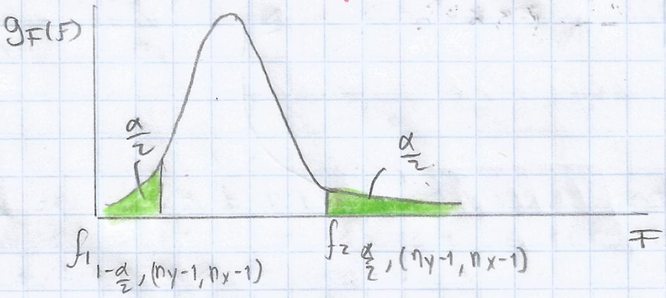
SI X Y Y SON VV. AA. INDEPENDIENTES CON DISTRIBUCIONES NORMALES CON MEDIAS μ_X Y μ_Y DESCONOCIDAS Y VARIANZAS σ_X^2 Y σ_Y^2 DESCONOCIDAS, RESPECTIVAMENTE, ENTONCES EL ESTADÍSTICO EMPLEADO ES.

$$F = \frac{\frac{S_y^2}{\sigma_y^2}}{\frac{S_x^2}{\sigma_x^2}} \qquad F \sim F(n_y-1,\ n_x-1)$$

DONDE

UTILIZANDO EL ESTADÍSTICO F SE OBTIENE EL INTERVALO DE CONFIANZA DE DOS LADOS CON UN COEFICIENTE $(1-\alpha)100\%$ PARA LA RELACIÓN DE VARIANZAS $\frac{\sigma_x^2}{\sigma_y^2}$ EL CUAL ES:

$$\frac{S_x^2}{S_y^2} F_{1-\frac{\alpha}{2},(n_y-1,\ n_x-1)} \leq \frac{\sigma_x^2}{\sigma_y^2} \leq \frac{S_x^2}{S_y^2} F_{\frac{\alpha}{2},(n_y-1,\ n_x-1)}$$



O BIEN UTILIZANDO EL RECÍPROCO

$$\frac{S_x^2}{S_y^2} \frac{1}{F_{\frac{\alpha}{2},(n_x-1,\ n_y-1)}} \leq \frac{\sigma_x^2}{\sigma_y^2} \leq \frac{S_x^2}{S_y^2} \frac{1}{F_{1-\frac{\alpha}{2}(n_x-1,\ n_y-1)}}$$

$\frac{\alpha}{2}$ = LO VOY A ENCONTRAR EN TABLAS

$1-\frac{\alpha}{2}$ = SERÁ DIFIC ENCONTRAR

SE EXTRAEN DOS MUESTRAS INDEPENDIENTES DE POBLACIONES NORMALES OBTENIENDOSE LOS SIG. DATOS.

$n_1 = 15$ $\quad$ $n_2 = 10$

$\bar{X}_1 = 300$ $\quad$ $\bar{X}_2 = 325$

$S_1^2 = 16$ $\quad$ $S_2^2 = 49$

CONSTRUIR UN INTERVALO DE CONFIANZA DE DOS LADOS DEL 95% CON RESPECTO A LA RELACIÓN DE LAS VARIANZAS $\frac{\sigma_1^2}{\sigma_2^2}$

$$\frac{S_1^2}{S_2^2}\frac{1}{f_{\frac{\alpha}{2},(n_1-1,\,n_2-1)}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} f_{\frac{\alpha}{2}(n_2-1,\,n_1-1)}$$

$$\frac{1}{f_{0.025,(14,9)}} = \frac{1}{3.8} = 0.263 \qquad f_{0.0025,(9,14)} = 3.209$$

$$\frac{16}{49}(0.263) \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{16}{49}(3.209)$$

$$0.085 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq 1.048$$

∴ EL INTERVALO OBTENIDO, $0.085 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq 1.048$ TIENE LA INTERPRETACIÓN ADICIONAL DE QUE, AL CONTENER EL UNO SE PUEDE ESTIMAR QUE LAS VARIANZAS POBLACIONALES SON IGUALES. EN ESTE CASO EL INTERVALO INDICA QUE LA VARIANZA σ_1^2 ES IGUAL QUE LA VARIANZA σ_2^2. POR LO QUE SE CONCLUYE LO SIGUIENTE:

$$\sigma_1^2 = \sigma_2^2$$

CELAYA

INTERVALOS DE CONFIANZA PARA UNA PROPORCIÓN

SI SE TOMA UNA MUESTRA DE TAMAÑO n DE UNA POBLACIÓN MUY GRANDE (O INFINITA), Y X OBSERVACIONES PERTENESEN A LA CLASE DE INTERES ENTONCES $\hat{p} = \frac{X}{n}$ ES UN ESTIMADOR PUNTUAL DE LA PROPORCIÓN DE LA POBLACIÓN QUE PERTENECE A LA CLASE EN CUESTIÓN, Y LA DISTRIBUCIÓN DE MUESTREO ES.

$$\hat{p} = \frac{X \rightarrow \text{EXITO}}{n \rightarrow \text{CASOS POSIBLES}} \qquad Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

DONDE $Z \sim N(0,1)$ NORMAL ESTANDAR.

