



# Estadístico

Oleth Antonio Ardila Jaime

CAP 1. Muestreo y métodos de muestreo

CAP 2. Distribuciones muestrales y teoría del límite central

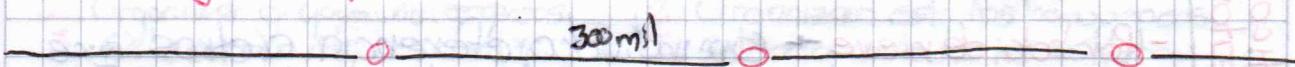
CAP 3. Pruebas de hipótesis con una sola muestra

CAP 4. Cálculo de tamaños demográficos

CAP 5. Pruebas de hipótesis para dos muestras

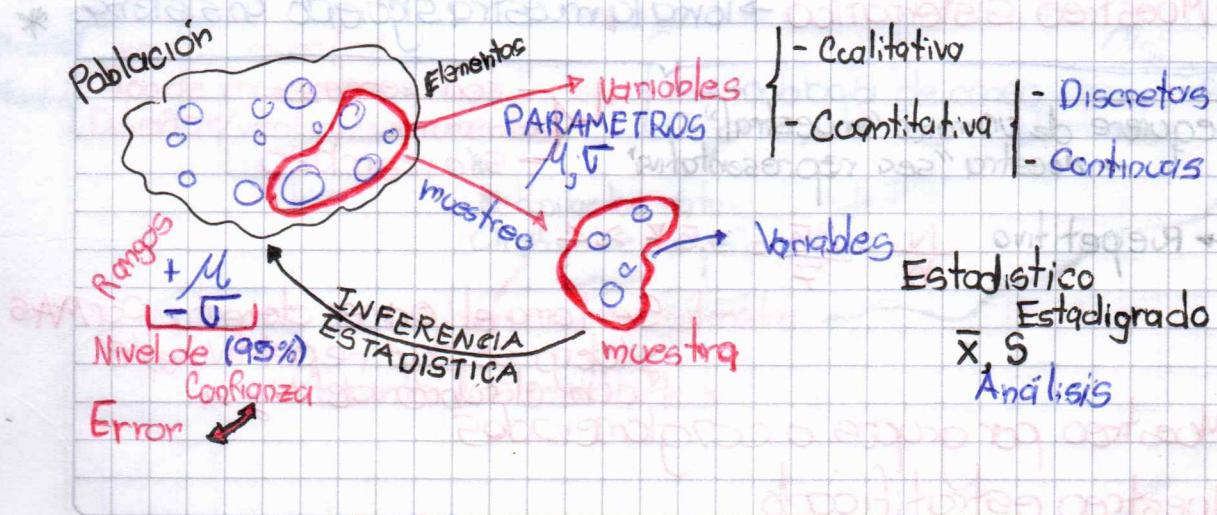
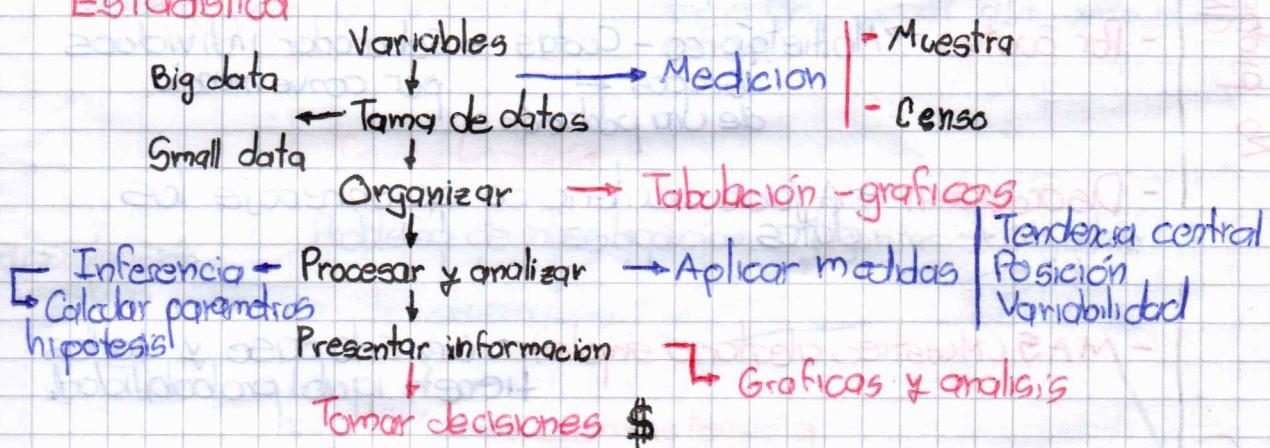
CAP 6. Análisis de varianza (ANOVA)

CAP 7. Regresión y correlación.



CAP 1: Muestreo y métodos de muestreo

Estadística



### No probabilístico (No aleatorio)

NO se puede calcular la probabilidad de que un elemento salga en la muestra.

#### Muestro

### Probabilístico (Aleatorio)

Calcular la probabilidad de que un elemento salga en la muestra.

#### No probabilístico (No aleatorio)

- Por conveniencia → Requiere un criterio/variable de conveniencia (Del investigador)
- Por bola de nieve → Un individuo referencia a otros que tengan una característica y se va multiplicando.
- Por cuotas → Multietápico - Cuotas → Seleccionar individuos grupos → por conveniencia de una población
- Discursivo → A juicio o criterio de quien recoje los datos.
- MAS (Muestreo aleatorio simple) → Un solo paso y todos tienen igual probabilidad.
- Muestreo sistemático → Toma la muestra siguiendo un sistema.

#### Probabilístico (Aleatorio)

Requiere de un marco "muestral"  
→ Muestra sea representativa

con reemplazo

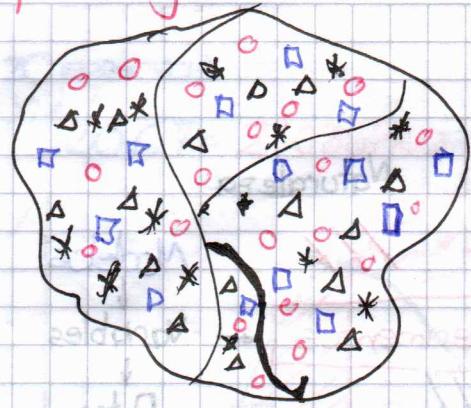
sin reemplazo

\* → Repetitivo  $\frac{N}{n} = \frac{25}{7} = 3.5 \neq 4$

Se toma el primer elemento por MAS  
y luego se sigue repetitivamente  
con el resto de  $N$  y  $n$ .

- Muestreo por grupos o conglomerados
- Muestreo estratificado

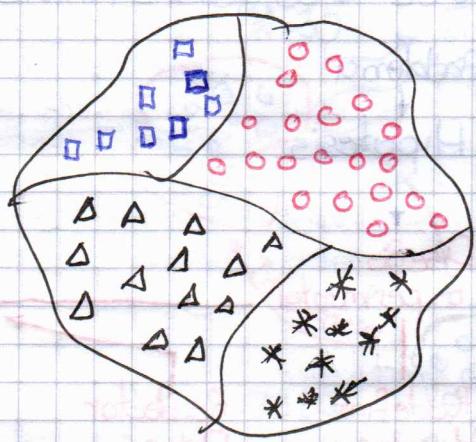
## Grupos/conglomerados



"Internamente heterogéneos"  
Revueltos

1. Marco muestral
2. Organizar grupos heterogéneos
3. Seleccionar algunos de los grupos para hacer muestreo
4. Seleccionar elementos dentro de los grupos por MAS o MS

## Estratos



"Internamente homogéneos"  
Clasificados

1. Marco muestral
2. Organizar estratos homogéneos
3. Seleccionar todos los estratos
4. Seleccionar elementos por MAS o MS
  - o MS pero el # de elementos debe ser proporcional al tamaño del estrato

## Investigación

Operacionalización  
de las variables



muestra/censo (Como se llegará a los datos)

Diseño  
Muestra

Diseño  
censo

marco muestral

marco censo

Tipo de muestreo - Cálculo tamaño

Metodología de censo

Diseñar/verificar instrumentos

Diseñar/censo  
Diseñar/verificar instrumentos

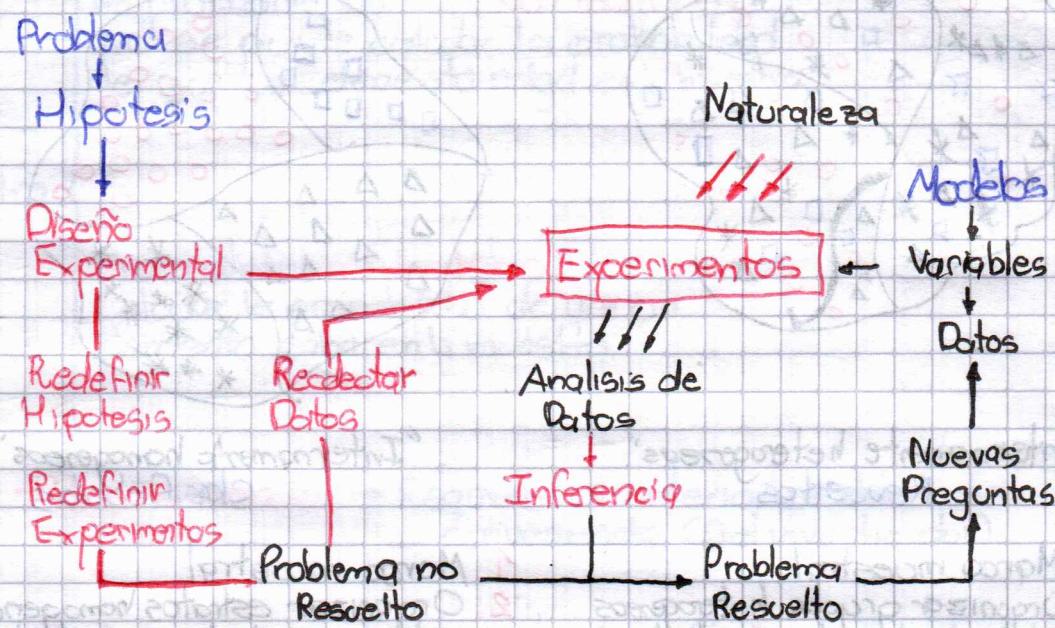
Recopilar los datos  
(Trabajo de campo)

Tabular, analizar, contrastar.

Presentar los resultados

## CAP 2: Distribuciones muestrales, teoría del límite central, ~~Normal~~

Ambito de la estadística inferencial



Estadística → Ing. Ambiental

Ambito de actuación

- Muestreo
- Estimación: Conocer un parámetro/valor
  - ↳ Intervalos
- Pruebas/test / contrastación de hipótesis
- Diseño experimental
- Inferencia Bayesiana

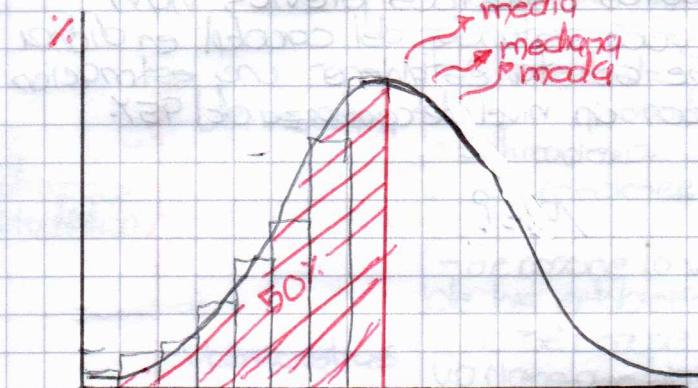
Parámetricos  
No paramétricos

# Distribuciones muestrales y el teorema del límite central

$N = 30$  personas

$n = 7$

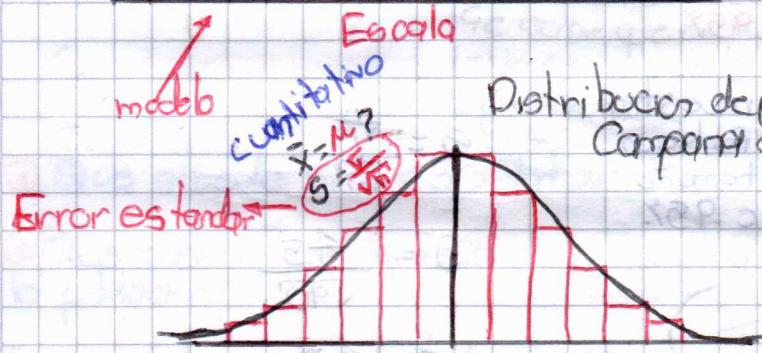
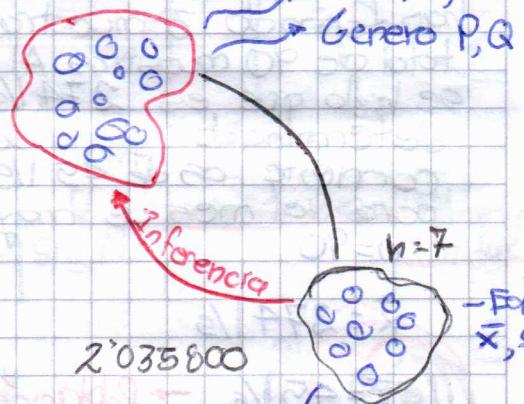
$$30 \times 7 = 2'035800 \text{ muestras}$$



