UNIVERSIDAD TECNOLÓGICA NACIONAL

FACULTAD REGIONAL CÓRDOBA

# Resumen de Probabilidad y Estadística

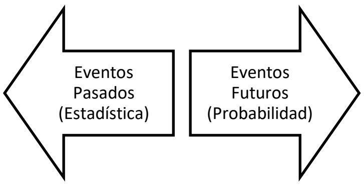
Hecho por Enrique Walter Philippeaux

Con material de clase del Cr. Alejandro Litvinoff

Año: 2021 – Curso: 2R3

# 1 CONTENIDO

# 2 CONCEPTOS GENERALES Y APLICADOS




**Probabilidad:** busca Datos numéricos más o menos subjetivos, evalúa posibilidades de evento futuro no determinístico (puede suceder o no)
**Estadística:** Usado para la toma de decisiones, usando datos pasados.
**Dato:** Es un valor que brinda cierta información

**Dato Estadístico:** usado para realizar el análisis estadístico.

**Método Estadístico:** Pasos para resolver un problema.

1. **Formulación del problema**
2. **Diseño del experimento o investigación**
   Ciertas técnicas sirven para resolver experimentos y otras para resolver investigación, siendo este último el más utilizado en esta materia.
3. **Recolección de datos**
   Como tomamos los datos para trabajarlos.
4. **Procesamiento y análisis de datos**
   Trabajamos los datos con distintas formas y herramientas
5. **Generalización e interferencia final**
   Sacamos conclusiones

Población y Muestra:

- **Población:** Conjunto de todos los elementos del objeto de estudio.
- **Muestra:** Parte de la población a estudiar que sirve para representarla. *Tiene que ser representativa de la población.*

**Razones para el muestreo:** Cuando trabajamos con muestras tenemos resultados aproximados. Razones:

- **Costo.**
- **Tiempo.**
- **Pruebas destructivas.**

**Razones para trabajar con la población:** Cuando necesito:

- **Exactitud de los resultados.**
- **Contar (Censo):** Cuando necesito contar la población.
- **Sensibilidad de los resultados.**

**Tipos de soportes:** Formas en que voy a mostrar los datos

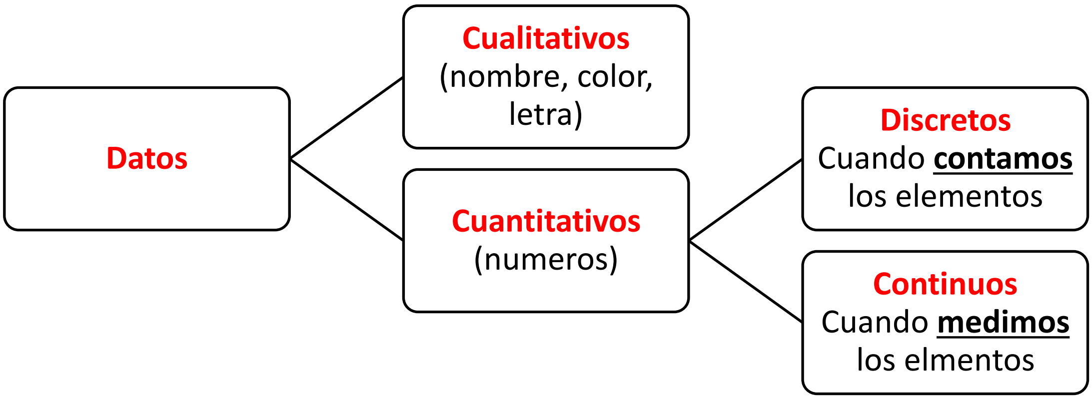
- **Enunciados:** Una oración, cuando tengo pocos datos que mostrar
- **Tablas**
- **Gráficos**

**Distribuciones categóricas:** muestra el número o proporción de observaciones que corresponden a cada clase cualitativa.

| | < 45 años | 45 años o más | Total |
|---|---|---|---|
| No interés | 0 (100%) | 1000 (0%) | 1000 (50%) |
| Interesado | 1000 (0%) | 0 (100%) | 1000 (50%) |
| Total | 1000 (100%) | 1000 (100%) | 2000 (100%) |

## 2.1 DATOS

### 2.1.1 Clasificación según características o según la variable




Hay formulas para datos discretos y otras para continuos.

### 2.1.2 Clasificación según presentación

#### *2.1.2.1 Datos Simples*

Están presentados individualmente

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 16 | 10 | 12 | 15 | 16 | 18 | 10 | 15 | 12 | 15 |

$$\boldsymbol{n} = cantidad\ de\ datos = tamaño\ de\ la\ muestra$$

$$\boldsymbol{N} = cantidad\ de\ datos = tamaño\ de\ la\ población$$

#### *2.1.2.2 Datos Repetidos*

Si tengo una gran cantidad de datos, y hay repetidos, se arman en una tabla según las veces que se repite cada uno:

| $\boldsymbol{X_i}$ | $\boldsymbol{f_i}$ |
|---|---|
| 10 | 2 |
| 12 | 2 |
| 15 | 3 |
| 16 | 2 |
| 18 | 1 |

$$X_i = valor\ del\ dato \qquad f_i = frecuencia\ del\ dato$$

*2.1.2.3 Datos Agrupados*

Datos presentados en intervalos, cuando no se repiten, y son una gran cantidad, se agrupan según a que clase pertenecen

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 16.1 | 10.4 | 12.5 | 15.3 | 16 | 18.5 | 10.5 | 15.2 | 12.1 | 15.6 |

El primer y último valor de cada clase deben estar incluidos en una sola clase, se debe definir de antemano. Para prevenir este problema, ver tabla 3. No puedo convertir los datos agrupados en simples.

| Tabla 1 | |
|---|---|
| **Clase** | $f_i$ |
| 10 a 13 | 4 |
| 13 a 16 | 4 |
| 16 a 19 | 2 |

| Tabla 2 | |
|---|---|
| **Clase** | $f_i$ |
| 10 a 13 | 4 |
| 13 a 16 | 3 |
| 16 a 19 | 3 |

| Tabla 3 | |
|---|---|
| **Clase** | $f_i$ |
| 10.1 a 13 | 4 |
| 13.1 a 16 | 4 |
| 16.1 a 19 | 2 |

**Conceptos cuando** *los limites reales coinciden con los nominales (tablas 1 y 2)*

- **Marca de clase:** Valor que tiene la clase: $MC = (LRI + LRS)/\,2$
- **Ancho de clase:** $c = LRS - LRI$

| Tabla 4 | | | | | |
|---|---|---|---|---|---|
| **Clase** | $f_i$ | LRI | LRS | MC | c |
| 10 a 13 | 4 | 10 | 13 | 11.5 | 3 |
| 13 a 16 | 4 | 13 | 16 | 14.5 | 3 |
| 16 a 19 | 2 | 16 | 19 | 17.5 | 3 |

**Conceptos cuando** *los limites reales* ***NO*** *coinciden con los nominales (tabla 3)*

- **Limite real inferior:** $LRI = (LNI\ de\ f_i + LNS\ de\ f_{i-1})\ /\ 2$
- **Limite real superior:** $LRS = (LNI\ de\ f_{i+1} + LNS\ de\ f_i)\ /\ 2$
- **Marca de clase:** Valor que tiene la clase: $MC = (LRI + LRS)/\,2$
- **Ancho de clase:** $c = LRS - LRI$ , lo calculo usando una de las clases del medio

| Tabla 5 | | | | | |
|---|---|---|---|---|---|
| **Clase** | $f_i$ | LRI | LRS | c | MC |
| 10.1 a 13 | 4 | *1 | 13.05 | 3 | 11.5 |
| 13.1 a 16 | 4 | 13 | 16.05 | 3 | 14.5 |
| 16.1 a 19 | 2 | 16 | *2 | 3 | 17.5 |

*1: No tengo limite real superior de la clase anterior. Uso: $\boldsymbol{LRI = LRS_{fi} - c_{fi}}$
*2: No tengo limite real inferior de la clase siguiente. Uso: $\boldsymbol{LRS = LRI_{fi} + c_{fi}}$

**Distribuciones de frecuencias:** Cuando trabajamos con datos repetidos o agrupados

**Frecuencia absoluta ($f_i$)**: cuantas veces se repite ese dato.
Cuando determinamos el tamaño de la muestra o población, esta será la **sumatoria de frecuencias absolutas**.

$$\boldsymbol{n} = cantidad\ de\ datos = tamaño\ de\ la\ muestra = \sum \boldsymbol{f_i}$$

$$\boldsymbol{N} = cantidad\ de\ datos = tamaño\ de\ la\ población = \sum \boldsymbol{f_i}$$