Y n Y p SON LOS PARAMETROS DE LA DISTRIBUCIÓN BINOMIAL

UTILIZANDO EL ESTADISTICO Z Y APROXIMANDO LA CANTIDAD $p(1-p)$ MEDIANTE EL ESTIMADOR PUNTUAL $\hat{p}(1-\hat{p})$ SE OBTIENE EL INTERVALO DE CONFIANZA DE DOS LADOS CON UN COEFICIENTE $(1-\alpha)100\%$ PARA LA PROPORCIÓN p ES.

$$\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ESTADISTICO

EJEMPLO

EN UNA MUESTRA AL AZAR DE 60 SECCIONES DE TUBO EN UNA PLANTA QUIMICA, 8 DE ELLOS MOSTRARON SEÑALES DE CORROSIÓN SERIA CONSTRUIR UN INTERVALO DE CONFIANZA DEL 95% PARA LA PROPORCIÓN DE LOS TRAMOS DE TUBO CON CORROSIÓN SERIA

$$\hat{p} = \frac{X}{n} = \frac{8}{60} \qquad Z_{\frac{\alpha}{2}} = Z_{0{,}025} = 1{,}96$$

$$\frac{8}{60} - 1{,}96\sqrt{\frac{\frac{8}{60}\left(1-\frac{8}{60}\right)}{60}} \leq p \leq \frac{8}{60} + 1{,}96\sqrt{\frac{\frac{8}{60}\left(1-\frac{8}{60}\right)}{60}}$$

$$0{,}04731 \leq p \leq 0.21939$$

IMPORTANTE:

$$4.7\% \leq p \leq 21.93\%$$

LA PROPORCIÓN NO PUEDE SER NEGATIVA

INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE PROPORCIONES

SI DOS MUESTRAS INDEPENDIENTES DE TAMAÑO n_x Y n_y SE EXTRAEN DE POBLACIONES INFINITAS CON DISTRIBUCIÓN DE BERNOULLI, X REPRESENTA EL NÚMERO DE OBSERVACIONES DE LA PRIMERA MUESTRA QUE CORRESPONDE A LA CLASE DE INTERES, y Y REPRESENTA EL NÚMERO DE OBSERVACIONES DE LA SEGUNDA MUESTRA QUE CORRESPONDE A LA CLASE EN CUESTIÓN, ENTONCES LA DISTRIBUCIÓN DE MUESTREO PARA LA DIFERENCIA DE PROPORCIONES ESTA DADA POR:

$$z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}}}$$

DONDE:

$z \sim N(0,1)$

DE LA FORMA ANTERIOR SE OBTIENE EL INTERVALO DE CONFIANZA DE DOS LADOS CON UN COEFICIENTE $(1-\alpha)100\%$ PARA LA DIFERENCIA DE PROPORCIONES $p_1 - p_2$ EL CUAL ES:

$$(\hat{p}_1 - \hat{p}_2) - \underbrace{z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}_{\text{ERROR}} \le p_1 - p_2 \le (\hat{p}_1 - \hat{p}_2) + \underbrace{z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}_{\text{ERROR}}$$

CELAYA

EJEMPLO

DOS GRUPOS DE 80 PACIENTES TOMARON PARTE DE UN EXPERIMENTO EN EL CUAL UN GRUPO RECIBIO PILDORAS QUE CONTENIAN UN ANTIALERGICO, MIENTRAS QUE OTRO GRUPO SE LE ADMINISTRO UN PLACEBO, ES DECIR, UNA PILDORA SIN DROGA ALGUNA EN EL GRUPO QUE RECIBIO EL MEDICAMENTO 23 EXHIBIERON SINTOMAS ALERGICOS. MIENTRAS QUE EN EL OTRO GRUPO 41 LOS EXHIBIERON. OBTENER EL INTERVALO DE CONFIANZA DEL 99% PARA LA DIFERENCIA DE PROPORCIONES.

$$\hat{P}_1 = \frac{X}{n_1} = \frac{23}{80} \qquad \hat{P}_2 = \frac{Y}{n_2} = \frac{41}{80} \qquad z_{\frac{\alpha}{2}} = z_{0.005} = 2.575$$

$$\left(\frac{23}{80} - \frac{41}{80}\right) \pm 2.575 \sqrt{\frac{\left(\frac{23}{80}\right)\left(1 - \frac{23}{80}\right)}{80} + \frac{\left(41/80\right)\left(1 - 41/80\right)}{80}}$$

$$-0.4119 \leq P_1 - P_2 \leq -0.031$$

SON DIFERENTES Y P_2 ES MAYOR QUE P_1 POR NO CONTENER AL 0

Hipotesis. Es una afirmación acerca de la distribución de probabilidad de una V.A.