$N = 30$

Edad/M, V

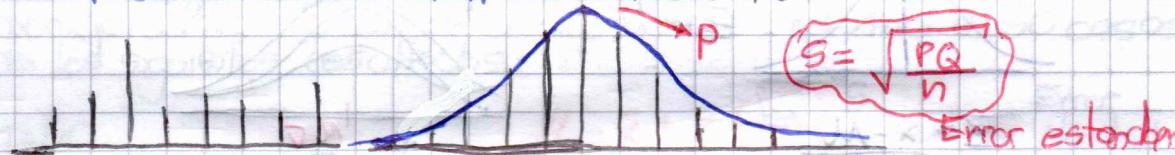
Genero P, Q



Distribución de probabilidad normal  
Campaña de Gauss

Distribución muestral

Teatro del límite central



If  $n \geq 30$  so distribution  $\approx$

Estimación:  $\rightarrow$  Variables cuantitativas

$\rightarrow$  Variables cualitativas

Parímetro = Edad de personas =  $\frac{\bar{x}}{V} \quad \left[ \begin{array}{l} \text{Puntual} \\ \text{de intervalo} \end{array} \right]$

= Proporción de personas  $> 20 = \frac{P}{Q}$

## Ejemplo:

Se desea estimar el caudal promedio mensual. Para ello se han tomado 3 muestras diarias durante el mes para un total de 90 muestras. Al calcular el caudal medio resultante se halló que es  $274 \text{ l/s}$ . Datos históricos previos han determinado que la desviación estandar del caudal en dicha corriente es de  $45 \text{ l/s}$ . Se solicita establecer una estimación para el mes de mayo con un nivel de confianza del 95%.

$$n=90$$

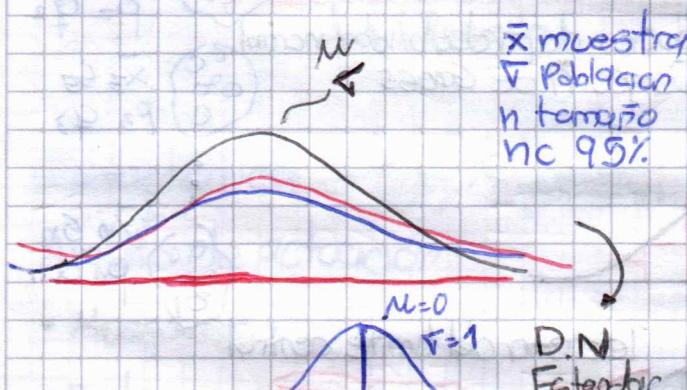
$$\bar{x} = 274 \text{ l/s}$$

$$M = ?$$

$$\sigma = 45 \text{ l/s} \rightarrow \text{Población}$$

Estimación

Nivel de confianza del 95%.

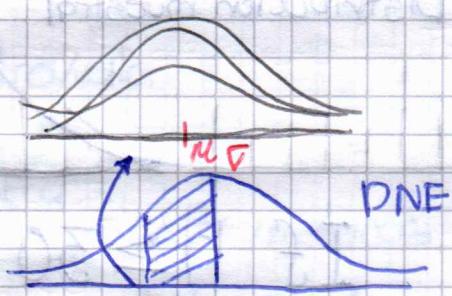


$$Z = \frac{\bar{x} - M}{\sigma}$$

$$S = \frac{\sigma}{\sqrt{n}}$$

$$S = \frac{45}{\sqrt{90}}$$

$$S = 4.74$$

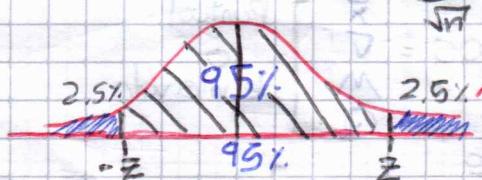


$$\text{nivel de confianza} \leftarrow Z = \frac{\bar{x} - M}{\sigma}$$

muestral

$$\frac{V}{\sqrt{n}}$$

$$M = \bar{x} \pm Z \left( \frac{V}{\sqrt{n}} \right)$$



$$Z = 1.9599$$

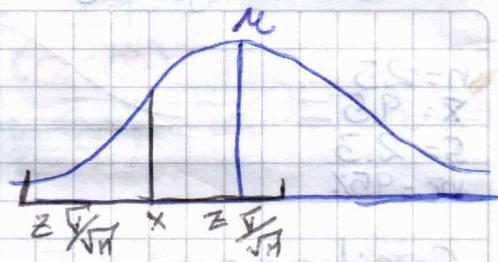
$$-1.9599 = -Z$$

$$\bar{M} = 274 + 1.96 \cdot 4.74 =$$

$$M = 274 - 1.96 \cdot 4.74 =$$

$$L_s = 283.29 \text{ l/s}$$

$$L_i = 264.7 \text{ l/s}$$



Estimación estadística

- Una muestra

- Estimación de promedios
- Estimación de proporciones

Si No

$n \geq 30$   $n < 30$  Población finita?

- Dos muestras

- Se conoce la varianza de las poblaciones

- No se conoce la varianza poblacional, pero se sabe que es igual entre las mismas

- No se conoce la varianza poblacional



Muestras pareadas

① Que sucede si  $n < 30$  datos?

② Que sucede si el tamaño de la muestra es grande respecto a la población?

La secretaría de salud de ecuador está interesada en conocer el consumo promedio de agua en  $\text{m}^3/\text{mes}$  de cierto barrio del municipio. Para ello realizó un muestreo sistemático en 50 casos con los siguientes resultados:

$$n = 50$$

$$\bar{x} = 9.5 \text{ m}^3/\text{mes}$$

$$S = 2.3 \text{ m}^3/\text{mes}$$

$$nc = 95\%$$

$$\bar{M} = \bar{x} \pm Z \left( \frac{S}{\sqrt{n}} \right)$$

$$\bar{M} = 9.5 \pm Z \left( \frac{2.3}{\sqrt{50}} \right)$$

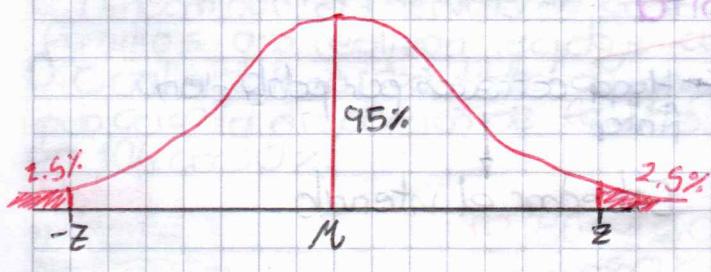
$$\begin{aligned} Z &= 1.95 \\ Z &= 1.96 \end{aligned}$$

Distribución normal  $95\% = nc$

$$\begin{aligned} \bar{M} &= 9.5 + 1.95 \left( \frac{2.3}{\sqrt{50}} \right) = 10.13 \\ \bar{M} &= 9.5 - 1.95 \left( \frac{2.3}{\sqrt{50}} \right) = 8.86 \end{aligned}$$

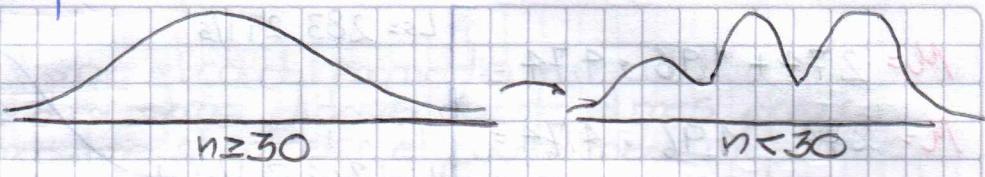
$$L_{\text{inf}} = 8.86$$

$$L_{\text{sup}} = 10.13$$



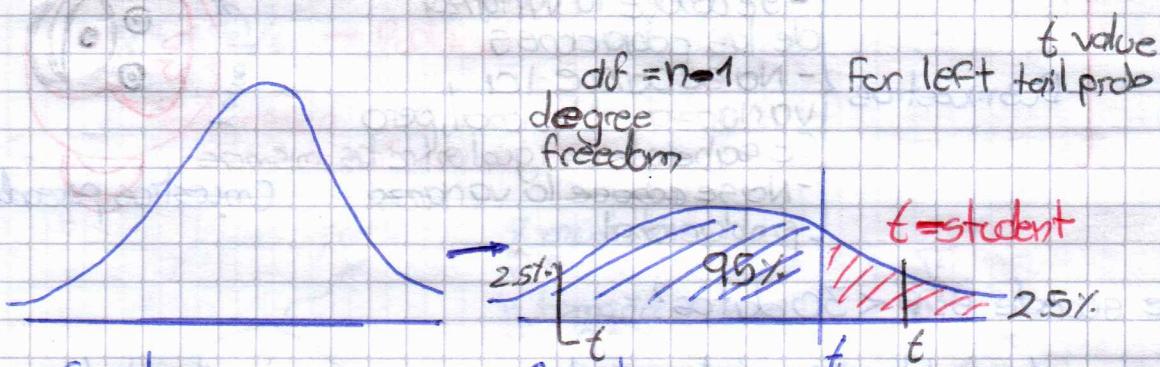
## Variación de ejemplo

$$\begin{aligned} n &= 25 \\ \bar{x} &= 9.5 \\ S &= 2.3 \\ NC &= 95\% \end{aligned}$$



### Consideraciones

- ① La población debe tener comportamiento normal
- ② Nueva distribución (modelo)



- Simétrica
- Asintótica eje x
- Mesocártica y/o leptocártica

- Simétrica
- Asintótica al eje x
- Platicártica
- Grados de libertad

$$M = \bar{x} \pm t \left( \frac{S}{\sqrt{n}} \right) \quad [\text{Error estándar de la muestra}] \sqrt{S}$$

$$\pm t = 2.06$$

$$M = 9.5 \pm t \left( \frac{2.3}{\sqrt{25}} \right)$$

$$L_{\text{sup}}/M = 9.5 + 2.06 \left( \frac{2.3}{\sqrt{25}} \right) = 10.44$$

$$L_{\text{inf}}/M = 9.5 - 2.06 \left( \frac{2.3}{\sqrt{25}} \right) = 8.55$$

### Corrección por población finita

$$n \geq 0.1N$$

(10%)

→ Hacer corrección por población finita

$$M = \bar{x} \pm \left[ \frac{E}{Z} \right] * \left( \frac{S}{\sqrt{n}} \right) * \left( \frac{N-n}{N-1} \right)$$

↓ Estrechar el intervalo

## Caso hipotético de $N=200$

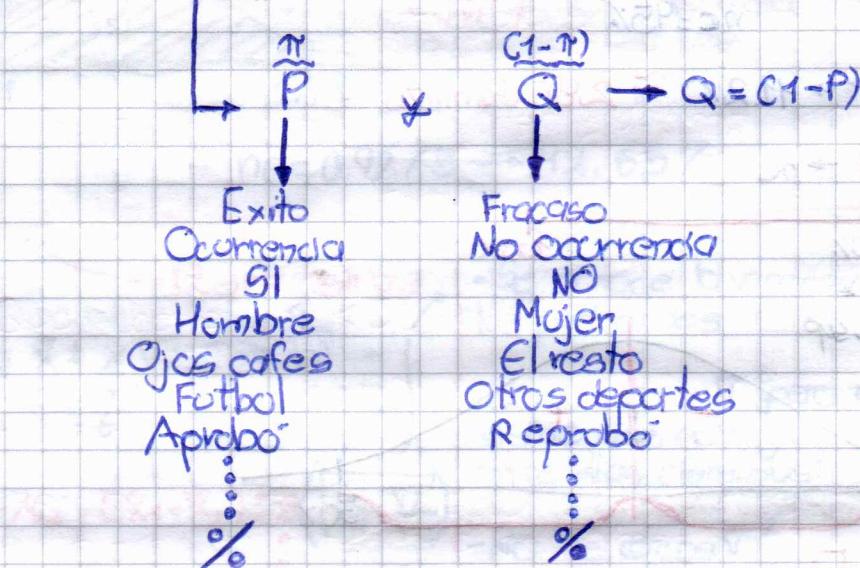
Donde  $n=25$

$$2.5 \geq 20(0.1 * N)$$

$$M = 9.5 \pm 2.063 \cdot \left( \frac{2.3}{\sqrt{25}} \right) \cdot \left( \frac{\sqrt{200-25}}{\sqrt{200-1}} \right)$$

$$\begin{aligned} L_{\text{sup}} &= 10.39 \\ L_{\text{inf}} &= 8.61 \end{aligned}$$

Estimación de proporciones  $\rightarrow$  Estimación de promedios  
 variables cuantitativas      variables cualitativas



Un equipo de ingenieros está determinando la preferencia o no de los habitantes de un municipio hacia los hábitos de reciclaje y separación en la fuente de residuos sólidos. Se ha formado una muestra de 80 familias y se encontró que el 20% de ellas manifestaron realizar reciclaje y separación en la fuente. A partir de esto informarán:

- A. La cantidad de familias que realizan reciclaje en la muestra.
- B. Si la cantidad de familias de todo el municipio es 2550, cuantas familias se esperaría que realicen reciclaje en dicha población.
- C. Determinar el intervalo de confianza para la población de familias que realizan reciclaje con un nivel de confianza del 99%.
- D. Cuál es la probabilidad de que en una muestra de dicha población la proporción de familias que realizan reciclaje esté en un 10% y un 30%.