**Frecuencia acumulada ($F_i$):** Se obtiene de acumular de menor a mayor a las frecuencias absolutas.

**Frecuencia relativa ($fr_i$):** $fr_i = f_i/n$

**Frecuencia acumulada relativa ($Fr_i$):** Debo tener los datos ordenados.

$\boldsymbol{Fr_i} = F_i/n$ o $\boldsymbol{Fr_i} = fr_i + Fr_{i-1}$

**Frecuencia relativa porcentual ($f\%_i$):** $\boldsymbol{f\%_i} = fr_i \times 100$

**Frecuencia acumulada relativa porcentual ($F\%_i$):**

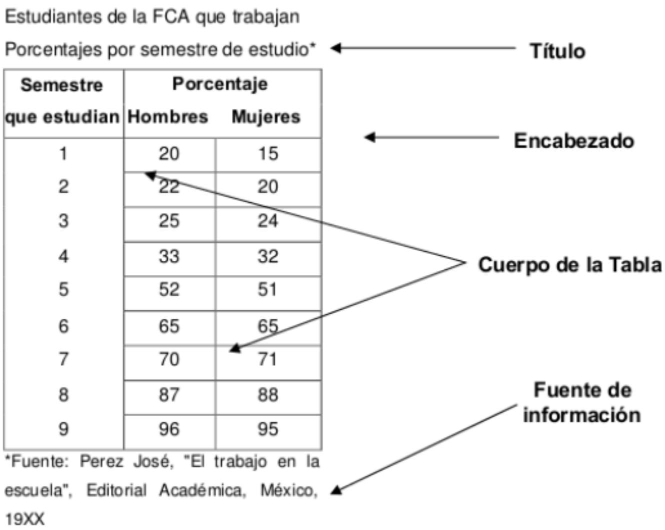
$\boldsymbol{F\%_i} = Fr_i \times 100$ o $\boldsymbol{F\%_i} = f\%_i + F\%_{i-1}$

| $X_i$ | $f_i$ | $F_i$ | $fr_i$ | $Fr_i$ | $f\%_i$ | $F\%_i$ |
|---|---|---|---|---|---|---|
| 10 | 2 | 2 | 0.2 | 0.2 | 20 | 20 |
| 12 | 2 | 4 | 0.2 | 0.4 | 20 | 40 |
| 15 | 3 | 7 | 0.3 | 0.7 | 30 | 70 |
| 16 | 2 | 9 | 0.2 | 0.9 | 20 | 90 |
| 18 | 1 | 10 | 0.1 | 1 | 10 | 100 |
| **n** | 10 | | 1 | | 100 | |

Color Verde: Siempre deben coincidir estos valores.
Color Rojo: Siempre da 1 conceptualmente. Deben coincidir.
Color Celeste: Siempre dará 100. Deben coincidir.

## 2.2 TABLAS Y GRÁFICOS

**Titulo:** Nos presenta la tabla y como están expresados los datos.
**Encabezamiento:** que contiene cada columna.
**Concepto:** Expresado en la primera columna.
**Cuerpo:** los datos
**Nota al pie:** Información importante, y la fuente


Estudiantes de la FCA que trabajan
Porcentajes por semestre de estudio*

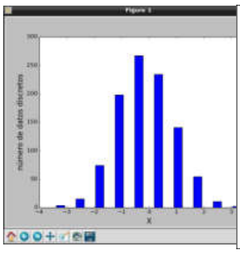
| Semestre que estudian | Porcentaje Hombres | Porcentaje Mujeres |
|---|---|---|
| 1 | 20 | 15 |
| 2 | 22 | 20 |
| 3 | 25 | 24 |
| 4 | 33 | 32 |
| 5 | 52 | 51 |
| 6 | 65 | 65 |
| 7 | 70 | 71 |
| 8 | 87 | 88 |
| 9 | 96 | 95 |

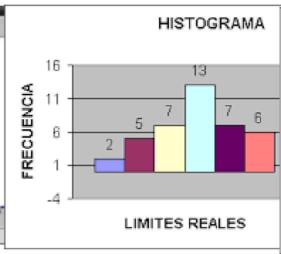
*Fuente: Perez José, "El trabajo en la escuela", Editorial Académica, México, 19XX

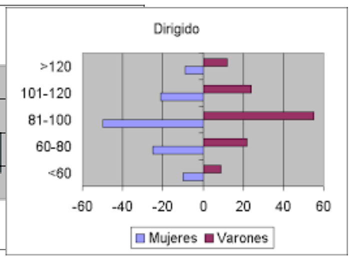


### 2.2.1 Histograma de frecuencias:

Gráficos de barras

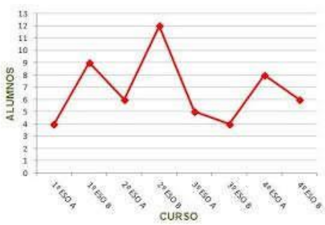



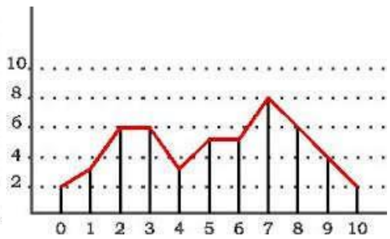



### 2.2.2 Polígono de frecuencias

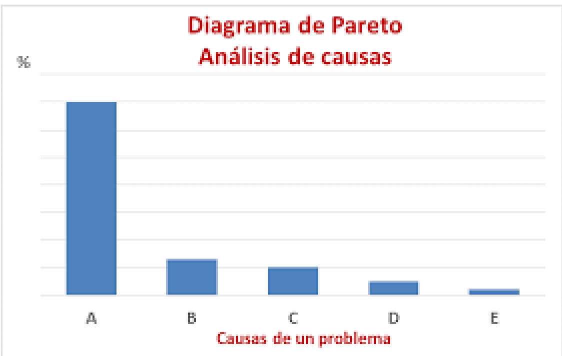
Puntos unidos por línea



### 2.2.3 Diagrama de Pareto

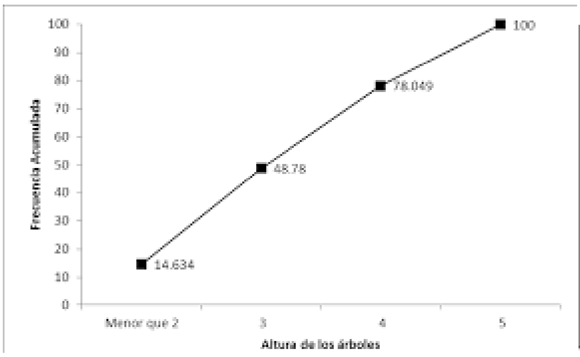
Al principio va lo mas importante (con mayor frecuencia)

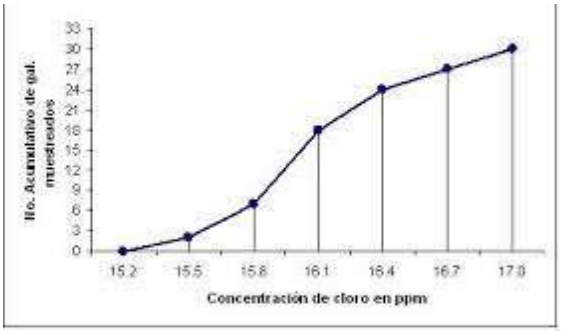



### 2.2.4 Gráfico de Ojiva

Un polígono de frecuencias acumuladas, (cualquiera de las frecuencias acumuladas)

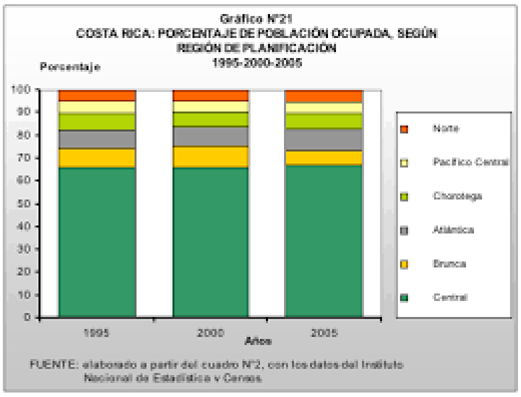
Puede presentar mesetas si no acumula, pero nunca puede tener pendiente negativa.

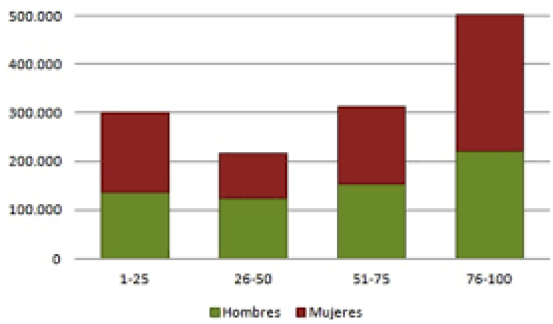



### 2.2.5 Barra Porcentual

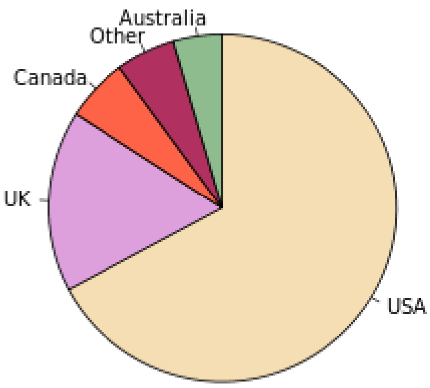
Tienen marcada cierta participación de los datos dentro de la barra.




### 2.2.6 Circulo Radiado

Grafico de torta, necesito tener calculadas las frecuencias relativas.




## 2.3 DISTRIBUCIÓN CATEGÓRICA

Cuando tenemos dos categorías son excluyentes entre sí y a su vez abarcan todas las posibilidades.