Errores de tipo I y tipo II

Las pruebas de hipótesis son parte de la inferencia estadistica y a menudo involucran a mas de un parámetro de la distribución.

Supongase por ejemplo que se desea estimar el promedio de la estatura de los alumnos de la F.I., y se pretende saber si el promedio es 1.67 o no lo es. Lo anterior se expresaria:

La hipotesis es el parametro

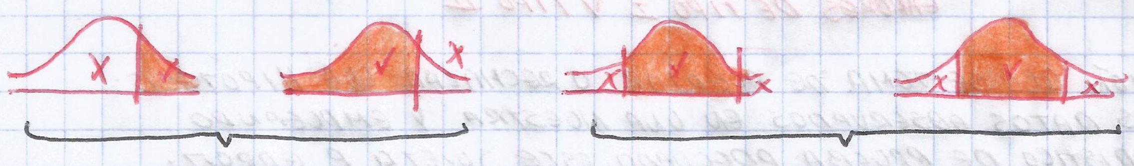
$H_0: \mu = 1,67\,[m]$ **Hipotesis nula** (siempre va el igual)

$H_1: \mu \neq 1.67\,[m]$ **Hipotesis alterna**

Contrastes de hipotesis

Existen 4 tipos de hipotesis, los cuales son:

①	②	③	④
$H_0: \theta \geq \theta_0$	$H_0: \theta \leq \theta_0$	$H_0: \theta = \theta_0$	$H_0: \theta \in (\theta_1, \theta_2)$
$H_1: \theta < \theta_0$	$H_1: \theta > \theta_0$	$H_1: \theta \neq \theta_0$	$H_1: \theta \notin (\theta_1, \theta_2)$



Prueba alternativa de un lado.

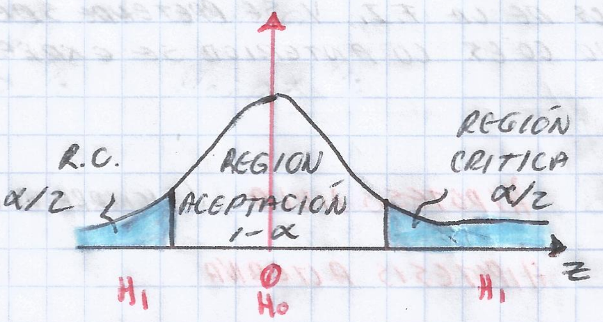
Prueba alternativa de dos lados.

CELAYA

REGIONES DE ACEPTACIÓN O RECHAZO.

PARA PROBAR UNA HIPÓTESIS ES NECESARIO SELECCIONAR UNA MUESTRA ALEATORIA, Y MEDIANTE UN ESTADÍSTICO DE PRUEBA ADECUADO DETERMINAR SI SE ACEPTA LA HIPÓTESIS H_0 O SE RECHAZA, ACEPTÁNDOSE LA ALTERNATIVA H_1. CON LA FINALIDAD DE ACEPTAR O RECHAZAR UNA HIPÓTESIS, DEBEN GENERARSE REGIONES DE ACEPTACIÓN Y RECHAZO, POR EJEMPLO, POR EJEMPLO PARA LA HIPÓTESIS SOBRE LA MEDIA POBLACIONAL PLANTEADA ANTERIORMENTE SE TIENE.

H_0 ES DONDE ESTA $1-\alpha$



UNA PRUEBA DE HIPÓTESIS PARA ALGUNA CARACTERÍSTICA DESCONOCIDA DE UNA POBLACIÓN ES CUALQUIER REGLA QUE PERMITE RECHAZAR O NO RECHAZAR UNA HIPÓTESIS NULA CON BASE A UNA MUESTRA ALEATORIA DE LA POBLACIÓN.

$$\bar{z} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \qquad \bar{x} = \frac{1}{n}\sum x_i \qquad \mu = \text{PROPOSICIÓN}$$

ERRORES DE TIPO I Y TIPO II

LA DESICIÓN QUE SE TOMA DE ACEPTAR O RECHAZAR UNA HIPÓTESIS SEGÚN LOS DATOS OBSERVADOS EN UNA MUESTRA Y EMPLEANDO UN ESTADÍSTICO DE PRUEBA ADECUADO, ESTE SUJETA A ERROR. EN PARTICULAR SE PUEDE COMETER DOS TIPOS DE ERRORES.