### Datos

$$n = 80 \text{ Fam}$$

$$N = 2550 \text{ Fam}$$

$$p = 20\% \text{ (Realizan reciclaje)}$$

$$q = 80\% \text{ (No realizan reciclaje)}$$

$$\text{Distribución normal} \rightarrow n \geq 30$$

$$\{ a. 80 \sim 20\% = 16$$

$$\{ b. 2550 \sim 20\% = 510 \text{ (se espera)}$$

Estimación puntual

M=4 ab anterior 2020

C.  $P = p \pm Z \left( \sqrt{\frac{pq}{n}} \right)$   
 ↓  
 95% Error estándar de la muestra

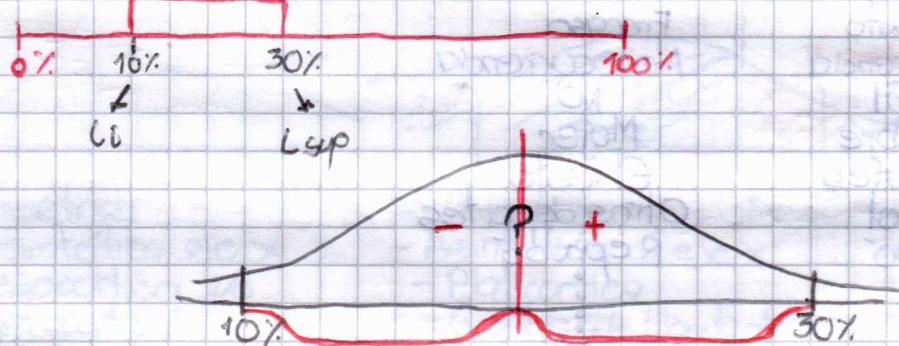
$$P = 0.2 \pm 1.96 \sqrt{\frac{0.2 \cdot 0.8}{80}}$$

$$L_s = 0.287 \rightsquigarrow 28.7\% \approx 732 \text{ familias}$$

$$P \quad n = 95\%$$

$$L_i = 0.112 \rightsquigarrow 11.2\% \approx 285 \text{ familias}$$

d. Probabilidad  $\rightarrow n_c$



$$Z = \frac{p - p_0}{\sqrt{\frac{pq}{n}}}$$

$$Z = \frac{0.3 - 0.2}{\sqrt{\frac{0.2 \cdot 0.8}{80}}} = 2.23$$

Al aplicar q la tabla o app el z en rango de -2.23 a 2.23 resulta,

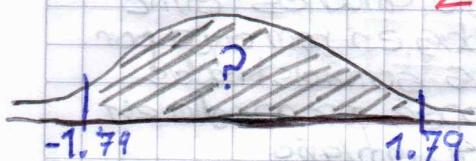
97.42%

## Ejemplo con promedios

Caudal en l/sec

Estimación  $\begin{cases} L_S = 277.5 \\ L_i = 260.5 \end{cases}$   $R = \bar{x} \pm Z \left( \frac{\sigma}{\sqrt{n}} \right)$

$n = 90$   
 $\sigma = 45 \text{ l/s}$



$Z = \frac{M - \bar{x}}{\left( \frac{\sigma}{\sqrt{n}} \right)} = \frac{277.5 - 269}{\left( \frac{45}{\sqrt{90}} \right)} = 1.79$

$NC = 0.9265 \rightarrow 92.65\%$

→ Dos muestras  $M = \bar{x} \pm Z \frac{\sigma}{\sqrt{n}}$

$M_1 - M_2 = (\bar{x}_1 - \bar{x}_2) \pm Z \left( \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$

- Se conocen varianzas de las poblaciones
- No se conoce  $\sigma^2$  pero se sabe que son iguales entre ellas (Pareadas)
- NO se conoce  $\sigma^2$

Diferencia entre los dos  $M_1 - M_2$

 $\sigma^2$  conocido $\sigma^2$  desconocido pero igual $\sigma^2$  desconocido

Dispersión  $\begin{cases} \text{Desviación } \sigma_1, \sigma_2, \text{ etc.} \\ \text{Varianza } \sigma_1^2, \sigma_2^2, \text{ etc.} \end{cases}$

$\checkmark$   $\checkmark$   $\checkmark$

No conocidos por población finita

$n \geq 30$        $n < 30$

Estimación de la diferencia de medias de dos poblaciones con varianza desconocida.

Una población

$M = \bar{x} \pm Z \left( \frac{\sigma}{\sqrt{n}} \right)$

Dos poblaciones

$M_1 - M_2 = (\bar{x}_1 - \bar{x}_2) \left( \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$

CARACTERISTICAS 1er caso  $\sigma^2$  conocida

1. No se estima con valor si no una diferencia.
2. Las varianzas deben conocerse
3.  $n_1 \geq 30$   $n_2 \geq 30$

Un grupo de ingenieros ambientales desea conocer las diferencias en las precipitaciones promedio anuales entre Colombia y otros países de América medidas en mm agua por año. Para ello se acordó, a partir de datos suministrados por el banco mundial que las precipitaciones promedio en Colombia en (1970 y 2015) fueron de 3240 mm/año

Colombia (1970 - 2015)

$$\bar{x} = 3240 \text{ mm/año}$$

$$\sigma^2 = 560 \text{ mm}^2/\text{año}$$

País	$\bar{x}$	$\sigma^2$
Costa Rica (1980 - 2015)	2926 mm	350
Brasil (1965 - 2015)	1761 mm	280
Ecuador (1980 - 2014)	2274 mm	347

A. Estimar la diferencia media entre las precipitaciones de Brasil y Colombia con un nivel de confianza de 97%.

B. Estimar la diferencia media entre las precipitaciones de Ecuador y Colombia con un nivel de confianza de 97%.

Brasil vs Colombia

$$\bar{x} = 1761 \text{ mm/año} \quad 3240 \text{ mm/año}$$

$$\sigma^2 = 280 \text{ mm}^2/\text{año} \quad 560 \text{ mm}^2/\text{año}$$

$$\sigma^2 = 78400 \text{ mm}^2/\text{año}^2 \quad 313600 \text{ mm}^2/\text{año}^2$$

Caso →

a. Dos poblaciones → Diferencia

b. Se conoce  $\sigma^2$

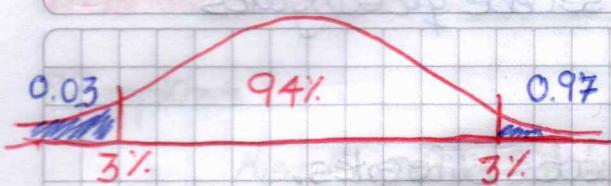
c.  $n_1$  y  $n_2 \geq 30$

$$n = 49$$

$$45$$

$$M_C - M_B = \bar{x}_C - \bar{x}_B \pm Z \left( \sqrt{\frac{\sigma_C^2}{n_C}} + \sqrt{\frac{\sigma_B^2}{n_B}} \right)$$

$n_C = 94\%$  ↗



$$+Z = 1.88 \\ -Z = -1.88$$

$$M_C - M_B = (3240 - 1761) \pm 1.88 \left( \sqrt{\frac{313600}{45} + \frac{78400}{49}} \right)$$

$$1304.97 \leq M_C - M_B \leq 1653.03$$

$$n_C = 94\%$$

$$\begin{cases} L_S = 1653.03 \\ n_C = 94\% \\ L_U = 1304.97 \end{cases}$$

### Ecuador Vs Colombia

$$\bar{x} = 2274 \text{ mm/año}$$

$$3240 \text{ mm/año}$$

$$\bar{V} = 377 \text{ mm/año}$$

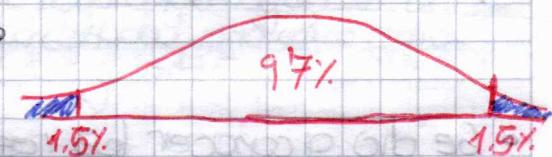
$$560 \text{ mm/año}$$

$$\bar{V}^2 = 142129 \text{ mm/año}$$

$$313600 \text{ mm/año}$$

$$n = 34 \text{ años}$$

$$45 \text{ años}$$



$$Z = 2.17$$

$$M_C - M_E = (3240 - 2274) \pm 2.17 \left( \sqrt{\frac{313600}{45} + \frac{142129}{34}} \right)$$

$$936.87$$

$$1195.12$$

La diferencia entre las precipitaciones entre Costa Rica y Colombia oscila entre 936.87 mm/año y 1195.12 mm/año con un nivel de confianza de 92%.

$$936.87 \leq M_C - M_E \leq 1195.12$$



OK  
3/3

Se desconoce  $\sigma_1$  y  $\sigma_2$  ( $\sigma_1^2$  y  $\sigma_2^2$ ) pero se debe que son iguales

## Muestras pareadas

- Misma población medida en dos períodos diferentes.
- Tener una población y dividirla en dos o más grupos

Estimación de la diferencia de medias con varianzas desconocidas pero iguales

$$M_1 - M_2 = (\bar{x}_1 - \bar{x}_2) \pm t_{(Z)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

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

Si se usan  
t-student  
 $g(v) = n_1 + n_2 - 2$

- Se dio a conocer los resultados de un análisis del peso de Calcio en cemento estandar y cemento contaminado con plomo. Los niveles bajos de Calcio indican que el mecanismo de hidratación del cemento queda bloqueado haciendo que el agua determine su estructura. Al tomar 10 muestras de cemento estandar se encontró que el peso promedio de calcio es de 90 g con una desviación de 5 g. Tomando 15 muestras de cemento contaminado con plomo se obtuvo un promedio de 87 g con una desviación de 4 g. Supongamos que el porcentaje de peso de calcio está distribuido de manera normal. Si las dos poblaciones tienen la misma desviación determinar el intervalo de confianza para la diferencia entre las medias, con un nivel de confianza de 95%.

<b>Estandar</b>	<b>Contaminado</b>
$n_1 = 10$ muestras	$n_2 = 15$ muestras
$\bar{x}_1 = 90$ g	$\bar{x}_2 = 87$ g
$S_1 = 5$ g	$S_2 = 4$ g

Distribución normal

$$\sigma_1 = \sigma_2$$

Intervalo de confianza?  $\alpha = 95\%$

## Identificar el caso

Reemplazo

$$Ms - Mc = (90 - 87) \pm t Sp \sqrt{\frac{1}{10} + \frac{1}{15}}$$

$$Sp^2 = \frac{(60-1) \cdot 5^2 + ((15-1) \cdot 4^2)}{10+15-2} = 19.52 = \sqrt{19.52} = 4.419$$

$$nc = 95\% = 2.068 = glcdf = 2.3$$

$$\begin{aligned} (-t)Ms - Mc &= -0.73 \\ (t)Ms - Mc &= 6.73 \end{aligned} \quad \left. \right\} nc = 95\%$$

- Un ingeniero ambiental sospecha que existe contenido de cobre en el agua de un acueducto cuando éste llega a los hogares, él quiere determinar siadicionalmente si dicha diferencia entre día y la noche. Los datos tomados son los siguientes.

	Día	Noche
n <sub>d</sub>	16	25
$\bar{x}$	50 ppm	14 ppm
s <sub>d</sub>	5 ppm	12 ppm

Estimar la diferencia de medias para el contenido de cobre con nc = 95%.

$$Sp = \sqrt{\frac{(24 \cdot 14) + (15 \cdot 25)}{39}} = \sqrt{98.23} = 9.91$$

$$\begin{aligned} (-t)M_N - M_d &= 96.96 \\ (t)M_N - M_d &= 83.43 \end{aligned} \quad \left. \right\} nc = 95\%$$

# Estimación de diferencias de medias $\bar{X}_1^2$ y $\bar{X}_2^2$ descorreladas

→ Siempre con distribución t

→ Supone comportamiento normal de las poblaciones

$$(M_1 - M_2) = (\bar{X}_1 - \bar{X}_2) \pm t * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Grados de libertad (d.f.) →  $g.l. = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left[ \frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1} \right]}$

Se aproxima al entero más cercano.