La dif. Con las no categóricas, es que cuando tenemos datos cualitativos únicamente podemos sacar cálculos y trabajar si la distribución es categórica

Un dato no puede pertenecer a dos clases distintas.

# 3 MEDIDAS DE POSICIÓN

## 3.1 MEDIA

**Media aritmética o simplemente media**. Promedio aritmético de las observaciones.

$$Media = \frac{Suma\ de\ los\ datos}{Numero\ de\ datos}$$

$$Media\ muestral = \bar{X}\ \ (la\ mas\ usada) \qquad\qquad Media\ poblacional = \mu$$

### 3.1.1 Fórmulas

| Datos Simples | Datos Repetidos | Datos Agrupados |
|---|---|---|
| $\bar{X} = \frac{\sum x_i}{n}$ | $\bar{X} = \frac{\sum x_i . f_i}{n}$ | $\bar{X} = \frac{\sum MC_i . f_i}{n}$ |

$$MC = marca\ de\ clase$$

### 3.1.2 Propiedades

1. La suma de las desviaciones respecto a la media aritmética, es igual a cero
   $\sum (x_i - \bar{x}) = 0 \qquad \sum (x_i - \bar{x}) \ .\ f_i = 0 \qquad \sum (MC_i - \bar{x}) \ .\ f_i = 0$
2. La suma de los cuadrados de las desviaciones es mínima cuando las desviaciones son obtenidas respecto de la media
   $\sum (x_i - \bar{x})^2 = Min \qquad \sum (x_i - \bar{x})^2 \ .\ f_i = Min \qquad \sum (MC_i - \bar{x})^2 \ .\ f_i = Min$
3. La media aritmética de una constante es igual a dicha constante
   $\bar{x}_{(k)} = K$
4. La media aritmética del producto de una constante por una variable es igual a la constante por la media de la variable
   $\bar{x}_{(y.k)} = K \ .\ \bar{x}_{(y)}$
5. La media aritmética de la suma de una variable más una constante es igual a la media de la variable más la constante
   $\bar{x}_{(y+k)} = K + \bar{x}_{(y)}$

## 3.2 MEDIANA

**Es el valor central de la variable de los datos ordenados.**

Es importante que los datos estén ordenados; Cuando los datos están ordenados ya sea de menor a mayor o de mayor a menor es el valor de la variable que supera a no más del 50 % de las observaciones y es superado por no más del 50 % de las observaciones

$$Mediana\ muestral = \tilde{x}\ \ (la\ mas\ usada) \qquad\qquad Mediana\ poblacional = Me$$

#### 3.2.1 Fórmulas

##### *3.2.1.1 Datos Simples*

Estas fórmulas te dan la ubicación y no el valor de la mediana.

| **Mediana muestral** | **Posición**(es) | **Valor** |
|---|---|---|
| **Cantidad de datos impar** | $\frac{(n+1)}{2}$ | Valor del dato |
| **Cantidad de datos par** | $\left(\frac{n}{2}\middle\vert\frac{n}{2}+1\right)$ | Promedio de las dos posiciones |

##### *3.2.1.2 Datos Repetidos*

Igual que con datos simples, pero uso la frecuencia acumulada para encontrar más fácil la posición. Tomo la cantidad de datos (sumatoria de fi + 1) y la divido a 2 (si es impar, usar tabla de arriba y tomar el promedio). Como los datos se encuentran agrupados de manera repetidos, utilizo la frecuencia acumulada. Apenas encuentre un numero mayor al que busco, esa línea de la tabla será el dato.

##### *3.2.1.3 Datos agrupados*

No encontramos la mediana, sino la <u>clase mediana</u>:

$$\tilde{x} = LRI + \frac{\frac{n}{2} - F_{i-1}}{f_i} \times c$$

La **mediana** es igual al limite real inferior + ( n/2 (cantidad de datos divido 2) – la Frecuencia Acumulada de la clase anterior a la mediana ) esto lo dividimos por la frecuencia de la clase mediana, y todo eso lo multiplicamos por el ancho de clase.

## 3.3 MODA

Es el valor $X_i$ que tiene la frecuencia más elevada, el valor que más se repite.

*Es el único valor estadístico que se puede utilizar con datos cualitativos.*

$$Moda\ muestral = \hat{x}\ \ (la\ mas\ usada) \qquad\qquad Moda\ poblacional = Me$$

#### 3.3.1 Fórmulas

| | |
|---|---|
| Datos Simples | Es el dato que más se repite. |
| Datos Repetidos | Es la frecuencia más alta.<br>$f_{max}$ |
| Datos Agrupados | Es la frecuencia más alta, pero así encontramos la clase modal.<br>Para calcular la moda hacemos:<br>$\hat{x} = LRI + \frac{d_1}{d_1 + d_2} \times c$<br>$d_1 = f_i - f_{i-1}$<br>$d_2 = f_i - f_{i+1}$<br>$d_1\ y\ d_2 \rightarrow Siempre\ positivos$ |

# 4 MEDIDAS DE DISPERSIÓN

Me indican que tan dispersos están los datos respectos a un valor de tendencia central. Normalmente este valor de tendencia central es la **media**.

## 4.1 RANGO O RECORRIDO

Es la medida de la variabilidad de una distribución más sencilla. Mide la amplitud de una distribución de frecuencias restándole al valor mayor el valor menor.

| $X_i$ | $f_i$ |
|---|---|
| 2 | 1 |
| 4 | 6 |
| 6 | 8 |
| 8 | 2 |
| 10 | 3 |
| | ***20*** |

$$Rango = Máx - Mín$$

$$Rango_{(Ejemplo)} = 8.00$$

## 4.2 DESVIACIÓN MEDIA (DM)

Es la media aritmética de los valores absolutos de las desviaciones de los valores de la variable con respecto de la media

### 4.2.1 Fórmulas

| Datos Simples | Datos Repetidos | Datos Agrupados |
|---|---|---|
| $DM = \frac{\sum\lvert x_i - \bar{X}\rvert}{n}$ | $DM = \frac{\sum\lvert x_i - \bar{X}\rvert . f_i}{n}$ | $DM = \frac{\sum\lvert MC_i - \bar{X}\rvert . f_i}{n}$ |

### 4.2.2 Ejemplo

| $X_i$ | $f_i$ | $X_i . f_i$ | $\lvert x_i - \bar{X}\rvert$ | $\lvert x_i - \bar{X}\rvert . f_i$ |
|---|---|---|---|---|
| 2 | 1 | 2 | 4 | 4 |
| 4 | 6 | 24 | 2 | 12 |
| 6 | 8 | 48 | 0 | 0 |
| 8 | 2 | 16 | 2 | 4 |
| 10 | 3 | 30 | 4 | 12 |
| | ***20*** | ***120*** | | ***32*** |

$Rango_{(Ejemplo)} = 8.00$ $Media_{(Ejemplo)} = 6.00$ $Desv.\ Media_{(Ejemplo)} = 1.60$

## 4.3 VARIANZA

Es la media aritmética de los cuadrados de las desviaciones de los valores de la variable con respecto de la media de la distribución. USAR SIEMPRE **n-1**

$$Variancia\ de\ muestra = S^2\ (+usada) \qquad Variancia\ Poblacional = \sigma^2 (sigma)$$

### 4.3.1 Fórmulas

| Datos Simples | Datos Repetidos | Datos Agrupados |
|---|---|---|
| $S^2 = \frac{\sum(x_i - \overline{X})^2}{n-1}$ | $S^2 = \frac{\sum(x_i - \overline{X})^2 \cdot f_i}{n-1}$ | $S^2 = \frac{\sum(MC_i - \overline{X})^2 \cdot f_i}{n-1}$ |

La varianza se puede calcular con **n** o **n-1**, siendo esta la más usada, si uso población, uso **n**.

### 4.3.2 Ejemplo

| $X_i$ | $f_i$ | $X_i \cdot f_i$ | $\lvert x_i - \overline{X}\rvert$ | $\lvert x_i - \overline{X}\rvert \cdot f_i$ | $(x_i - \overline{X})^2$ | $(x_i - \overline{X})^2 \cdot f_i$ |
|---|---|---|---|---|---|---|
| 2 | 1 | 2 | 4 | 4 | 16 | 16 |
| 4 | 6 | 24 | 2 | 12 | 4 | 24 |
| 6 | 8 | 48 | 0 | 0 | 0 | 0 |
| 8 | 2 | 16 | 2 | 4 | 4 | 8 |
| 10 | 3 | 30 | 4 | 12 | 16 | 48 |
| | ***20*** | ***120*** | | ***32*** | | ***96*** |

$$Rango = 8.00 \qquad Media = 6.00 \qquad DM = 1.60 \quad Var = 5.0526$$

### 4.3.3 Propiedades

1. La varianza es siempre una cantidad no negativa
   V(x) ≥ 0
2. La varianza de una constante es cero
   V(K) = 0
3. La varianza del producto de una constante por una variable es igual al cuadrado de la constante por la varianza de la variable
   V (KX) = K² . V(X)
4. La varianza de la suma de una variable más una constante es igual a la varianza de la variable
   V(X + K) = V(X)

## 4.4 DESVIACIÓN ESTÁNDAR O TÍPICA

$Variancia\ de\ muestra = S\ (+usada)$ $\qquad$ $Variancia\ Poblacional = \sigma(sigma)$

Se suele usar en vez de variancia para analizar resultados.