1) CUANDO LA HIPÓTESIS NULA H_0 SE RECHAZA SIENDO QUE ES VERDADERA SE COMETE UN ERROR DE TIPO I

2) MIENTRAS QUE SI SE ACEPTA LA HIPÓTESIS NULA H_0 CUANDO ES FALSA ENTONCES SE COMETE UN ERROR DEL TIPO II.

TABLA 1 TIPOS DE ERROR EN LAS PRUEBAS DE HIPÓTESIS.

	Si la hipótesis es / y la conclusión es	Estado de la naturaleza: H_0 es verdadera	H_0 es falsa
Desición	No se rechaza H_0	Ningun error	Error tipo II
	Si se rechaza H_0	Error tipo I	Ningun error

Las probabilidades de cometer errores del tipo I y II se denotan mediante α y β respectivamente, es decir.

$$P(\text{rechazar } H_0 \mid H_0 \text{ es verdadera}) = P(\text{error tipo I}) = \alpha$$

$$P(\text{aceptar } H_0 \mid H_0 \text{ es falsa}) = P(\text{error tipo II}) = \beta$$

TABLA 2. PROBABILIDAD DE ERROR EN LAS PRUEBAS DE HIPÓTESIS.

Si la hipótesis es / y la conclusión es	H_0 es verdadera	H_0 es falsa
No se rechaza H_0	$1-\alpha$	β
Si se rechaza H_0	α	$1-\beta$

La potencia de una prueba de hipótesis estadística es la probabilidad de rechazar una hipótesis falsa, es decir

$$\text{Potencia de una prueba} = P(\text{rechazar } H_0 \mid H_0 \text{ falsa})$$

CELAYA

PRUEBAS DE HIPOTESIS.

ADEMAS DE LAS PRUEBAS DE HIPÓTESIS UNILATERALES Y BILATERALES COMO FUERON LAS PRIMERAS ECUACIONES, LAS PRUEBAS SE CLASIFICAN EN SIMPLES Y COMPUESTAS.

LAS HIPOTESIS SIMPLES SON AQUELLAS QUE ESPECIFICAN EL VALOR DEL PARAMETRO A L QUE SE REFIEREN, POR EJEMPLO:

$H_0: p = \frac{1}{2}$ → PROPORCIÓN

LAS HIPOTESIS COMPUESTAS SON AQUELLAS QUE NO ESPECIFICAN EL VALOR DEL PARAMETRO, POR EJEMPLO:

$H_0: p > 1/2$ $\quad H_0: p \neq 1/2$

EJERCICIO

EL TIEMPO TRANSCURRIDO x ENTRE DOS SEÑALES CONSECUTIVAS DE UN CONTADOR Geiger DE PÁRTICULAS RADIOACTIVAS, ES UNA V.A. CON DISTRIBUCIÓN EXPONENCIAL CON PARÁMETRO λ

$$f_x(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{eoc.} \end{cases}$$

A FIN DE PROBAR LA HIPOTESIS H_0 DE QUE PARA UN MATERIAL EN PARTICULAR $\lambda = 2$, CONTRA LA ALTERNATIVA H_1 DE $\lambda = 1$, SE REALIZA UNA SOLA OBSERVACIÓN DE X Y SE DECIDE NO RECHAZAR H_0 ⇔ EL VALOR OBSERVADO DE X OCURRE EN EL INTERVALO $[0, 1]$. CALCULAR LOS TAMAÑOS DE LOS ERRORES TIPO I Y II.

$H_0: \lambda = 2$ PARA NO RECHAZAR TIENE $[0, 1]$

$H_1: \lambda = 1$

PARA TIPO I

$$\alpha = P(X > 1 / \lambda = 2) = \int_1^\infty 2e^{-2x} dx = e^{-2} \Rightarrow \alpha = e^{-2} \approx 0.135$$

PARA TIPO II

$$\beta = P(0 \leq X \leq 1 / \lambda = 1) = \int_0^1 e^{-x} dx = 1 - e^{-1} \qquad \beta = 1 - e^{-1} \approx 0.632$$