- Un grupo de ingenieros desarrolló para extraer cierto metal. Este proceso es considerablemente más lento y posiblemente logre menor resistencia del metal, por tanto se hace necesario estimar la diferencia entre el metal antiguo y el nuevo. Para ello se han seleccionado 2 muestras de 12 probetas de cada proceso y se someterán a pruebas así:

(1) Antiguo	446	401	476	421	459	438	481
	411	456	429	459	445		
(2) Nuevo	462	448	435	465	429	472	453
	459	427	468	452	447		
Resistencia $\text{kg/cm}^2$							

$$\bar{X}_1 = 443.33 \quad s_1 = 24.824$$

$$M_2 - M_1 = ?$$

$$\bar{X}_2 = 451.41 \quad s_2 = 14.939$$

$$g.l. = \frac{\left( \frac{24.824^2}{12} + \frac{14.939^2}{12} \right)}{\left[ \frac{\left( \frac{24.824^2}{12} \right)^2}{12-1} + \frac{\left( \frac{14.939^2}{12} \right)^2}{12-1} \right]} =$$

## Capítulo 3. Pruebas o test de hipótesis

### Hipótesis

Supuesto que responde a la pregunta o intérprete a investigar.

Muestreo  
Requiere prueba (contraste) de hipótesis. (Hay incertidumbre)

Aceptar      Rechazar

Censo  
No requiere test de hipótesis solo "comparar" datos.

### Clasificación

#### Pruebas de hipótesis

##### Paramétricas

Utilizan medidas como:

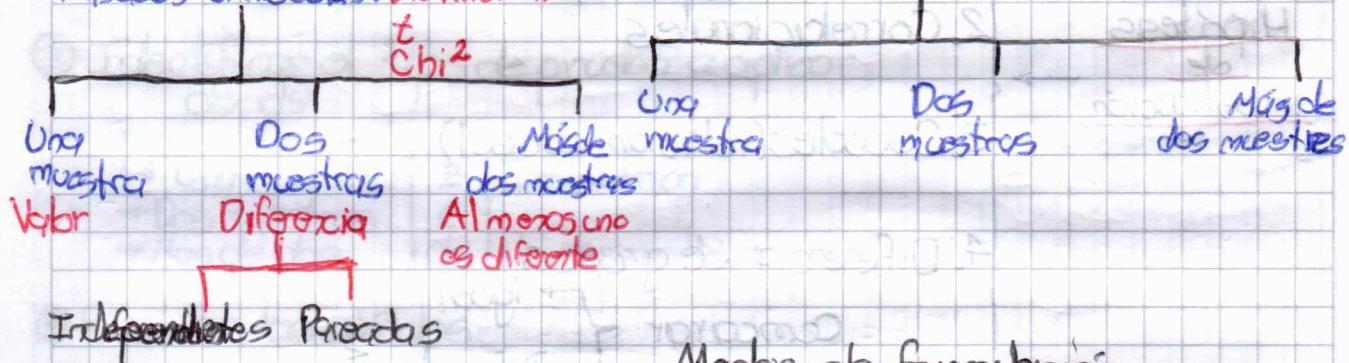
Promedio, desvío, varianza, mediana.

##### Numericas

Basadas en modelos: Dist. normal

##### No paramétricas

Utiliza medidas y modelos NO convencionales.



### Tipos de hipótesis

#### Hipótesis de investigación

Vinculadas a un proceso investigativo  
(Coherentes con los objetivos y las variables)

#### Modos de formulación

Afirmativo: Se expresa como si estuviera sucediendo

VARIABLES → Expreso →  
Futuro simple

Condicional:

~~Por los aerosoles~~

El uso de aerosoles dañina la capa de ozono.

El uso de aerosoles crea un déficit en la capa de ozono.

Si se encuentran restos de aerosoles en el agua este, es una contaminación.

### Características

- Refiere a situación real medible, posible.
  - Debe incluir variables de la investigación.
  - Clara y creíble
  - Revisiones técnicas para medirlas y probarlas.
- variable
- Dependientes
  - Independientes
  - Intervenientes

Hipótesis de investigación

1. Descriptivas del valor de una variable

2. Correlacionales

3. Causales (relación causal)

4. Diferencia de grupos

Comparar      → igual

Diferente  $\neq$

Hipótesis estadísticas (Números y simbolos)

No iguales  $> < \leftrightarrow$  Hipótesis Alterna  
(Simbolo)

Igual  $\leq \geq \rightarrow$  Hipótesis Nula  
(Simbolo)

Esta que se "ataca" la que se demuestra

Acepto Rechazo

## Pruebas/test de hipótesis

### Procedimiento de prueba de hipótesis

Hipótesis → Investigación → Hipótesis estadísticas

$H_0$	Acepta
$H_a$ o $H_1$	Rechaza

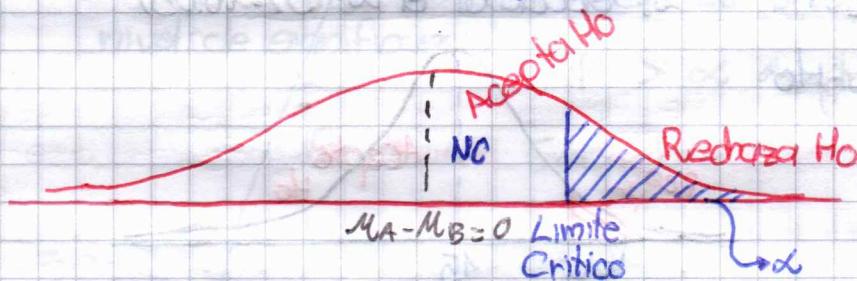
### ① Formular las hipótesis estadísticas

→ Hipótesis nula  
Hipótesis alterna

$$H_A: \mu_A - \mu_B > 0 \rightarrow \mu_A > \mu_B$$

$$H_0: \mu_A \leq \mu_B \rightarrow \mu_A - \mu_B \leq 0$$

### ② Seleccionar $\alpha$ (nivel de significancia) ( $1 - \alpha = N. \text{confianza}$ )



### ③ Identificar el tipo de prueba a aplicar

→ una muestra	$Z$	Paramétrica
→ Dos muestras	$t$	
→ Muestras	chi	No paramétrica

### ④ Realizar los cálculos

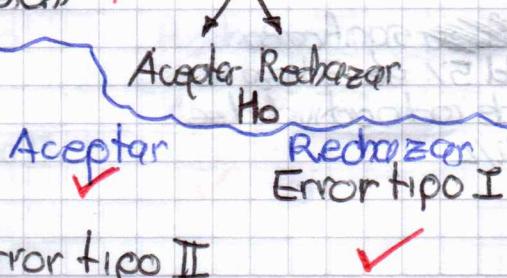
- 1. Valor crítico
- Tres caminos
  - 2.  $Z(t)$  crítico (valor normalizado)
  - 3. C (Computador) P probabilidad Pvalue

### ⑤ Verificar o comprobar la hipótesis (Tomar una decisión)

#### ERRORES POSIBLES

$H_0$  Verdadera

$H_0$  Falsa



De acuerdo a la normatividad vigente el nivel de contaminantes radioactivos en agua potable debe ser inferior a 15 PCi/L. La evidencia histórica sugiere que el abastecimiento de agua en una ciudad es por lo tanto estable. Para probar esto se realizó un muestreo midiendo los niveles de radioactividad.

a) Identifique la hipótesis nula y la alterna para el valor promedio de la población.

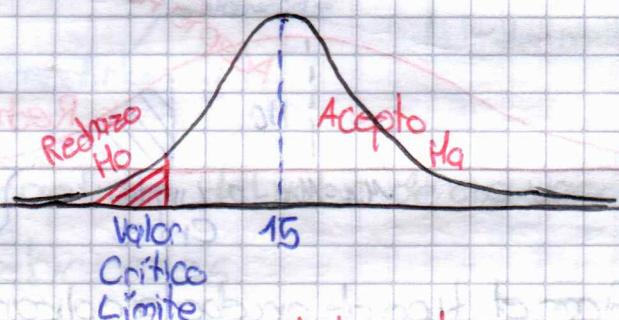
b) Suponga que se tomaron 44 muestras en diferentes días y se encontró una desviación estándar de 5.8 PCi/L con un comportamiento normal con un valor promedio de 14.7 PCi/L. Pruebe la hipótesis formulada.

$$1. \bar{M} < 15 \text{ PCi/L} : H_0$$

2. Seleccionar el  $\alpha = 5\% (0.05)$

$$\bar{M} \geq 15 \text{ PCi/L} : H_a$$

Formular hipótesis



3. → Paramétrica  
→ Dist. Normal Z  
→ Una muestra

4. Calcular

- Valor crítico 1
- Z crítico 2
- Probabilidad (Computadora) 3

$$1. \bar{M} = \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

error  
nivel de confianza  $1 - \alpha$

$$\alpha = 0.05 \\ Z = 1.64 \text{ (seguir)}$$

$$\text{Valor Crítico} = \bar{M} + z \frac{\sigma}{\sqrt{n}}$$

$$V_{Cr} = 15 + (-1.64) \frac{5.8}{\sqrt{44}}$$

$$V_{Cr} = 14.8022$$

5. Tomar decisión

Con un nivel de significancia de 5% se rechaza  $H_0$  por tanto el nivel de radioactividad es inferior a 15 PCi/L



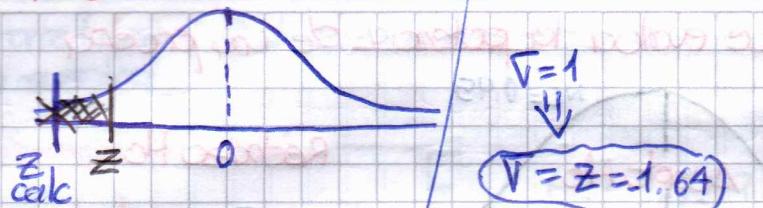
El resultado es correcto

Ho se rechaza

Norma

## 2. Z critico

Dist. Normal Estándar



hasta 1.64 desvest.

$$Z_{calulado} = \frac{\bar{x} - \mu}{\left(\frac{V}{\sqrt{n}}\right)} = \frac{14.7 - 15}{\left(\frac{0.8}{\sqrt{991}}\right)} = -2.48$$

Misma  
Cocobisico

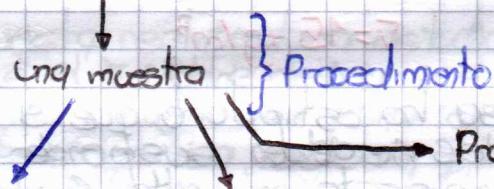
## 3. Ordenador

$\alpha$  Vs  $P$   
nivel de significancia

 $P < \alpha$  Rechazo $P > \alpha$  Acepta

Software: Geogebra

## TEST DE HIPOTESIS



Ayer paramétrico Utilizando la  
utilizando la distribución t  
normal Z

$n \geq 30$  datos  $n < 30$  datos

Requisito: Comportamiento  
normal de la  
población.