La desviación típica o standard, es la raíz cuadrada, con signo positivo, de la varianza.

$$S = \sqrt{S^2}$$

### 4.4.1 Fórmulas

| Datos Simples | Datos Repetidos | Datos Agrupados |
|---|---|---|
| $S^2 = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}}$ | $S^2 = \sqrt{\frac{\sum(x_i - \bar{X})^2 \cdot f_i}{n-1}}$ | $S^2 = \sqrt{\frac{\sum(MC_i - \bar{X})^2 . f_i}{n-1}}$ |

### 4.4.2 Ejemplo

| $X_i$ | $f_i$ | $X_i \cdot f_i$ | $\lvert x_i - \bar{X}\rvert$ | $\lvert x_i - \bar{X}\rvert . f_i$ | $(x_i - \bar{X})^2$ | $(x_i - \bar{X})^2 . f_i$ |
|---|---|---|---|---|---|---|
| 2 | 1 | 2 | 4 | 4 | 16 | 16 |
| 4 | 6 | 24 | 2 | 12 | 4 | 24 |
| 6 | 8 | 48 | 0 | 0 | 0 | 0 |
| 8 | 2 | 16 | 2 | 4 | 4 | 8 |
| 10 | 3 | 30 | 4 | 12 | 16 | 48 |
| | ***20*** | ***120*** | | ***32*** | | ***96*** |

$Rango = 8.00$ $\quad$ $Media = 6.00$ $\quad$ $DM = 1.60$ $\quad$ $Var = 5.0526$ $\quad$ $DS = 2.2478$

### 4.4.3 Propiedades

1. La desviación típica es siempre un valor no negativo

   DS(x) ≥ 0

2. La desviación típica de una constante es cero

   DS(K) = 0

3. Si a todos los valores de la variable se multiplican por una misma constante, la desviación típica queda multiplicada por el valor absoluto de dicha constante

   DS (KX) = K .DS(X)

4. Si a todos los valores de la variable se le suma una misma constante la desviación típica no varía.

   DS(X + K) = DS(X)

### 4.4.4 Coeficiente de variación

El coeficiente de variación es la relación entre la desviación típica de una muestra y su media

$$C.V. = \frac{S}{\bar{X}}$$

# 5 MOMENTOS

- Es otro enfoque para la explicación del Análisis Estadístico Simple
- Son los valores esperados de ciertas funciones, un generador de fórmulas
- Lo necesitamos para definir medidas de Asimetría y Kurtosis
- Los momentos de orden par serán positivos y los impares positivos o negativos

## 5.1 MOMENTO DE UNA VARIABLE

Se define como momento de orden "**r**" de una variable "**x**" a:

$$\overline{X^r} = \frac{\sum x_i^r}{n} \qquad datos\ repetidos\ y\ agrupados: \qquad \overline{X^r} = \frac{\sum x_i^r . f_i}{n}$$

**Casos particulares:** $\overline{X^0} = 1 \quad ; \quad \overline{X^1} = \bar{X} = Media$

## 5.2 MOMENTO RESPECTO DE UN VALOR

Al calcularlos para el coeficiente Fisher y coeficiente de Kurtosis usar n-1 en el denominador.

***Este es el más utilizado.*** Se define como momento de orden "**r**" de una variable "**x**" respecto de un cierto valor **K**:

### 5.2.1 Fórmulas

| Datos Simples | Datos Repetidos | Datos Agrupados |
|---|---|---|
| $m_r = \frac{\sum (x_i - K)^r}{n}$ | $m_r = \frac{\sum (x_i - K)^r . f_i}{n}$ | $m_r = \frac{\sum (MC_i - K)^r . f_i}{n}$ |

*Lo más frecuente es que **K** sea la media.*

**Casos particulares:** $m_0 = 1 \quad ; \quad m_1 = 0 \quad ; \quad m_2 = s^2$

### 5.2.2 Ejemplo

| $X_i$ | $f_i$ | $X_i . f_i$ | $(x_i - \bar{X})^2$ | $(x_i - \bar{X})^2 . f_i$ | $(x_i - \bar{X})^5$ | $(x_i - \bar{X})^5 . f_i$ |
|---|---|---|---|---|---|---|
| 2 | 1 | 2 | 16 | 16 | -1024 | -1024 |
| 4 | 6 | 24 | 4 | 24 | -32 | -192 |
| 6 | 8 | 48 | 0 | 0 | 0 | 0 |
| 8 | 2 | 16 | 4 | 8 | 32 | 64 |
| 10 | 3 | 30 | 16 | 48 | 1024 | 3072 |
| | ***20*** | ***120*** | | ***96*** | | ***1920*** |

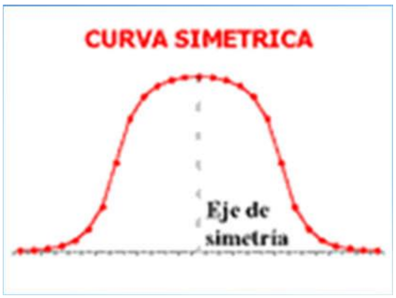
$$Rango_{(Ejemplo)} = 8.00 \quad Media_{(Ejemplo)} = 6.00$$

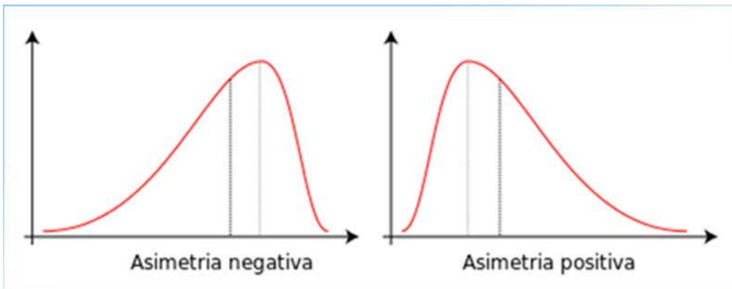
$$m_5 = \frac{\sum (x_i - 6)^5 . f_i}{20} = 96$$

# 6 MEDIDAS DE FORMA

## 6.1 ASIMETRÍA O SESGO

El concepto de asimetría o sesgo se refiere a si la curva que forman los valores de la serie presenta la misma forma a izquierda y derecha de un valor central (media aritmética).




$Media = Mediana = Moda$ $\qquad$ $Simetrica.$

$Asimetria\ negativa:$ $\qquad$ $Media < Mediana < Moda$

$Asimetria\ positiva:$ $\qquad$ $Moda < Mediana < Media$

### 6.1.1 Coeficientes de Pearson

Se basan en relaciones entre las medidas de centralización:

#### 6.1.1.1 Fórmula

$$S_{K1} = \frac{\bar{X} - \hat{x}}{s} \quad (media - moda)$$

$$S_{K2} = \frac{3(\bar{X} - \tilde{x})}{s} \quad 3(media - mediana)$$

- Si $S_K > 0$, la distribución es asimétrica positiva o a la derecha.
- Si $S_K < 0$, la distribución es asimétrica negativa o a la izquierda
- Si $S_K = 0$, la distribución es simétrica (la moda, la media y la mediana son iguales)

#### 6.1.1.2 Ejemplo

| $X_i$ | $f_i$ | $X_i \cdot f_i$ | $(x_i - \bar{X})^2$ | $(x_i - \bar{X})^2 . f_i$ | $F_i$ |
|---|---|---|---|---|---|
| 2 | 6 | 12 | 19.36 | 116.16 | 6 |
| 4 | 11 | 44 | 5.76 | 63.36 | 17 |
| 6 | 8 | 48 | 0.16 | 1.28 | 25 |
| 8 | 17 | 136 | 2.56 | 43.52 | 42 |
| 10 | 8 | 80 | 12.96 | 103.68 | 50 |
| | ***50*** | ***320*** | | ***328*** | |

| | |
|---|---|
| $\bar{X}$ | 6.4 |
| $\hat{x}$ | 8.0 |
| $S^2$ | 6.6939 |
| $S$ | 2.5873 |
| $\tilde{x}$ | 7.0 |
| $S_{K1}$ | -0.6184 |
| $S_{K2}$ | -0.6957 |

### 6.1.2 Coeficiente de asimetría de Fisher

#### *6.1.2.1 Fórmula*

$$a_3 = \frac{m_3}{S^3}\ (formula\ de\ momento) \qquad a_3 = \frac{\frac{\sum(x_i - \bar{X})^3 . f_i}{n}}{S^3}$$

- Si $a_3 > 0$, la distribución es asimétrica positiva o a la derecha.
- Si $a_3 < 0$, la distribución es asimétrica negativa o a la izquierda
- Si $a_3 = 0$, la distribución es simétrica

#### *6.1.2.2 Ejemplo*

| $\boldsymbol{X_i}$ | $\boldsymbol{f_i}$ | $\boldsymbol{X_i . f_i}$ | $\boldsymbol{(x_i - \bar{X})^2}$ | $\boldsymbol{(x_i - \bar{X})^2 . f_i}$ | $\boldsymbol{(x_i - \bar{X})^3}$ | $\boldsymbol{(x_i - \bar{X})^3 . f_i}$ |
|---|---|---|---|---|---|---|
| 2 | 6 | 12 | 19.36 | 116.16 | -85.184 | -511.104 |
| 4 | 11 | 44 | 5.76 | 63.36 | -13.824 | -152.054 |
| 6 | 8 | 48 | 0.16 | 1.28 | -0.064 | -0.512 |
| 8 | 17 | 136 | 2.56 | 43.52 | 4.096 | 69.632 |
| 10 | 8 | 80 | 12.96 | 103.68 | 46.656 | 373.248 |
| | ***50*** | ***320*** | | ***328*** | | ***-220.8*** |

| | |
|---|---|
| $\boldsymbol{\bar{X}}$ | 6.4 |
| $\boldsymbol{S^2}$ | 6.6939 |
| $\boldsymbol{S}$ | 2.5873 |
| $\boldsymbol{m_3}$ | -4.4160 |

Acá ocurre un problema, ya que al momento lo calculamos dividiendo por 50, y la varianza con la que trabajamos, tiene en el denominador 49.