PRUEBAS DE HIPÓTESIS PARA EL PARÁMETRO DE LA MEDIA

CASO 1

μ

σ

CASO 2

μ

σ

Dist Z: P.H PARA LA MEDIA DE 1 POBLACIÓN NORMAL CUANDO SE CONOCE LA VARIANZA POBLACIONAL σ^2

P.H. PARA LA MEDIA DE 1 POBLACIÓN NORMAL CUANDO SE DESCONOCE LA VARIANZA POBLACIONAL σ^2 — DISTRIBUCIÓN t

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$T = \frac{\bar{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}}$$

CASO 1 P.H. PARA LA MEDIA DE 1 POBLACIÓN NORMAL CUANDO SE CONOCE LA VARIANZA POBLACIONAL σ^2

CUANDO SE DESEA REALIZAR UNA HIPÓTESIS CON RESPECTO A LA MEDIA DE UNA VARIABLE ALEATORIA X, DEBE SUPONER CON DISTRIBUCIÓN NORMAL, YA SEA PORQUE X SE DISTRIBUYE NORMALMENTE O POR EL TEOREMA DEL LÍMITE CENTRAL. SI SE CONSIDERA QUE LA MEDIA μ SE DESCONOCE PERO SE CONOCE LA VARIANZA σ^2, ENTONCES LA HIPÓTESIS BILATERAL PUEDE FORMULARSE COMO:

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

DONDE μ_0 ES UNA CONSTANTE ESPECÍFICA, Y EL ESTADÍSTICO DE PRUEBA ES

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$ DONDE $Z \sim N(0, 1)$

CECAYA

EJEMPLO

CONSIDÉRESE UNA POBLACIÓN CON DISTRIBUCIÓN NORMAL CON PARÁMETROS μ DESCONOCIDO Y $\sigma^2 = 4$. CON BASE EN UNA MUESTRA DE 30 OBSERVACIONES EN LA CUAL $\bar{x} = 10$ Y $S^2 = 3.5$, DETERMINAR SI ES CORRECTO SUPONER QUE $\mu = 12$ CON UN NIVEL DE SIGNIFICANCIA de 0.01

LA PRUEBA DE HIPÓTESIS ESTADÍSTICA ES:

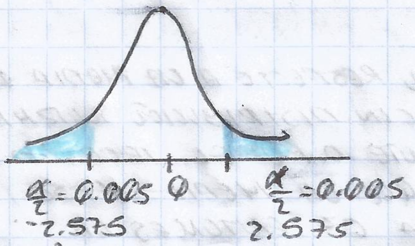
$H_0: \mu = 12$ ✗

$H_1: \mu \neq 12$ ✓

EL ESTADÍSTICO DE PRUEBA ES $\rightarrow$ ESTANDARIZANDO

$$\bar{X} \sim N(\mu, 4/30) \qquad z_0 = \frac{\bar{X} - \mu_0}{\sqrt{\frac{4}{30}}} \sim N(0,1)$$

Y LAS REGIONES CRÍTICAS DE ACEPTACIÓN



$P(Z > z) = \frac{\alpha}{2}$

$z = 2.575$

Y EVALUANDO EL ESTADÍSTICO DE PRUEBA PARA LA MUESTRA DADA Y SUPONIENDO CIERTA H_0 SE TIENE.

$$z_0 = \frac{\bar{x} - \mu_0}{\sqrt{\frac{4}{30}}} = \frac{10 - 12}{\sqrt{\frac{4}{30}}} = -5.47$$

DE DONDE SE OBSERVA QUE EL ESTADÍSTICO SE ENCUENTRA FUERA DE LA REGIÓN DE ACEPTACIÓN $-5.47 < -2.575$ $(z_0 < -z)$

$\therefore$ SE CONCLUYE QUE, CON BASE EN ESTA MUESTRA, NO PARECE ADECUADO SUPONER $\mu = 12$ POR LO QUE H_0 SE RECHAZA

CASO 2 P.H. PARA LA MEDIA DE UNA POBLACIÓN NORMAL CUANDO SE DESCONOCE LA VARIANZA POBLACIONAL σ^2

CUANDO EN LA PRÁCTICA NO SE CONOCE LA VARIANZA POBLACIONAL σ^2, PUEDE SUSTITUIRSE SU VALOR POR S_{n-1}^2 SI LA MUESTRA ES GRANDE ($n \geq 30$) SIN TENER UN EFECTO PERJUDICIAL DE CONSIDERACIÓN.
SI LA VARIANZA SE DESCONOCE Y LA MUESTRA ES PEQUEÑA ($n < 30$) ENTONCES EL ESTADÍSTICO DE PRUEBA ES:

$n \geq 30$ σ^2 DESCONOCIDA.

$$Z = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

$n < 30$ σ^2 DESCONOCIDA

$$T_0 = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

SIEMPRE QUE LA POBLACIÓN TENGA DIST. NORMAL.