$n \geq 30$   
Comportamiento  
normal

Ho Verdadera

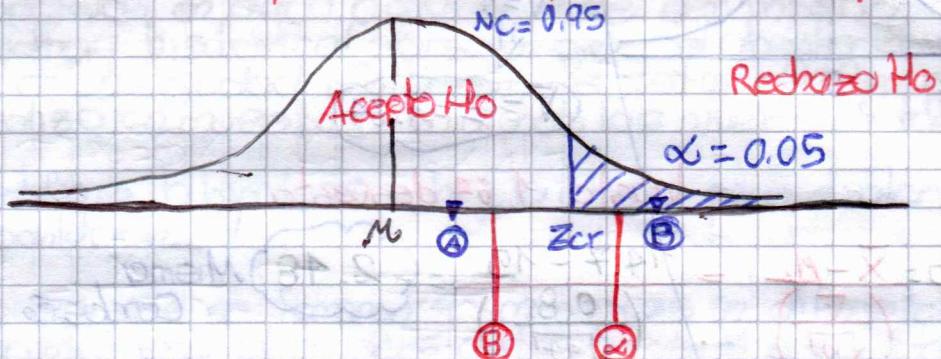
Acepto  
Ⓐ ✓Rechazo  
Error tipo I  
Alfa  $\alpha$ 

Ho Falsa

Error tipo II  
Beta  $\beta$ 

✓ Ⓑ

→ Existe un test que evalúa la potencia de una prueba



El estándar aceptado para el polvo de Cadmio en el lugar de trabajo debe ser inferior a  $200 \text{ mg/m}^3$ . Para supervisar los niveles de Cadmio en una empresa se toman muestras del polvo en el aire durante tres horas en intervalos de 10 min.

$$200 \text{ mg/m}^3$$

muestras  $\rightarrow 3h \rightarrow c/10 \text{ minutos}$   $n=18$  muestras

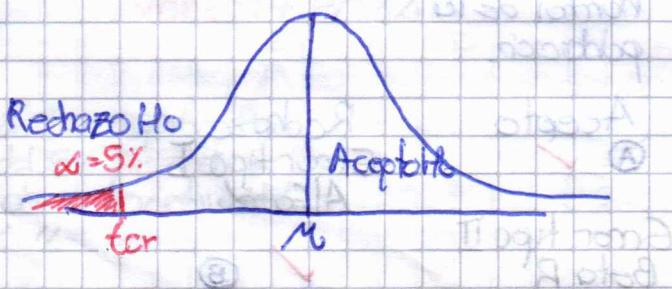
El cálculo del promedio el resultado obtenido fue de  $195 \text{ mg/m}^3$ . A partir de datos de otras empresas se sabe que el nivel del polvo de Cadmio en aire tiene distribución normalmente y se conoce de un estudio previo que la desviación de este elemento en el aire es de  $15 \text{ mg/m}^3$ .

$$\bar{x} = 195 \text{ mg/m}^3 \quad \sigma = 15 \text{ mg/m}^3$$

La gerencia de la empresa ha establecido que si los niveles de Cadmio son inaceptables se deben suspender actividades laborales. Esto conlleva un costo muy alto. ¿Cuál sería su recomendación?

$$\text{④ H0: } M_0 < 200 \text{ mg/m}^3 \quad M_0 \geq 200 \text{ mg/m}^3 : \text{H1}$$

$$\text{② } \alpha = 5\%$$



## Prueba de hipótesis con una muestra

③ Paramétrico Distribución t  
Una muestra

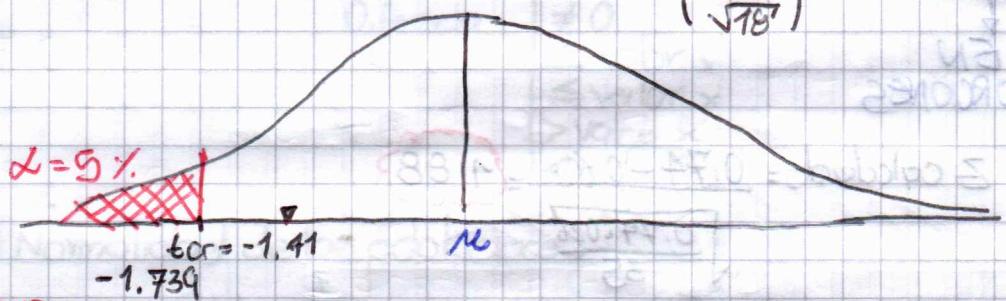
④ Crítico  $\rightarrow \alpha$

Ejemplo  $\rightarrow$  Fórmula

$$t = -1.739 \\ df = 17$$

$$t_{cr} = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

$$t_{calc} = \frac{195 - 200}{\left(\frac{15}{\sqrt{18}}\right)} = -1.41$$



⑤ Decisión

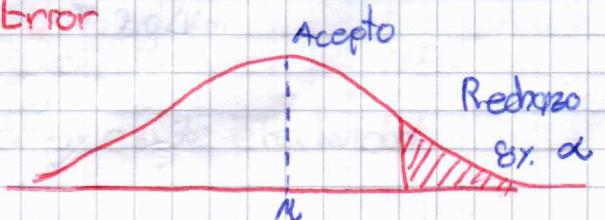
Rpta: Se acepta la  $H_0$  con un nivel de significancia del 5%, por tanto hay evidencia estadística para afirmar que el polvo de Cadmio supera  $200 \text{ mg/m}^3$ .

En una encuesta realizada a estudiantes de la U Libre (35) se encontró que el 74% debe mejorar sus hábitos de estudio. Si esto es cierto, la universidad está dispuesta a desarrollar un programa de apoyo. La rectora y los decanos quieren determinar si el porcentaje verdadero es superior al 60%. Antes de implementar el programa y hacer la inversión, realizar la prueba de hipótesis.

$n = 35$ .

74% mejorar hábitos de estudio      60%       $P > 60\%$   
cuantos son? **26 estudiantes**

- ①  $P > 60\%: H_0$     ②  $\alpha = 8\%: \text{Error}$   
 $P \leq 60\%: H_1$



③ Una muestra, proporciones  $n \geq 30$

④ Cálculo

$$Z_{\text{cr}} \rightarrow \alpha = 8\% \quad Z_{\text{critico}} = 1.405$$

$$\text{Zcalc} \Rightarrow \frac{\bar{X} - N}{\left(\frac{P}{\sqrt{n}}\right)} \sim -\frac{P - P}{\sqrt{\frac{P \cdot Q}{n}}}$$

↓  
No son  
PROPORCIONES

$$Z_{\text{calculado}} = \frac{0.74 - 0.60}{\sqrt{\frac{0.79 \cdot 0.21}{35}}} = 1.88$$

⑤ Decisión:

Rta: Con un  $\alpha$  de 8%, se concluye con la proporción de proporción de estudiantes que obtienen mejoras sus hábitos de estudio es superior a 60%.

## Pruebas de hipótesis para dos poblaciones

$n_1 \text{ y } n_2 \geq 30$   
Comportamiento normal

$n_1 \text{ y } n_2 < 30$   
comportamiento normal

Muestras pareadas  
comportamiento normal

$n_1 \text{ y } n_2 \geq 30$   
Proporciones  
comportamiento normal

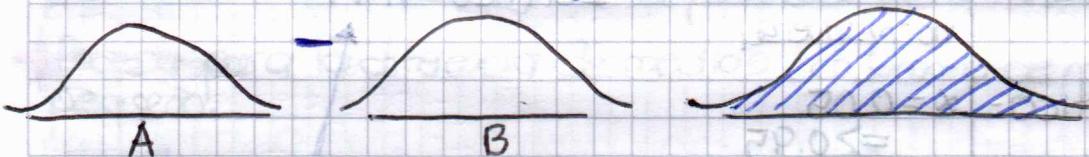
### CONSIDERACIONES

- Se va a comparar la diferencia no valores individuales.

$$\begin{array}{l|l} M_A - M_B & = 0 \\ \text{Diferencia} & \neq 0 \\ & \geq \text{valor } x \\ & \leq \text{valor } x \\ & > \text{valor } x \\ & < \text{valor } x \end{array}$$

- Normalidad de las poblaciones

Poblaciones de origen normal



- Prueba de hipótesis para la diferencia de medios con tamaños de muestras grandes  $n_1 \text{ y } n_2 \geq 30$ .

- Un equipo de ingeniería ha desarrollado un dispositivo para medir las emisiones de CO<sub>2</sub> en los automóviles. Para ello se tomó una muestra de 100 vehículos, un grupo A que incluye el dispositivo (60) y 40 que no, grupo B.

Grupo A :  $\bar{x}: 12.5 \text{ g/Km}$  (recorrido)  $s: 8.3 \text{ g/Km}$

Grupo B :  $\bar{x}: 13.2 \text{ g/Km}$  (recorrido)  $s: 7.2 \text{ g/Km}$

Se le ha pedido a Ud comprobar si existe efectividad.

seguimiento de los datos estadísticos

①

$$H_0: \mu_B - \mu_A > 0 \quad H_1: \mu_B - \mu_A \leq 0$$

sin con

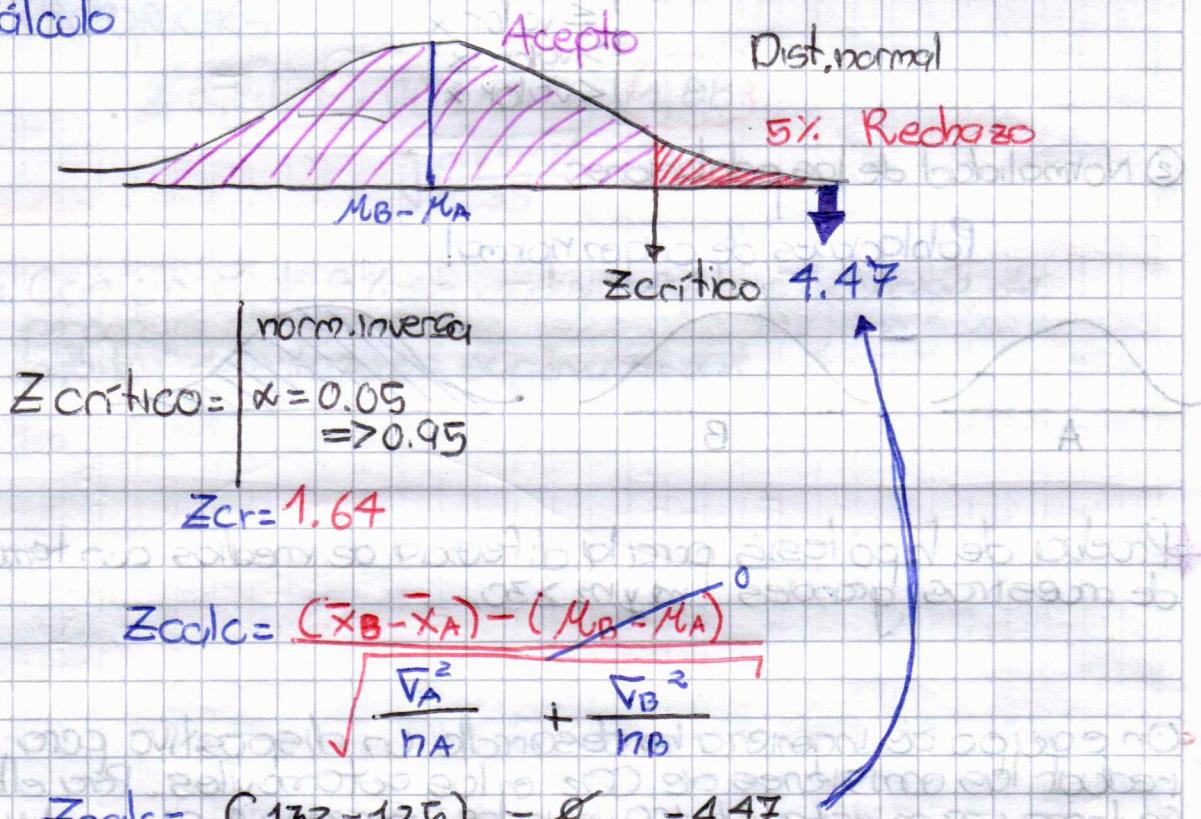
② Nivel de significancia: 5%  $\alpha = 5\%$

③  $n_1, n_2 > 30$

Comportamiento normal

Dos muestras  $\rightarrow$  Dospoblaciones

④ Cálculo



⑤ Análisis

Con un  $\alpha = 5\%$  se rechaza  $H_0$ , por tanto el nuevo dispositivo está reduciendo el promedio de emisión de los vehículos.

01/09/2013

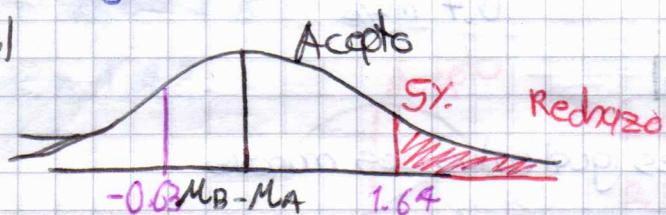
$$\textcircled{1} \quad H_0: M_B - M_A > 8$$

$$H_1: M_B - M_A \leq 8$$

\textcircled{2} Nivel de significancia  $\alpha = 5\%$

\textcircled{3} Igual

\textcircled{4}



$$Z_{\text{crítico}} = 1.64$$

$$Z = \text{calculado} = -0.63$$

\textcircled{5} Análisis

Con un  $\alpha = 5\%$ . se acepta la hipótesis  $H_0$ , por tanto el nuevo dispositivo no reduce en más de  $8\text{g/km}$  el promedio de emisión.

\* Prueba para la diferencia de medias con tamaño de muestra pequeño.

Consideraciones

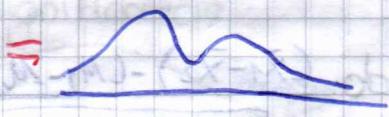
$$\textcircled{1} \quad n_1 \text{ y/o } n_2 < 30$$

\textcircled{2} Cambio en el cálculo del error estándar

↳ Se obtiene  $s_p = \sqrt{s_1^2 + s_2^2}$   
(Supone)

\textcircled{3} Comportamiento normal, necesario, de las dos poblaciones.

$$n_1 < 30 \quad n_2 < 30$$



yecaderna97@gmail.com.

carolina.