Por eso, se debe trabajar ambos con denominador 50:

| | |
|---|---|
| $\boldsymbol{S^2}$ | 6.56 |
| $\boldsymbol{S}$ | 2.5612 |

| | | |
|---|---|---|
| $\boldsymbol{a_3}$ | -0.2628 | *Resultado Correcto* |

O también podemos calcular el momento con n-1 y nos ahorramos calculo.

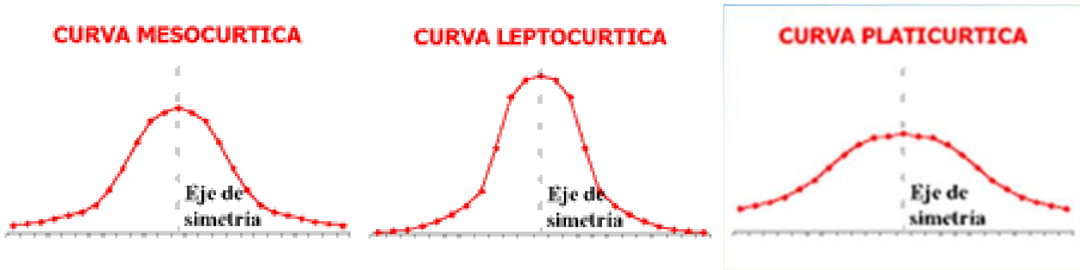
| | |
|---|---|
| $\boldsymbol{m_3}$ | -4.5061 |

| | | |
|---|---|---|
| $\boldsymbol{a_3}$ | -0.2602 | *Resultado incorrecto pero aceptado* |

En la practica usaremos esta última forma.

## 6.2 KURTOSIS

Analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución




### 6.2.1 Coeficiente de Kurtosis

#### *6.2.1.1 Fórmula*

$$a_4 = \frac{m_4}{S^4}\ (formula\ de\ momento) \qquad a_4 = \frac{\frac{\sum (x_i - \bar{X})^4 . f_i}{n}}{S^4}$$

- Si $a_4 > 3$, Distribución leptocúrtica, presenta un elevado grado de concentración alrededor de los valores centrales de la variable. Es más apuntada que la normal
- Si $a_4 < 3$, Distribución platicúrtica, presenta un reducido grado de concentración alrededor de los valores centrales de la variable. Es más aplastada que la normal
- Si $a_4 = 3$, Distribución mesocúrtica, presenta un grado de concentración medio alrededor de los valores centrales de la variable. Es igual que la normal

#### *6.2.1.2 Ejemplo*

| $X_i$ | $f_i$ | $X_i . f_i$ | $(x_i - \bar{X})^2$ | $(x_i - \bar{X})^2 . f_i$ | $(x_i - \bar{X})^4$ | $(x_i - \bar{X})^4 . f_i$ |
|---|---|---|---|---|---|---|
| 2 | 6 | 12 | 19.36 | 116.16 | 374.8096 | 2248.8576 |
| 4 | 11 | 44 | 5.76 | 63.36 | 33.1776 | 364.9536 |
| 6 | 8 | 48 | 0.16 | 1.28 | 0.0256 | 0.2048 |
| 8 | 17 | 136 | 2.56 | 43.52 | 6.5536 | 111.4112 |
| 10 | 8 | 80 | 12.96 | 103.68 | 167.9616 | 1343.6928 |
| | ***50*** | ***320*** | | ***328*** | | ***4069.12*** |

| | |
|---|---|
| $m_1$ | 81.3824 |

| | | |
|---|---|---|
| $a_4$ | 1.8911 | *Resultado Correcto* |

Como en el coeficiente de Fisher, hacemos el cálculo utilizando **n-1** para el denominador del momento:

| | |
|---|---|
| $m_4$ | 83.0433 |

| | | |
|---|---|---|
| $a_4$ | 1.8533 | *Resultado incorrecto pero aceptado* |

# 7 PROBABILIDAD

## 7.1 CONCEPTOS

- **Espacio muestral:**
  El conjunto de todos los resultados posibles de un experimento estadístico se llama espacio muestral, y se representa por el símbolo S
- **Suceso:**
  Cada elemento en un espacio muestral se llama Suceso o punto muestral.
  Si el espacio muestral tiene un numero finito de elementos, podemos listarlos
- **Eventos:**
  Uno o más elementos forman un evento. Un evento es un subconjunto de un espacio muestral
- **Regla de Laplace:**
  Si realizamos un experimento aleatorio en el que hay n elementos, todos igualmente probables, entonces si A es un suceso, la probabilidad de que ocurra el suceso A es:

$$P(A) = \frac{número\ de\ casos\ favorables\ a\ A}{número\ de\ casos\ posibles}$$

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

- **Eventos independientes:**
  *La ocurrencia de un evento no afecta en nada a la ocurrencia de otro.*
  *La probabilidad de que pase B sabiendo que haya pasado A, es igual a la de B.*
  *La probabilidad de que pase A sabiendo que haya pasado B, es igual a la de A.*

$$P\left(\frac{B}{A}\right) = P(B) \quad y \quad P\left(\frac{A}{B}\right) = P(A)$$

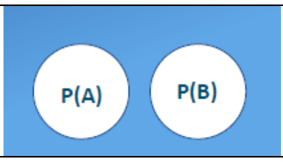
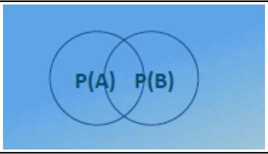
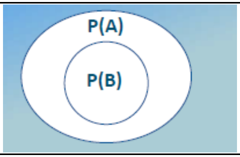
- **Eventos mutuamente excluyentes:**
  *Si ocurre A, no puede ocurrir B. Si ocurre B, no puede ocurrir A.*
  *La probabilidad de que ocurran A y B al mismo tiempo es 0.*

$$P(A \cap B) = 0$$

- **Probabilidad de un evento:**
  *Se utilizarán herramientas de análisis de problemas, en especial la <u>teoría de conjuntos</u>.*

## 7.2 TEORÍA DE CONJUNTOS

*Es una forma de graficar problemas para poder entenderlos de una forma mejor.*

| | | |
|---|---|---|
|   |   |   |
| **Eventos mutuamente excluyentes.**<br>*No hay espacio en común por que no hay ninguna posibilidad de que ocurran A y B mutuamente.* | **No son eventos mutuamente excluyentes.**<br>*Tienen un espacio en común. Pueden o no ocurrir A y B mutuamente.* | **No son eventos mutuamente excluyentes.**<br>*Un espacio abarca al otro.*<br>*Si ocurre B, ocurre A.* |
| **No son eventos independientes.**<br>*Por qué, si pasa A, la probabilidad de que pase B es 0. Si pasa B, la probabilidad de que pase A es 0.* | **Pueden o no ser independientes.**<br>*Son independientes únicamente cuando la intersección es igual a la probabilidad de A por la probabilidad de B.*<br>$P(A \cap B) = P(A).P(B)$ | **No son eventos independientes.**<br>*La probabilidad de B si pasa A, es mucho alta que si no pasa B.*<br>*Si no pasa A, la probabilidad de B es 0.*<br>*Si pasa B, la probabilidad de A es 1.* |

### 7.2.1 Reglas de adición

Adicionar dos probabilidades es la probabilidad de una unión de sucesos.

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

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

### 7.2.2 Probabilidad condicional

Sean A y B dos sucesos tal que $P(A) = 0$ , se llama probabilidad de B condicionada a A, $P(B/A)$, a la probabilidad de B tomando como espacio muestral A, es decir, la probabilidad de que ocurra B dado que ha sucedido A.