PRUEBAS DE HIPÓTESIS SOBRE LA IGUALDAD DE DOS MEDIAS

CASO 1

$\mu_1 = \mu_2$
σ_1^2, σ_2^2

P.H SOBRE LA IGUALDAD DE MEDIAS DE 2 POBLACIONES NORMALES CUANDO SE CONOCEN LAS VARIANZAS POBLACIONALES σ_1^2, σ_2^2

Z_0 (MUESTRA GRANDE)

CASO 2

$\mu_1 = \mu_2$
$\sigma_1^2 = \sigma_2^2$

P.H SOBRE LA IGUALDAD DE MEDIAS DE 2 POBLACIONES NORMALES CUANDO SE DESCONOCEN LAS VARIANZAS POBLACIONALES, PERO SE SABE QUE SON IGUALES. $\sigma_1^2 = \sigma_2^2$

$\frac{\sigma_1^2}{\sigma_2^2} = 1$

T_0

CASO 3

$\mu_1 = \mu_2$
$\sigma_1^2 \neq \sigma_2^2$

P.H. SOBRE LA IGUALDAD DE MEDIAS DE 2 POBLACIONES NORMALES CUANDO SE DESCONOCEN LAS VARIANZAS POBLACIONALES, PERO SE SABE QUE SON DIFERENTES.

$\frac{\sigma_1^2}{\sigma_2^2} \neq 1$

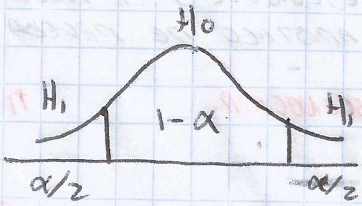
T_0^*

·CELAYA

CASO 1: PH SOBRE IGUALDAD DE DOS MEDIAS POBLACIONALES NORMALES CUANDO SE CONOCEN LAS VARIANZAS POBLACIONALES σ_1^2, σ_2^2

SI SE DESEA PROBAR QUE LAS MEDIAS DE DOS POBLACIONES (CON DISTRIBUCIONES NORMALES) SON IGUALES, ENTONCES EL ESTADISTICO DE PRUEBA ES;

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n}}}$$



LA PRUEBA CON ALTERNATIVA DE DOS LADOS ES:

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

CUANDO LAS VARIANZAS POBLACIONALES SE DESCONOCEN, SE PUEDE UTILIZAR LAS VARIANZAS MUESTRALES PARA LAS POBLACIONES, SIEMPRE QUE LAS MUESTRAS SEAN GRANDES.

SI SE DESEA PROBAR LA DIFERENCIA DE MEDIAS, ENTONCES EL ESTADISTICO SE MODIFICA RESTANDO LA DIFERENCIA DE MEDIAS.

(AQUI YO PROPONES LA $\mu_1 - \mu_2$).

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

CASO 2 P.H. SOBRE IGUALDAD DE MEDIAS DE 2 POBLACIONES NORMALES CUANDO SE DESCONOCEN LAS VARIANZAS POBLACIONALES σ_1^2, σ_2^2 PERO SE SABE QUE SON IGUALES. $\sigma_1^2 = \sigma_2^2$

MUESTRAS PEQUEÑAS DE POBLACIONES NORMALES Y VARIANZAS DESCONOCIDAS PERO IGUALES.

EL ESTADÍSTICO DE PRUEBA ES:

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

DONDE

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \quad \text{y} \quad T_0 \sim t_{(n_1+n_2-2)}$$

CASO 3 P.H. SOBRE IGUALDAD DE MEDIAS DE 2 POBLACIONES NORMALES CUANDO SE DESCONOCEN LAS VARIANZAS POBLACIONALES σ_1^2, σ_2^2 PERO SE SABE QUE SON DIFERENTES $\sigma_1^2 \neq \sigma_2^2$.

MUESTRAS PEQUEÑAS DE POBLACIONES NORMALES Y VARIANZAS DESCONOCIDAS PERO DIFERENTES.

CUANDO LAS VARIANZAS SON DIFERENTES, ENTONCES NO EXISTE UN ESTADÍSTICO t EXACTO PARA REALIZAR LA PRUEBA SOBRE LA IGUALDAD DE MEDIAS! SIN EMBARGO, UNA BUENA APROXIMACIÓN LA PROPORCIONA EL ESTADÍSTICO.