# EJEMPLO

	MARCA A	MARCA B
n <sub>i</sub>	10	8
X̄ <sub>i</sub>	3.1 mg/u	2.7 mg/u
S <sub>i</sub>	0.5 mg/u	0.7 mg/u

Comportamiento normal

Contenido de nicotina es igual en las dos marcas

① Hipótesis:

$$M_A = M_B \rightarrow M_A - M_B = 0 \quad H_0$$

$$M_A \neq M_B \rightarrow M_A - M_B \neq 0 \quad H_A$$

② Definir el nivel de significación  $\alpha = 5\%$

③ Tipo de prueba

$n_1, n_2 < 30 \rightarrow$  Distribución t-student

Comportamiento

④ Cálculo



t crítico |  $\alpha = 5\%$

$$df | n_1 + n_2 - 2 = 16$$

Si solo se conoce  $s$  (desv. estándar) siempre se trabajara la dis. t. QSI

$$n_1, n_2 \geq 30$$

Suponer  $\bar{v}_1 = \bar{v}_2$

$$n_1, n_2 < 30$$

t calculado: 
$$\frac{(\bar{x}_1 - \bar{x}_2) - (M_1 - M_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

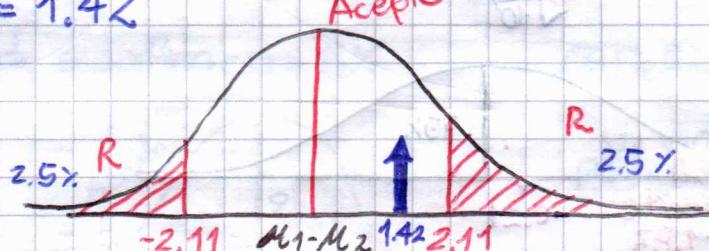
$$\Rightarrow s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$Sp^2 = \frac{9 \cdot 0.5^2 + 7 \cdot 0.7^2}{16} = 0.355 \rightarrow Sp = 0.60$$

$$t_{\text{calc}} = \frac{(3.1 - 2.7) - 0}{0.60 \sqrt{\frac{1}{8} + \frac{1}{10}}} = 1.42$$

$$t_{\text{calc}} = 1.42$$



### ⑤ Análisis

Se acepta  $H_0$  con un  $\alpha = 5\%$ . Luego no hay diferencia significativa en el contenido de nicotina de los dos muescos.

→ Muestras pareadas (misma individuo antes vs después)  
Pruebas de hipótesis para muestras pareadas

Experimento

### EJERCICIO

Un gimnasio está ofreciendo un programa de reducción de clorofila y pérdida de peso. Un cliente específico ha pedido que le demuestren la efectividad del programa. Para eso el gimnasio le proporcionó información de datos y le aseguró una pérdida de peso de 10 kg o más en 4 meses. Los datos son:

	1	2	3	4	5	6	7	8	9	10	
ANTES	94.5	101	110	103.5	97	98.5	96.5	101	104	116.5	Normalidad de la
DESPUES	85	89.5	101.5	96	86	80.5	87	93.5	93	102	población
DIFERENCIA	9.5	11.5	8.5	7.5	11	8	9.5	7.5	11	14.5	

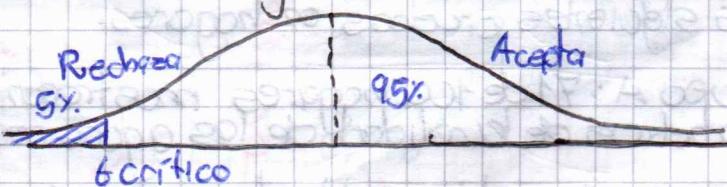
El cliente desea probar con un nivel de significancia del 5% la pérdida de peso obtenida.

### ① Hipótesis

$$H_0: \mu_A - \mu_D \geq 10$$

$$H_a: \mu_A - \mu_D < 10$$

### ② Nivel de significancia



### ③ Tipo de prueba

$$\bar{x} = 9.85 \text{ kg}$$

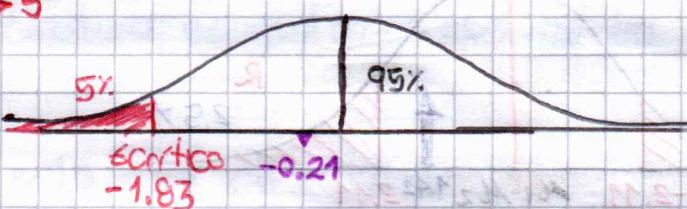
$$S = 2.199 \text{ kg}$$

$$\begin{aligned} &\text{I} \quad \text{DIFERENCIA} \geq 10 \\ &\text{II} \quad \text{DIFERENCIA} < 10 \end{aligned}$$

## ④ Octubre

$$t_{\text{crítico}} = -1.83$$

$$t_{\text{calc}} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{9.85 - 10}{\frac{2.199}{\sqrt{10}}} = -0.21$$



## ⑤ Análisis

Con un  $\alpha = 5\%$ , se acepta la  $H_0$ , por tanto hay evidencia estadística significativa pero débil para afirmar que la pérdida de peso es mayor o igual a 10kg.

## \* Prueba de hipótesis para dos poblaciones $\rightarrow$ Proporciones

### Consideraciones

- ① Comportamiento normal
- ② Siempre  $n_1$  y  $n_2 \geq 30$  datos  
 $n_1$  y  $n_2 \leq 30$  no hay solución.
- ③ Variables cualitativas (% proporciones)

### Ejemplo:

Una empresa que realiza control de parámetros de calidad del agua ha lanzado un nuevo sistema para el tratamiento de aguas residuales domésticos. La empresa ha efectuado los siguientes pruebas en hogares.

Grupo A: 71 de 100 hogares mostraron mejora en los parámetros de la calidad de los aguas residuales

Grupo B: Se utilizó el tratamiento ofrecido por otros empresas. 58 de 90 hogares mostraron mejora en los parámetros de calidad

Se pide que con un nivel de significación de 5%, se compruebe

Si el sistema funciona.

### ① Hipótesis

Mejorat<sup>s</sup> Nomedero  $P_{Nuevo} - P_{TRAD} > 0 \quad H_a$

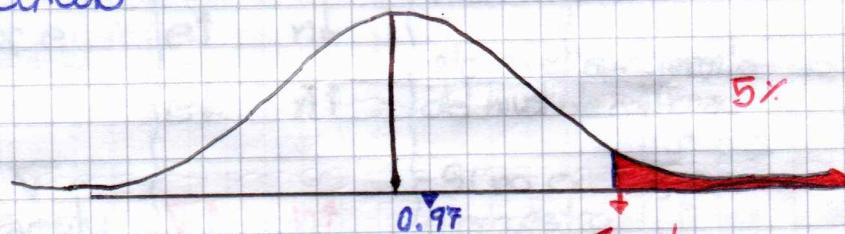
$P_{Nuevo} - P_{TRAD} \leq 0 \quad H_0$

② Nivel de significancia = 5%

③ Definir el hipótesis

$n_1 \text{ y } n_2 \geq 30$   
Proporciones  
Cualitativo

### ④ Cálculo



$$Z_{\text{crítico}} = 1.6448$$

$$Z_{\text{crítico}} = 1.64$$

### ⑤ Análisis

$$Z_{\text{calculado}} = \frac{(P_{Nuevo} - P_T) - (P_N - P_T)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

$\overset{\uparrow}{P_1 \ Q_1} \quad \overset{\uparrow}{P_2 \ Q_2}$  → Ponderada ←

Se acepta  $H_0$  cond = 5%,  
no hay evidencia estadística para afirmar que el sistema funciona mejor.

$$P_{Nuevo} = \frac{71}{100} = 71\% \sim 0.71$$

$$P_{TRAD} = \frac{58}{90} = 64.4\% \sim 0.644$$

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$\hat{p} = \frac{100 * 0.71 + 90 * 0.644}{190}$$

$$\hat{p} = 0.678$$

$$\hat{Q} = 1 - \hat{p} = 1 - 0.678 = 0.322$$

$$Z_{\text{calc}} = \frac{(P_{Nuevo} - P_T) - (P_N - P_T)}{\sqrt{\frac{\hat{p} \hat{Q}}{n_1} + \frac{\hat{p} \hat{Q}}{n_2}}}$$

$$Z_{\text{calc}} = \frac{(0.71 - 0.644) - 0}{\sqrt{\frac{0.678 * 0.322}{100} + \frac{0.678 * 0.322}{90}}} = 0.97 *$$

## CALCULO DE TAMAÑO DE MUESTRA

Medios (Var. Continuas)  
Estrategia  
Proporciones (Var. Categoricas)

$$n = \frac{Z^2 \cdot V^2}{e^2} \quad Z = \frac{\bar{x} - \mu}{\sigma} \rightarrow n = \frac{Z^2 \cdot V^2}{e^2}$$

$\frac{V}{\sqrt{n}}$  → Error

DD MM AA  
10/09/13

Variable	Influencia	Observación	
Nivel de confianza ( $\alpha$ )	$Z \uparrow$	$n \uparrow$	Lo seleccionan los investigadores.
error e	$e \uparrow$	$n \uparrow$	Lo seleccionan los investigadores
$\sigma$	$\sigma \uparrow$	$n \uparrow$	No es el complemento del nivel de confianza
Desviación	$\sigma \uparrow$	$n \uparrow$	- Si no existe $\sigma$ , sacar premuestra. $n_{pre} \geq 30$ si es posible
$N$	Corrección por población finita	$n \uparrow$	- Si existe estecho previo se saca $\sigma$
Tamaño de población	$N \uparrow$	$n \uparrow$	$\rightarrow$ Aparece si la población es finita
	$N \downarrow$	$n \uparrow$	$0.1N \leq n$

$$n = \frac{Z^2 PQ}{e^2} \quad \rightarrow \text{X.}$$

- PQ (Varianibilidad)  $PQ \uparrow n \downarrow$  - si no conoce PQ larga muestra  
 $PQ \downarrow n \uparrow$  - Conoce PQ c/estudios previos  
- Superar máxima variabilidad (E. conservadora)  
 $P=0.5 \quad Q=0.5$

Ejemplo:

Población: Estudiantes  $\Rightarrow N = 1200$



Estudio

Edad promedio al ingresar

$$n = ?$$

$$\sqrt{se\ corrije} = 3.2$$

Hay datos

Preferencia por programas de ciencias sociales

No hay datos  $\rightarrow$  Promuestra

$$n_p = 30$$

SI/NO

$$19 \approx P = 65\%$$

$$11 \approx Q = 35\%$$

Cálculo

- $n_c = 95\% \rightarrow Z$  (dos colas)

$$Z = 1.96$$

- error =  $\pm 1$  año

- error =  $\pm 4$

$$n = \frac{Z^2 \cdot \bar{e}^2}{e^2}$$

$$n = \frac{(1.96^2 * 3.2^2)}{1^2} = 39.34 \approx 40$$

Siempre  
hacer  
arriba.

$$n = \frac{(1.96^2 * (0.65 * 0.35))}{(0.04)^2} = 546.23$$

547

Corrección por población finita

$$n_{corregido} = \frac{n_0 * N}{C_{n_0} + N - 1} = \frac{547 * 1200}{(547 + 1200 - 1)} = 376$$

Solo si hay  
limitaciones

## Análisis de varianza

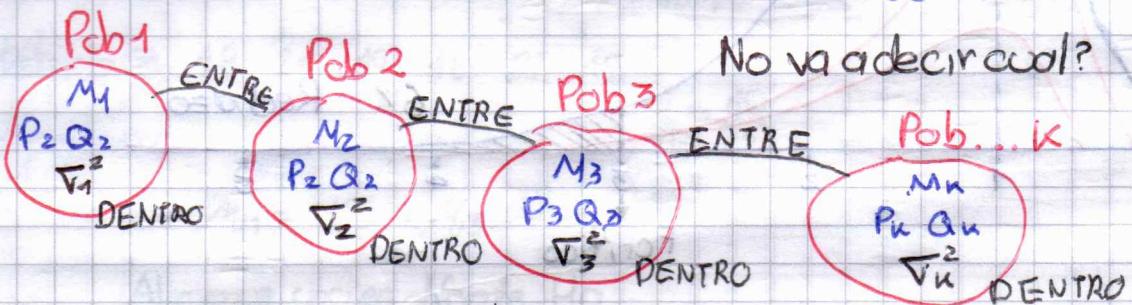
### ANOVA

Test hipótesis

Una población → Comprobar su valor

Dos poblaciones → Verificar la diferencia de dos grupos.

Más de obs → Identificar si al menos uno es diferente.



No va a decir cuál?