*La probabilidad de B, habiendo pasado A.*

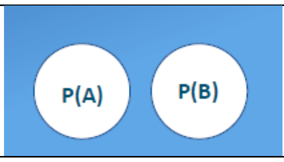
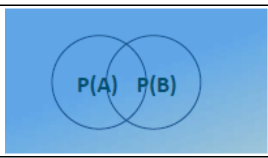
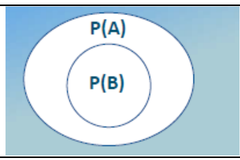
$$P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)} \quad si \quad P(A) > 0$$

Para eventos independientes: $P(B/A) = P(B)$

### 7.2.3 Reglas de multiplicación

Si pasa un evento y otro simultáneamente

$$P(A \cap B) = P(A).P(\frac{B}{A})$$

| | | |
|---|---|---|
|   |   |   |
| $P(A \cap B) = 0$ | $P(A \cap B) = P(A).P(\frac{B}{A})$ | $P(A \cap B) = P(B)$ |
| | **Sólo cuando son independientes:**<br>$P(A \cap B) = P(A).P(B)$ | |

## 7.3 TÉCNICAS DE CONTEO

Sirven para definir el número de casos favorables, y el número de casos posibles.

Las técnicas de conteo son un ordenamiento de los elementos de un conjunto. Pueden tener dos formas dependiendo del orden, y dos formas dependiendo si hay elementos repetidos.

| | **El orden no importa** | **El orden importa** |
|---|---|---|
| **Se pueden repetir los elementos** | Combinaciones con repetición | Permutaciones con repetición |
| **No se pueden repetir los elementos** | Combinaciones sin repetición | Permutaciones sin repetición |

Ejemplo: Tenemos un grupo de 3 elementos $(A; B; C)$ y queremos hacer grupos de dos elementos:

| Combinaciones sin repetición $_n\boldsymbol{C}_x$ | Combinaciones con repetición $_n\boldsymbol{Cr}_x$ | Permutaciones sin repetición $_n\boldsymbol{P}_x$ | Permutaciones con repetición $_n\boldsymbol{Pr}_x$ |
|---|---|---|---|
| A-B<br>A-C<br>B-C | A-B<br>A-C<br>B-C<br>A-A<br>B-B<br>B-C | A-B B-A<br>A-C C-A<br>B-C C-B | A-B B-A<br>A-C C-A<br>B-C C-B<br>A-A<br>B-B<br>C-C |
| **Formula**<br>$_nC_x = \frac{n!}{x!\,(n-x)!}$ | **Formula**<br>$_nCr_x = \frac{(n+x-1)!}{x!\,(n-1)!}$ | **Formula**<br>$_nP_x = \frac{n!}{(n-x)!}$ | **Formula**<br>$_nPr_x = n^x$ |
| $_3C_2 = \frac{3!}{2!\,(3-2)!}$<br>$_3C_2 = 3$ | $_3Cr_2 = \frac{(3+2-1)!}{2!\,(3-1)!}$<br>$_3Cr_2 = 6$ | $_3P_2 = \frac{3!}{(3-2)!}$<br>$_3P_2 = 6$ | $_3Pr_2 = 3^2$<br>$_3Pr_2 = 9$ |

$$n = cant. total\ de\ elementos$$

$$x = cant. de\ elementos\ x\ grupo$$

La función **factorial** se representa con un signo de exclamación "**!**" detrás de un número. Esta exclamación dice **que** hay **que** multiplicar todos los números enteros positivos **que** hay entre ese número y el 1.

$5! = 5.4.3.2.1 \quad ; \quad 3! = 3.2.1 \quad ; \quad 0! = 1$

$$\frac{5!}{3!} = 5.4$$

## 7.4 DIAGRAMAS DE ÁRBOL

El diagrama de árbol es una representación gráfica de los posibles resultados del experimento, el cual consta de una serie de pasos, donde cada uno de estos tiene un número infinito de maneras de ser llevado a cabo.

### 7.4.1 Construcción

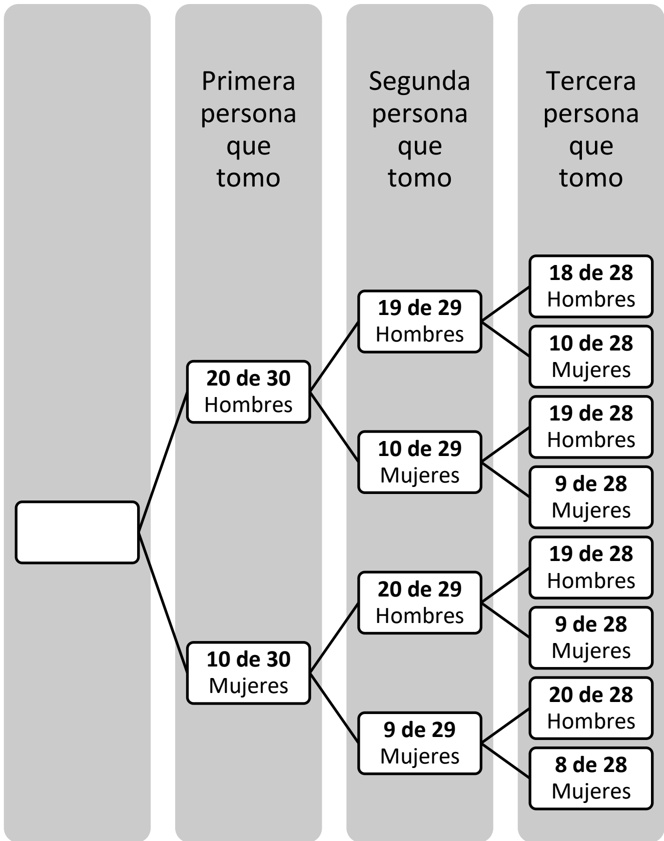
Para la construcción de un diagrama en árbol se partirá poniendo una rama para cada una de las posibilidades, acompañada de su probabilidad. En el final de cada rama parcial se constituye a su vez, un nudo del cual parten nuevas ramas, según las posibilidades del siguiente paso, salvo si el nudo representa un posible final del experimento (nudo final).

Hay que tener en cuenta: que la suma de probabilidades de las ramas de cada nudo ha de dar 1. Para los cálculos de probabilidad: multiplicamos las probabilidades si se trata de ramas contiguas (horizontal) y se suman las probabilidades entre las ramas (vertical)

### 7.4.2 Ejemplo

Ejemplo: Si tenemos en clase 10 mujeres y 20 varones y necesitamos elegir 3 representantes:

*Cada vez que sacamos una persona, la probabilidad de los restantes van a cambiar.*



## 7.5 TABLA DE CONTINGENCIA

Se emplean para registrar y analizar la relación entre dos o más variables, habitualmente de naturaleza cualitativa

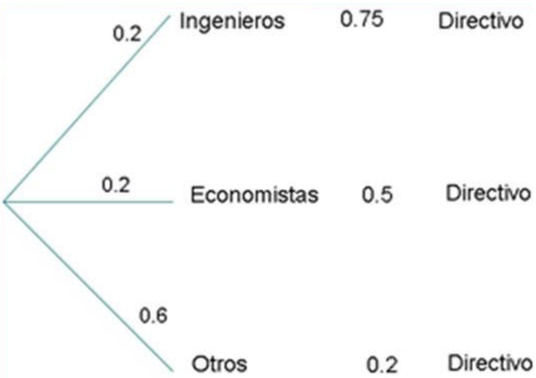
Cuenta las observaciones por múltiples variables categóricas. Las filas y columnas de las tablas corresponden a estas variables categóricas.

### 7.5.1 Ejemplo

Si tenemos en clase 15 mujeres y 25 varones y sabemos que tercio de las mujeres son mayores de 21 años y 15 hombres son mayores.

| | **Hombres** | **Mujeres** | ***SUMA*** |
|---|---|---|---|
| **Mayores** | 0.375 | 0.125 | *0.5* |
| **Menores** | 0.25 | 0.25 | *0.5* |
| ***SUMA*** | *0.625* | *0.375* | *1* |

## 7.6 TEOREMA DE BAYES




- **Las probabilidades P(D) se denominan probabilidades a priori.**
  *La probabilidad de encontrar, tomando una **persona** al azar, a un ingeniero que sea directivo es de* $0{,}2 \,.\, 0{,}75 = 0{,}15$

- **Las probabilidades P(D/I) se denominan probabilidades a posteriori.**
  *La probabilidad de encontrar un directivo, tomando un **ingeniero** al azar, es de* $0{,}75$

- **Las probabilidades P(I/D) se denominan verosimilitudes.**
  *La probabilidad de encontrar un ingeniero, tomando un **directivo** al azar, se calcula abajo:*

---

Si $A_1, A_2, \ldots, A_n$ son: Sucesos mutuamente excluyentes Y cuya unión es el espacio muestral ($A_1 \; A_2 \; \ldots \; A_n = E$ ). Y B es otro suceso.