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

EL CUAL TIENE DISTRIBUCIÓN APROXIMADAMENTE t, i.e. $T_0^* \sim t_{(v)}$: DONDE EL NUMERADOR DE GRADOS DE LIBERTAD ESTÁ DADO POR:

$$v \approx \frac{\left[\left(\frac{S_1^2}{n_1}\right) + \left(\frac{S_2^2}{n_2}\right)\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2+1}} - 2$$

Y DEBE UTILIZARSE EL ENTERO MÁS CERCANO.

CELAYA

EJERCICIO IGUALDAD DE MEDIAS

MEDICIONES RESPECTO AL ESFUERZO CORTANTE OBTENIDAS A PARTIR DE PRUEBAS DE COMPRESIÓN INDEPENDIENTES PARA DOS TIPOS DE SUELOS DIERON LOS RESULTADOS SIGUIENTES (MEDICIONES EN TONELADA POR METRO CUADRADO).

SUELO TIPO I

$n_1 = 36$
$\bar{y}_1 = 1.65$
$s_1 = 0.26$

SUELO TIPO II

$n_2 = 35$
$\bar{y}_2 = 1.43$
$s_2 = 0.22$

¿DIFIEREN LOS DOS SUELOS CON RESPECTO AL ESFUERZO CORTANTE PROMEDIO, A NIVEL DE SIGNIFICACIÓN DE 1%?

PRIMERO PLANTEAR HIPÓTESIS $H_0: \mu_1 - \mu_2 = 0$ ✗ $H_1: \mu_1 - \mu_2 \neq 0$ ✓

SEGUNDO ¿CUÁL ESTADÍSTICO DEBEMOS USAR?

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

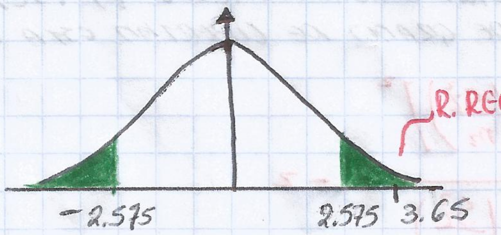
EL ESTADÍSTICO DE PRUEBA ES

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

VALUANDO

$$z_0 = \frac{1.65 - 1.43}{\sqrt{\frac{(0.26)^2}{36} + \frac{(0.22)}{35}}} = 3.65$$

CON $\alpha = 0.01$ $P(Z > z) = \alpha/2$ $P(Z > z) = 0.005$ $z = 2.575$



PRUEBAS DE HIPOTESIS PARA LA VARIANZA

SI SE DESEA PROBAR LA VARIANZA DE UNA POBLACIÓN CON DISTRIBUCIÓN NORMAL, ENTONCES EL ESTADISTICO DE PRUEBA ES:

$$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

DONDE S^2 ES LA VARIANZA MUESTRAL Y $\chi^2 \sim \chi^2_{(n-1)}$

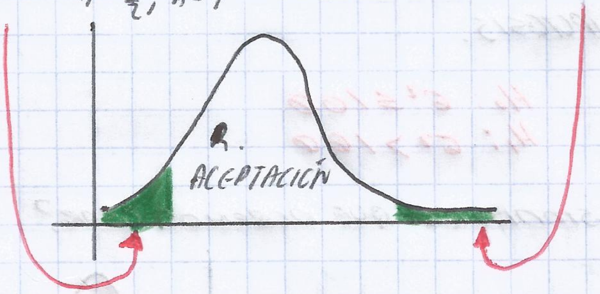
LA PRUEBA DE HIPOTESIS DE DOS LADOS ES:

$H_0: \sigma^2 = \sigma_0^2$

$H_1: \sigma^2 \neq \sigma_0^2$

Y LA HIPOTESIS NULA SE RECHAZA SI $\chi_0^2 > \chi^2_{\frac{\alpha}{2}, n-1}$

O BIEN $\chi_0^2 < \chi^2_{1-\frac{\alpha}{2}, n-1}$



CELAYA.