Si todas son iguales:

$$H_0: M_1 = M_2 = M_3 = \dots = M_k$$

$H_a$ : Al menos uno es diferente

$$H_0: P_1 = P_2 = P_3 = \dots = P_k$$

$H_a$ : Al menos uno es diferente

ANOVA

CHI-CUADRADO

$\chi^2$

TI-CUADRADO

### CONSIDERACIONES / REQUISITOS

① Normalidad de las poblaciones

② Si todas las poblaciones son iguales

$$\hat{M} = M_1 = M_2 = M_3 = \dots = M_k$$

Si solo uno es  $\neq$   $\Rightarrow$  Rechazo  $H_0$

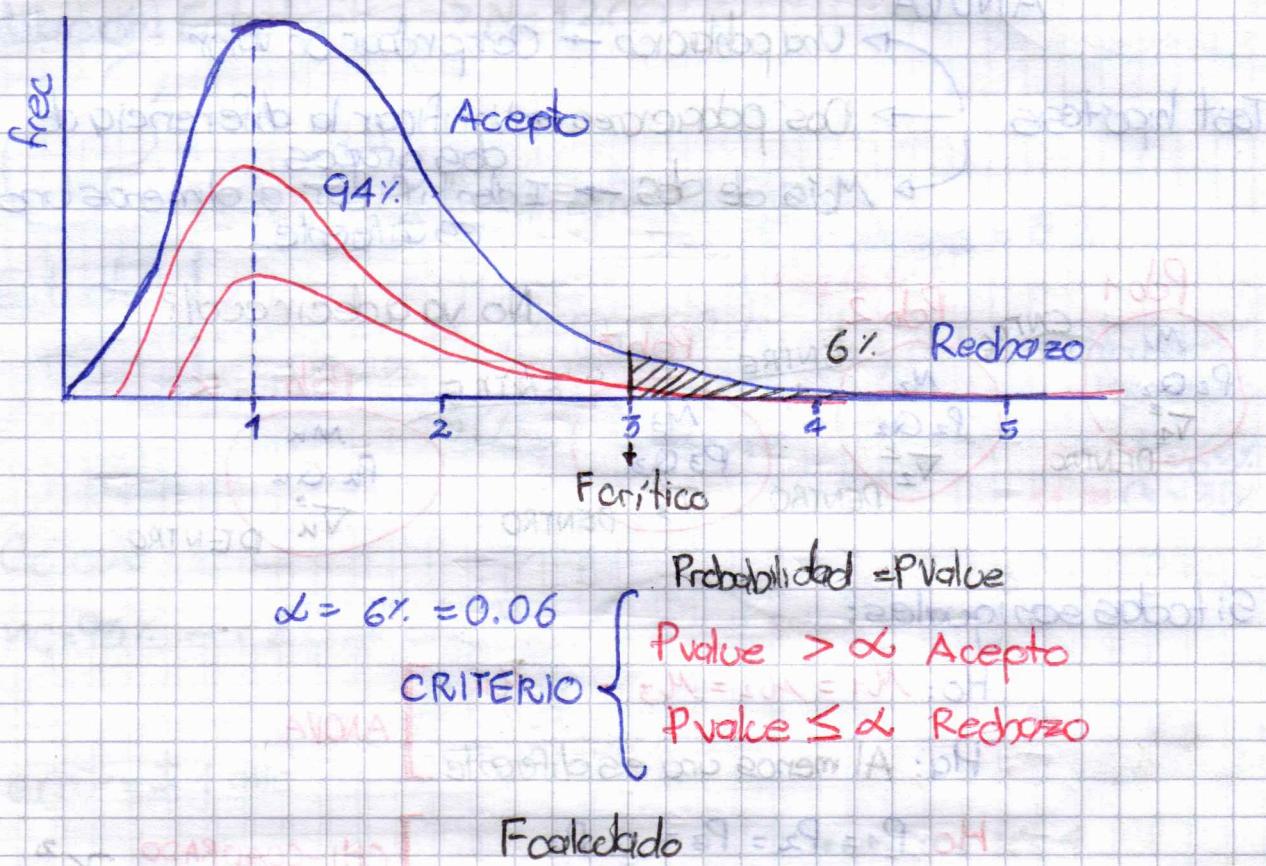
③ Si son iguales

$$V_1^2 = V_2^2 = V_3^2 = \dots = V_k^2 \rightarrow \text{Varianza}$$

$$F = \frac{\text{Var Entre}}{\text{Var Dentro}} \approx 1 \text{ Si son iguales}$$

$\neq \gg$  Son diferentes

## Distribución F (Fisher)



## Ejercicio ANOVA:

El psicólogo de una empresa quiere evaluar tres diferentes métodos de entrenamiento para empleados nuevos.

Método 1: Asigna un empleado nuevo con un trabajador experimental para que este lo asista.

Método 2: Ubicar a todos los empleados nuevos en tres secciones de entrenamiento separadas de la planta.

Método 3: Realizar capacitación utilizando películas de entrenamiento y materiales de aprendizaje programado.

El psicólogo escogió aleatoriamente 16 empleados nuevos asignados a los 3 métodos y registró su productividad media después de terminar los programas de entrenamiento.

## UNIDADES PRODUCIDAS

Método 1 15 18 19 22 11

$$\bar{x}_1 = 17 \\ s_1^2 = 17.5$$

Existen diferencias excepto  
a la efectividad de los  
tres métodos?

Método 2 22 27 18 21 17

$$\bar{x}_2 = 21 \\ s_2^2 = 15.44$$

Método 3 18 24 19 16 22 15

$$\bar{x}_3 = 19 \\ s_3^2 = 12$$

Hipótesis

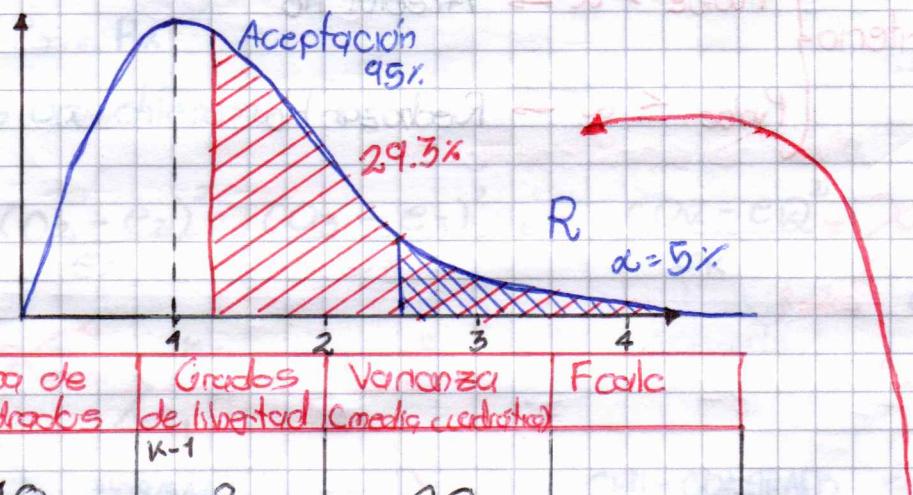
$$\mu_1 = \mu_2 = \mu_3 \quad H_0$$

Al menos uno es diferente  $H_a$ 

Nivel de significancia

$$\alpha = 0.05 \quad 5\%$$

Tabla ANOVA



Fuente de Variación	Suma de Cuadrados numerador	Grados de libertad	Varianza entre (media cuadrática)	Fcalc
ENTRE tratamiento	40	$k-1$	20	$\frac{2.0}{14.8} = 1.35 \Rightarrow 0.293$
DENTRO de tratamiento	192.4	$n-k$	14.8	Pvalue = 29.3%
TOTAL	232.4	15		

Cálculos

$$\text{Varianza entre } \bar{x} = 19 \\ \text{Gran promedio} \quad \bar{x}_1 = 17 \quad \bar{x}_2 = 21 \quad \bar{x}_3 = 19$$

$$V_{\text{entre}} = \frac{(17-19)^2 \times 5 + (21-19)^2 \times 5 + (14-19)^2 \times 6}{3-1}$$

$$\Gamma^z_{\text{entre}} = 20$$

$$\text{Ventre} = \sum_{k=1}^n (\bar{x}_k - \bar{x})^2 \cdot n_k$$

$$\nabla^2_{\text{DENTRO}} = \frac{(17.5 * 4) + (15.44 * 4) + (12 * 5)}{(16 - 3)}$$

$$\nabla^2_{\text{DENTRO}} = 14.8$$

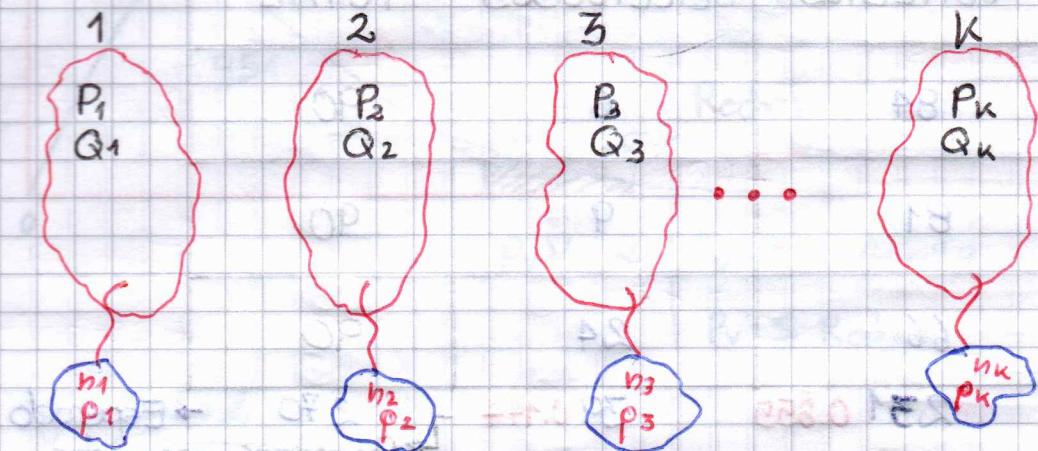
$$\nabla_{\text{DENTRO}} = \frac{\sum (S_u^2 \cdot (C_{u-1}))}{h_T - K}$$

Criterios  $P\text{value} > \alpha \rightarrow \text{Aceptar } H_0$

$P\text{value} \leq \alpha \rightarrow$  Rechazar  $H_0$

# Prueba CHI-CUADRADO / JI-CUADRADO

Comparación de proporciones, tres o más poblaciones



Esperado vs Observado (Real)

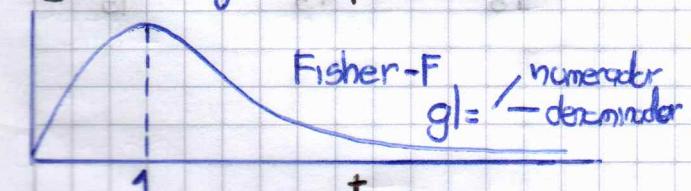
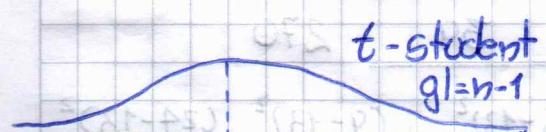
Hipótesis

$$H_0: P_1 = P_2 = P_3 = \dots = P_K$$

$H_a$ : Al menos uno diferente

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \dots + \frac{(O_K - E_K)^2}{E_K} = \chi^2$$

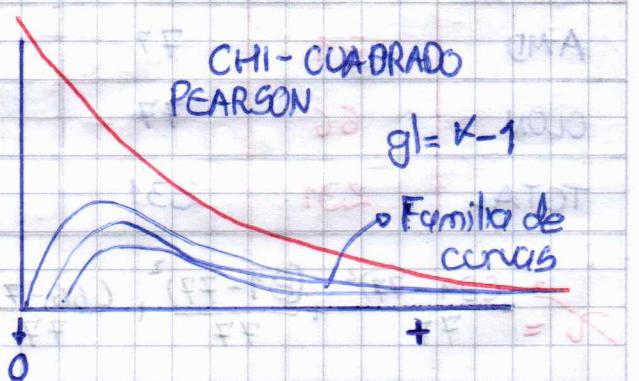
$$\chi^2 = \sum_{n=1}^K \frac{(O_n - E_n)^2}{E_n}$$



CHI-CUADRADO  
PEARSON

$$gl = K - 1$$

Familia de curvas



Ejemplo:

ORIAGO-AUD-16\ORIAGO-AUD-14

Marca	Correctos	Defectuosos	TOTAL
INTEL	84	6	90
AMD	81	9	90
CLON	66	24	90

231 0.855      39 0.144      270 → Esperado

→ Proporciones esperadas

Pregunta:

Existe una marca que tenga diferencias de calidad con las otras?