La probabilidad de un suceso especial $A_i$ , dado un suceso $B$, la probabilidad de este se da por la ecuación:

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i).P\left(\frac{B}{A_i}\right)}{P(A_1).P\left(\frac{B}{A_1}\right) + P(A_2).P\left(\frac{B}{A_2}\right) + \cdots + P(A_n).P\left(\frac{B}{A_n}\right)}$$

---

La probabilidad de que sea ingeniero multiplicado por la condicional de que sea directivo, sabiendo que es ingeniero *(es decir, parto de la base de que sé que es directivo, y quiero saber la probabilidad de que sea ingeniero):*

$$P\left(\frac{ingeniero}{directivo}\right) = \frac{0{,}2\,.\,0{,}75}{0{,}2\,.\,0{,}75 + 0{,}2\,.\,0{,}5 + 0{,}6\,.\,0{,}2} = 0{,}405$$

# 8 VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

## 8.1 CONCEPTOS

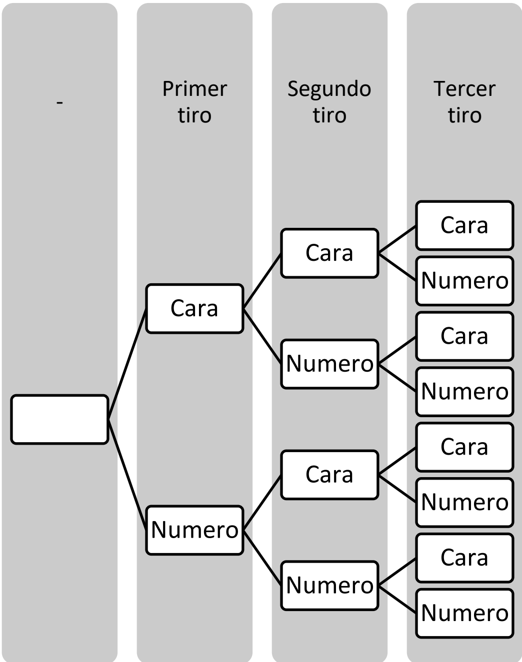
### 8.1.1 Variable aleatoria

Una variable aleatoria es toda función que asocia un número real con cada elemento del espacio muestral, cuyos valores posibles pueden representar los posibles resultados de un experimento aún no realizado.

### 8.1.2 Distribuciones de probabilidad

La Distribución de probabilidad de una variable aleatoria es la representación de las probabilidades de todos y cada uno de los posibles valores que tome dicha variable.

*Ejemplo:* Tres tiros de una moneda de 2 caras




La posibilidad de que salga 3 veces cara es de

$$0{,}5 * 0{,}5 * 0{,}5 = 0{,}125$$

Con esto elaboramos una distribución de probabilidad:

Los valores posibles: cantidad de **caras** que tengo en cada rama, es decir: cantidad de **caras** que me pueden salir.

Con esto planteo todos los resultados posibles de la variable aleatoria y la probabilidad de cada uno ellos.

| $X_i$ | $P$ |
|---|---|
| 0 | 0,125 |
| 1 | 0,375 |
| 2 | 0,375 |
| 3 | 0,125 |
| | 1 |

La distribución de probabilidad nos sirve para aplicar modelos.

### 8.1.3 Modelos según variable

Hay modelos de distribución de probabilidades **Discretos** para *variables discretas,* y **Continuos** para *variables continuas.*

# 9 DISTRIBUCIONES DE PROBABILIDAD DISCRETAS

## 9.1 DISTRIBUCIÓN DE BERNOULLI

Se utiliza para el caso más sencillo en el que la variable aleatoria es dicotómica, es decir que puede tomar sólo dos valores. Estos **dos únicos valores son el 0 y el 1**. Este experimento consiste en observar una sola unidad experimental y clasificarla en una de dos categorías mutuamente excluyentes (si sucede una no puede suceder la otra) y exhaustivas (no puedo tener 0 1 2, solo 0 = desfavorable y 1 = favorable).

## 9.2 DISTRIBUCIÓN BINOMIAL

Es una repetición del experimento de Bernoulli.

Un experimento de Bernoulli se caracteriza por ser dicotómico, esto es, sólo son posibles dos resultados. A uno de estos se denomina éxito y tiene una probabilidad de ocurrencia **p** y al otro, fracaso, con una probabilidad **q = 1 - p**.

Distribución binomial es una **distribución de probabilidad discreta** que cuenta el número de éxitos en una secuencia de ***n*** *ensayos de Bernoulli independientes* entre sí, con una probabilidad fija **p** de ocurrencia del éxito entre los ensayos.

**p es constante entre los ensayos.**

Tiene dos parámetros **n** y **p**, se escribe: $\boldsymbol{b}\ (\boldsymbol{x};\ \boldsymbol{n};\ \boldsymbol{p})$ y $\boldsymbol{B}\ (\boldsymbol{x};\ \boldsymbol{n};\ \boldsymbol{p}) = probabilidad\ acumulada$

### 9.2.1 Condiciones

Se deben cumplir todas:

1. **Existe una serie de n ensayos.**
   Si fuera un solo ensayo no sería una distribución binomial
2. **En cada ensayo hay sólo dos posibles resultados**
   Si tengo mas de dos resultados no puedo aplicar la distribución binomial.
3. **En cada ensayo, los dos resultados posibles son mutuamente excluyentes**
   No hay ninguna posibilidad de que puedan pasar los dos eventos mutuamente. La intersección es 0.
4. **Los resultados de cada ensayo son independientes entre sí:**
   La probabilidad del primer ensayo no afecta la probabilidad del segundo ensayo.
5. **La probabilidad de cada resultado posible en cualquier ensayo es la misma de un ensayo a otro.**
   La probabilidad se mantiene constante.

### 9.2.2 Como se calcula

Se usa lo que se llama la **Función de Probabilidad**: $f_{(x)} = {}_nC_x * p^x * q^{n-x}$

- $p^x$ Probabilidad favorable elevado a la x
- $q^{n-x}$ Probabilidad desfavorable elevado a la n-x

También existe lo siguiente:

$Esperanza\ E_{(x)} = n * p$ *Valor más probable que saldrá del experimento si lo realizo una sola vez.*

$Varianza\quad V_{(x)} = n * p * q$ *Variabilidad que va a tener ese valor de esperanza.*

### 9.2.3 Ejemplo

Calcular la probabilidad de obtener 2 caras al lanzar una moneda 3 veces.

$$f_{(x)} = {}_nC_x * p^x * q^{n-x}$$
$$b_{(2;3;0;5)} = {}_3C_2 * p^2 * q^1$$
$$b_{(2;3;0;5)} = 3 * 0{,}5^2 * 0{,}5$$
$$b_{(2;3;0;5)} = 0{,}375$$

| $X_i$ | $P$ |
|---|---|
| 0 | 0,125 |
| 1 | 0,375 |
| 2 | 0,375 |
| 3 | 0,125 |
| | 1 |

## 9.3 DISTRIBUCIÓN HIPERGEOMÉTRICA

La distribución hipergeométrica es aplicable a muestreos sin reemplazo y la binomial a muestreos con reemplazo (la probabilidad no permanece constante).

Se denomina $\boldsymbol{h}\ (\boldsymbol{x}; \boldsymbol{n}; \boldsymbol{a}; \boldsymbol{N})$, y $\boldsymbol{H}\ (\boldsymbol{x}; \boldsymbol{n}; \boldsymbol{a}; \boldsymbol{N})$ para la acumulada.

$$\boldsymbol{h}\ (\boldsymbol{x}; \boldsymbol{n}; \boldsymbol{a}; \boldsymbol{N}) = \frac{{}_aC_x * {}_{N-a}C_{n-x}}{{}_NC_n} = \left[\frac{\boldsymbol{Total\ de\ favorables * Total\ de\ desfavorables}}{\boldsymbol{Total\ de\ grupos\ posibles}}\right]$$

- $\boldsymbol{N}$: Cantidad total de elementos con los que trabajo
- $\boldsymbol{a}$: Cantidad de elementos favorables que tengo en el total de la población.

$Esperanza\ E_{(x)} = n * P$ *Igual a la probabilidad de ocurrencia que tenga en el Primer elemento.*

- $P = \frac{\boldsymbol{a}}{\boldsymbol{N}}$

$Varianza \quad V_{(x)} = n * P * Q * \frac{N-n}{N-1}$

- $\frac{N-n}{N-1}$ Es un factor de corrección.

### 9.3.1 Ejemplo

Calcular la probabilidad obtener 3 pelotitas blancas al sacar 4 pelotas de una bolsa con 5 pelotitas blancas y 2 negras.

- **x** va a ser el valor que yo quiero calcular; la probabilidad de obtener **3** pelotitas blancas.
- **n** la cantidad de experimentos que yo realice; tamaño de la muestra; saco **4** pelotitas.
- **a** los elementos favorables que tengo en la población tengo **5** pelotitas blancas de 7 en total.
- **N** total de elementos que tengo en la población. Tengo **7** pelotitas.

$$h_{(3;4;5;7)} = \frac{{}_5C_3 * {}_2C_1}{{}_7C_4}$$

$$h_{(3;4;5;7)} = \frac{10 * 2}{35} = \mathbf{0,5714}$$

| $\boldsymbol{X_i}$ | $\boldsymbol{h}$ |
|---|---|
| 0 | - |
| 1 | - |
| 2 | 0,2857 |
| 3 | 0,5714 |
| 4 | 0,1429 |
| | 1,0000 |

*Al calcular las posibilidades de la hipergeométrica tengo que ver cuales son los valores de la variable posible. Debo plantear si lo que me pregunta es un resultado posible o si es 0.*

## 9.4 Distribución Poisson

Es una **distribución de probabilidad discreta** que expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo o a lo largo de un espacio de observación $\boldsymbol{P}\ (\boldsymbol{x}\ ;\ \boldsymbol{\lambda})$

- $\boldsymbol{x}$ Variable que tengo que calcular
- $\boldsymbol{\lambda}$ Promedio de ocurrencia *que tengo de este evento en este periodo de tiempo o espacio de observación.*

*Yo no mido los espacios de observación o periodos de tiempo, sino que cuento la cantidad de eventos que ocurren en estos.*

$$p_{(x;\,\lambda)} = \frac{e^{-\lambda} * \lambda^{x}}{x!}$$

- $Esperanza\ E_{(x)} = \lambda$
- $Varianza\quad V_{(x)} = \lambda$

### 9.4.1 Ejemplo

Si un banco recibe en promedio 6 cheques sin fondo por día, ¿cuáles son las probabilidades de que reciba 4 cheques sin fondo en un día dado?