EJERCICIO:

LA DISPERSIÓN O VARIANZA, DE TIEMPO DE ACARREO EN UN PROYECTO DE CONSTRUCCIÓN SON DE GRAN IMPORTANCIA PARA EL SOBRESTANTE, YA QUE LOS TIEMPOS MUY VARIABLES DE ACARREO ORIGINAN PROBLEMAS EN LA PROGRAMACIÓN DE LOS TRABAJOS. EL ENCARGADO DE LOS TRASPORTES DICE QUE EL INTERVALO DE TIEMPO DE ACARREO NO DEBE SER MAYOR QUE 40 MINUTOS (ESTE INTERVALO ES LA DIFERENCIA ENTRE EL TIEMPO MAYOR Y EL MENOR). SI SE SUPONE QUE ESTOS TIEMPOS DE ACARREO ESTAN DISTRIBUIDOS EN FORMA APROXIMADAMENTE NORMAL, EL SOBRESTANTE CREE QUE LA AFIRMACIÓN ACERCA DE LOS LÍMITES QUIERE DECIR QUE LA DESVIACIÓN ESTANDAR σ DEBE SER APROXIMADAMENTE 10 MINUTOS. SE MIDIERÓN EN REALIDAD 15 TIEMPOS DE ACARREO Y SE OBTUVO UN PROMEDIO DE 142 MINUTOS Y UNA DESVIACIÓN ESTANDAR DE 12 MINUTOS. ¿PODRÁ REFUTARSE LA AFIRMACIÓN DE $\sigma = 10$ EN EL NIVEL DE SIGNIFICANCIA DEL 5%?

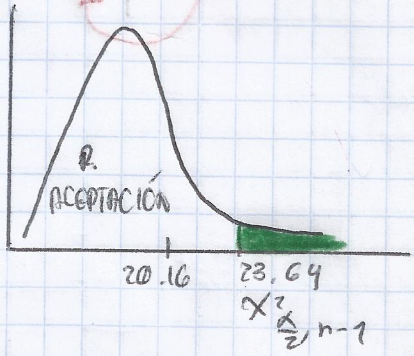
PRIMERO: PLANTEAR LA HIPÓTESIS.

$H_0: \sigma \leq 10$ $\quad$ $H_0: \sigma^2 \leq 100$

$H_1: \sigma > 10$ $\quad$ $H_1: \sigma^2 > 100$

SEGUNDO ¿CUÁL ES EL ESTADÍSTICO QUE DEBEMOS USAR?

$$\chi_0^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{(14)(12)^2}{100} = 20.16$$



LA REGIÓN DE RECHAZO ES $\chi^2 > \chi^2_{0.05,14}$

$\chi^2_{0.05,14} = 23.6848$

PUESTO QUE

$\chi_0^2 < \chi^2_{0.05}$ $\quad$ $20.16 < 23.69$

NO SE RECHAZA. CON BASE EN LA INFORMACIÓN DE LA MUESTRA NO HAY SUFICIENTE EVIDENCIA PARA CONCLUIR QUE LA DESVIACIÓN ESTANDAR ES SUPERIOR A 10 MINUTOS, CON $\alpha = 0.05$

PRUEBAS DE HIPÓTESIS PARA LA IGUALDAD DE VARIANZAS.

PARA PROBAR LA IGUALDAD DE DOS VARIANZAS DE POBLACIONES NORMALES CON PARAMETROS μ_1, σ_1^2, μ_2 Y σ_2^2 Y SE UTILIZA EL ESTADÍSTICO.

$$F_0 = \frac{S_1^2}{S_2^2}$$

DONDE $F \sim F_{n_1-1, n_2-1}$

Y LA PRUEBA DE DOS LADOS QUEDARÍA COMO:

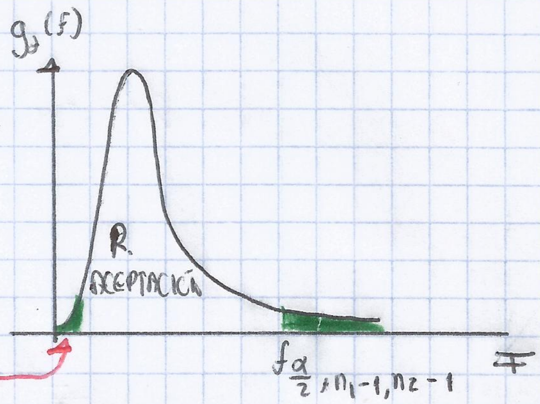
$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$

LA HIPÓTESIS SERÍA RECHAZADA SI

$$f_0 > f_{\frac{\alpha}{2}, n_1-1, n_2-1}$$

O BIEN

$$f_0 < f_{1-\frac{\alpha}{2}, n_1-1, n_2-1}$$



PRUEBAS DE HIPÓTESIS SOBRE UNA PROPOSICIÓN

ES UN CASO PARTICULAR DE LA MEDIA POR LO QUE NO DEBE SORPRENDER QUE EL ESTADÍSTICO DE PRUEBA SEA MUY SIMILAR

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Y LA HIPÓTESIS PUEDA PLANTEAR COMO:

$$H_0: p = p_0$$

$$H_1: p \neq p_0$$

PARA UNA PRUEBA DE DOS LADOS