Hipótesis:

$$H_0: P_{\text{Intel}} = P_{\text{AMD}} = P_{\text{Clon}}$$

$$H_a: Al \text{ menos una diferente}$$

Nivel de significancia

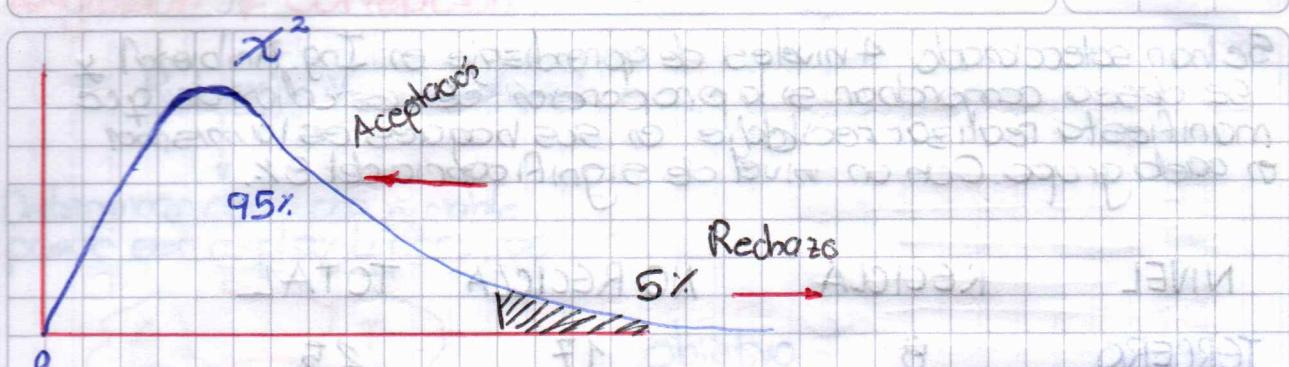
$$\alpha = 5\%$$

## CÁLCULOS

MARCA	Correctos observado	Esperado	Defectuosos observado	Esperado	Total
INTEL	84	$\frac{0.855 \times 90}{77}$	6	$\frac{0.144 \times 90}{13}$	90
AMD	81	77	9	13	90
CLON	66	77	24	13	90
TOTAL	231	231	39	39	270

$$\chi^2 = \frac{(84-77)^2}{77} + \frac{(81-77)^2}{77} + \frac{(66-77)^2}{77} + \frac{(6-13)^2}{13} + \frac{(9-13)^2}{13} + \frac{(24-13)^2}{13}$$

$$\chi^2 = 16.72$$



$$\begin{array}{l|l} P_v < \alpha & | P_v > \alpha \text{ Acepta} \\ & | P_v \leq \text{Rechazo} \end{array}$$

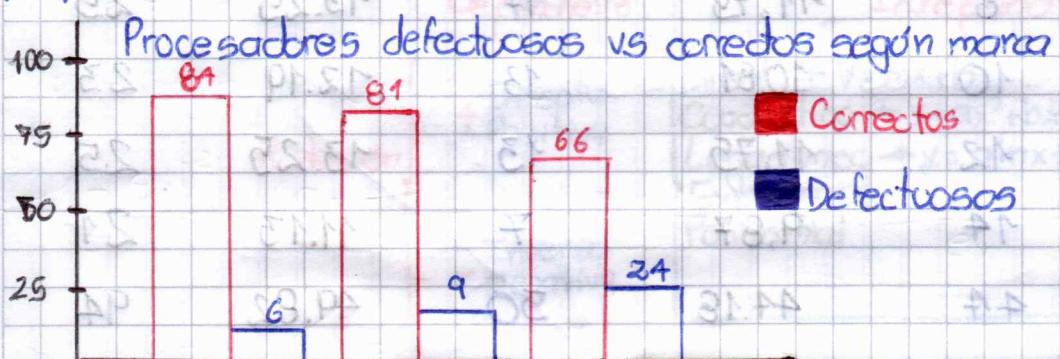
$$gl = 3 - 1 = 2$$

$$\chi^2 = 16.72 \rightarrow P\text{value}$$

$$P\text{value calc} = 2.344 \times 10^{-4}$$

$$0,0002344 << 0,05$$

Rta: Se rechaza  $H_0$ , por tanto con un  $\alpha = 5\%$  al menos uno de los grupos tiene diferente proporción.



Se han seleccionado 4 niveles de aprendizaje en Ing. Ambiental y se desea comprobar si la proporción de estudiantes que manifiesta realizar reciclaje en sus hogares es la misma en cada grupo. Con un nivel de significación del 5%.

NIVEL	RECICLA	NO RECICLA	TOTAL
TERCERO	8	17	25
QUINTO	10	13	23
SEPTIMO	12	13	25
NOVENO	14	7	21
TOTAL	44	50	94

NIVEL	Recicla Observado	Español Observado	No recicla Observado	Español Observado	Total
Tercero	8	11.75	17	13.25	25
Quinto	10	10.81	13	12.19	23
Septimo	12	11.75	13	13.25	25
Noveno	14	9.87	7	11.13	21
Total	44	44.18	50	49.82	94

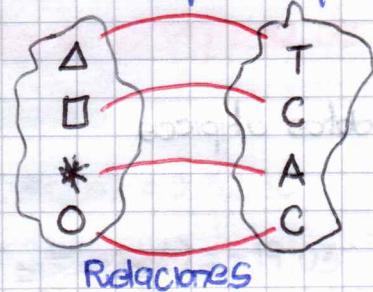
$$\chi^2 = \frac{(8-11.75)^2}{11.75} + \frac{(10-10.81)^2}{10.81} + \frac{(12-11.75)^2}{11.75} + \frac{(14-9.87)^2}{9.87} + \frac{(17-13.25)^2}{13.25} + \frac{(13-12.19)^2}{12.19} + \frac{(17-13.25)^2}{13.25} + \frac{(7-11.13)^2}{11.13}$$

$$\chi^2 =$$

# Regresión y Correlación

## Relación vs Causalidad

Determinar como una variable puede ser explicada por otras



### Ecuaciones

- Lineal
- Logarítmica
- Cuadrática
- Exponencial
- Polinómica

## Relación

Var. Dependiente

- + Directa Positiva
- Inversa Negativa

## Objetivo

- Determinar un modelo matemático (Aprox) que explique el comportamiento de una variable en función de otras

## Regresión simple

una variable

1 Variable Explicada

## Regresión multiple

muchas variables

Estadística Multivariada

Var. Independiente

$$\begin{aligned} \text{Lineal} &\rightarrow y = mx + b \\ \text{Cuadrática} &\rightarrow y = ax^2 + bx + c \\ \text{Logarítmica} &\rightarrow y = a \ln x + b \\ \text{Exponencial} & \\ \text{Polinomial} & \end{aligned}$$

Simple  
Curva

Múltiple  
Curva explicada por  
muchas

## ► Resultados esperados

① Ecuación de predicción

② Coeficiente de regresión (correlación)

③ Grado de explicación de una variable en función de otra.

$$R^2 \quad | \quad 1$$

Coeficiente de  
determinación

% con que la variable independiente explica a la dependiente.

Norma

Excepción

$y$  $x$ 

Dependiente  
Peso(kg)

Independiente  
Estatura(cm)

Intervinientes

Edad

Género

Her. genética

D<sub>189</sub>

Nivel. Act. Física

Dependiente Peso(kg)	Independiente Estatura(cm)
94	184
75	163
72	172
70	167
57	170
70	185
X 45	180
68	173
70	170
65	174
80	170
65	172

→ Cuidado con los datos atípicos

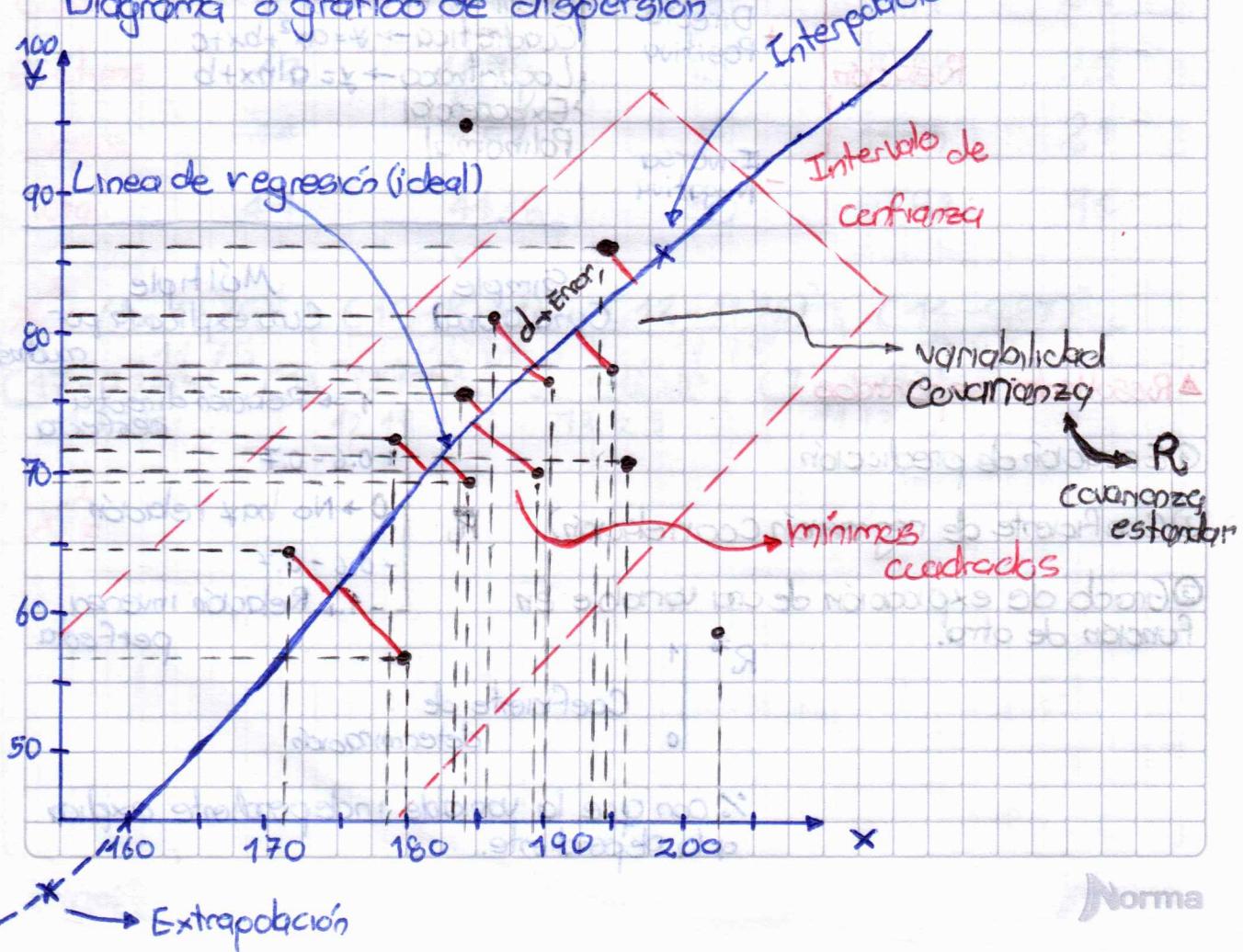
mod

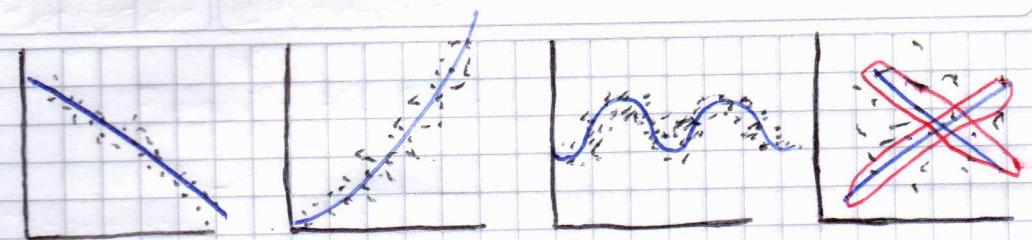
mod

K

K

Diagrama o gráfico de dispersión





Mediante el ingreso de datos a distablos

① Ecuación

$$y = ax + b$$

$$a(\text{slope}) = 0.5238$$

$$b(\text{intercept}) = -19,03$$

Error estandar pendiente  $\approx 0.4507$

$$y = 0.5238x - 19.03$$

$$r^2 = 0.1304 \rightsquigarrow 13,04\%$$

$$r = 0.3612$$

$$b = \frac{\sum xy - n \bar{x} \bar{y}}{\sum x^2 - n \bar{x}^2}$$

$$a = \bar{y} - b \bar{x}$$

$$Se = \sqrt{\frac{\sum y^2 - a \sum y + b \sum xy}{n - 2}}$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$M = \bar{x} \pm z \left( \frac{\sigma}{\sqrt{n}} \right)$$

$$M = \bar{x} \pm t \left( \frac{\sigma}{\sqrt{n}} \right)$$

Corrección finita

$$M = \bar{x} \pm t \left( \frac{\sigma}{\sqrt{n}} \right) \cdot \left( \sqrt{\frac{N-n}{N-1}} \right)$$

Población Proportiones-

$$P = p \pm z \left( \sqrt{\frac{pq}{n}} \right)$$

Dos Poblaciones:

$$M_B - M_A = (\bar{x}_B - \bar{x}_A) \pm z \left( \sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}} \right)$$