$\boldsymbol{X} = \boldsymbol{4}, \boldsymbol{\lambda} = \boldsymbol{6}\ cheques\ sin\ fondo\ por\ día$

$$p_{(x=4;\,\lambda=6)} = \frac{e^{-6} * 6^{4}}{4!} = 0{,}13392$$

Si un banco recibe en promedio 6 cheques sin fondo por día, ¿cuáles son las probabilidades de que reciba 10 cheques sin fondos en **dos días** cualquiera consecutivos?

$\boldsymbol{X} = \boldsymbol{10}, \boldsymbol{\lambda} = \boldsymbol{12}\ cheques\ sin\ fondo\ por\ día$ SOLAMENTE PUEDO AJUSTAR EL $\boldsymbol{\lambda}$

$$p_{(x=10;\,\lambda=12)} = \frac{e^{-12} * 12^{10}}{10!} = 0{,}104953$$

## 9.5 La Distribución Multinomial

Se puede ver como una generalización del Binomial en el que, en lugar de tener dos posibles resultados, tenemos r resultados posibles. Supongamos que el resultado de una determinada experiencia puede ser n valores distintos: cada uno de ellos con probabilidad p1, p2, ..., pi, respectivamente.

## 9.6 La Distribución Geométrica

Esta distribución es un caso especial de la Binomial, ya que se desea que ocurra un éxito por primera y única vez en el último ensayo que se realiza del experimento, para obtener la fórmula de esta distribución, haremos uso de un ejemplo.

## 9.7 Distribución de Pascal

Supongamos que ahora se está interesado en conocer en que prueba x ocurre el k éxito en una sucesión de pruebas independientes de Bernoulli con probabilidad de éxito p.

## 9.8 TABLAS

### 9.8.1 Distribución binomial

$$b(x; n; p)$$

- $x$ valor de la variable que queríamos calcular.
- $n$ cantidad de experimentos o el tamaño de la muestra.
- $p$ probabilidad favorable de ocurrencia en cada uno de los experimentos.

**DISTRIBUCION BINOMIAL**

| n | p / x | 0,01 | 0,05 | 0,1 | 0,15 | 0,2 | 0,25 | 0,3 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0,9900 | 0,9500 | 0,9000 | 0,8500 | 0,8000 | 0,7500 | 0,7000 |
| 1 | 1 | 0,0100 | 0,0500 | 0,1000 | 0,1500 | 0,2000 | 0,2500 | 0,3000 |

| n | p / x | 0,01 | 0,05 | 0,1 | 0,15 | 0,2 | 0,25 | 0,3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 0 | 0,9801 | 0,9025 | 0,8100 | 0,7225 | 0,6400 | 0,5625 | 0,4900 |
| 2 | 1 | 0,0198 | 0,0950 | 0,1800 | 0,2550 | 0,3200 | 0,3750 | 0,4200 |
| 2 | 2 | 0,0001 | 0,0025 | 0,0100 | 0,0225 | 0,0400 | 0,0625 | 0,0900 |

### 9.8.2 Distribución binomial acumulada

*Solo cambia el titulo y las probabilidades. ¡LAS TABLAS SON DIFERENTES!*

| n | x | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0,9900 | 0,9500 | 0,9000 | 0,8500 | 0,8000 | 0,7500 | 0,7000 |
| 1 | 1 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 |

| n | p / x | 0,01 | 0,05 | 0,1 | 0,15 | 0,2 | 0,25 | 0,3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 0 | 0,9801 | 0,9025 | 0,8100 | 0,7225 | 0,6400 | 0,5625 | 0,4900 |
| 2 | 1 | 0,9999 | 0,9975 | 0,9900 | 0,9775 | 0,9600 | 0,9375 | 0,9100 |
| 2 | 2 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 | 1,0000 |

### 9.8.3 Distribución Poisson

$$p(x; \lambda)$$

- $x$ Variable que busco (columna izquierda)
- $\lambda$ Lambda (fila superior)

**DISTRIBUCION DE POISSON**

| X \ λ | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 |
|---|---|---|---|---|---|---|---|
| 0 | 0,904837 | 0,818731 | 0,740818 | 0,670320 | 0,606531 | 0,548812 | 0,496585 |
| 1 | 0,090484 | 0,163746 | 0,222245 | 0,268128 | 0,303265 | 0,329287 | 0,347610 |
| 2 | 0,004524 | 0,016375 | 0,033337 | 0,053626 | 0,075816 | 0,098786 | 0,121663 |
| 3 | 0,000151 | 0,001092 | 0,003334 | 0,007150 | 0,012636 | 0,019757 | 0,028388 |
| 4 | 0,000004 | 0,000055 | 0,000250 | 0,000715 | 0,001580 | 0,002964 | 0,004968 |
| 5 | 0,000000 | 0,000002 | 0,000015 | 0,000057 | 0,000158 | 0,000356 | 0,000696 |
| 6 | 0,000000 | 0,000000 | 0,000001 | 0,000004 | 0,000013 | 0,000036 | 0,000081 |
| 7 | 0,000000 | 0,000000 | 0,000000 | 0,000000 | 0,000001 | 0,000003 | 0,000008 |

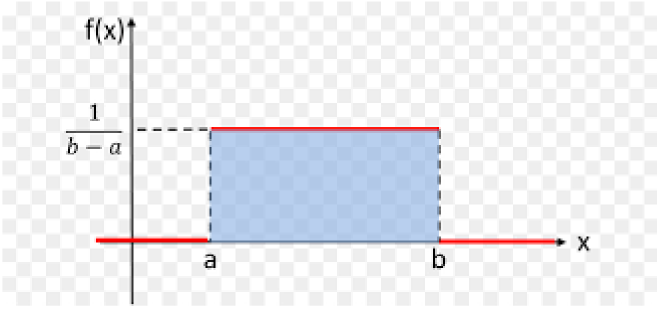
### 9.8.4 Distribución Poisson acumulada

Lo mismo, con otra tabla.

# 10 DISTRIBUCIONES DE PROBABILIDAD CONTINUAS

## 10.1 DISTRIBUCIÓN UNIFORME

Es la más simple de todas las distribuciones continuas de probabilidad. Se caracteriza por una función de densidad que es "plana", y por ello la probabilidad es uniforme en un intervalo cerrado.

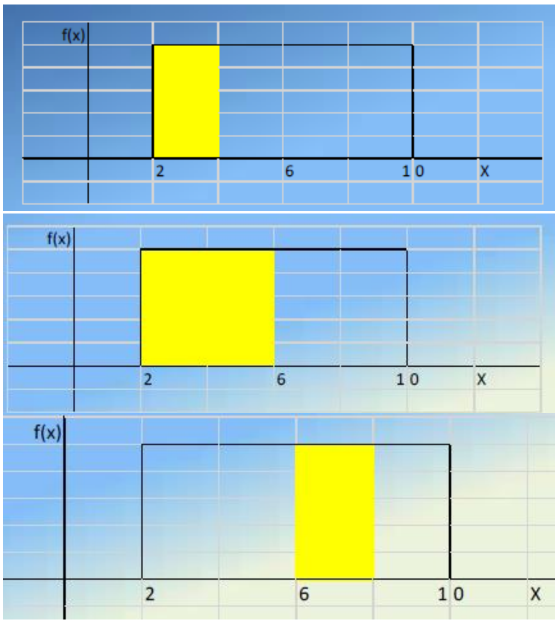



Esta no tiene mucha aplicación práctica.

En el intervalo **ab** hay una probabilidad de ocurrencia.

### 10.1.1 Ejemplo

Si tengo una probabilidad de ocurrencia uniforme entre 2 y 10:




Entre 2 y 4, la probabilidad es 0,25.

Entre 2 y 6, la probabilidad es de 0,5.

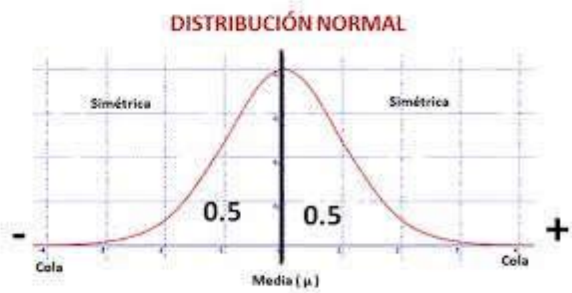
Entre 6 y 8, la probabilidad es de 0,25.

Entre 2 y 10, la probabilidad es 1.

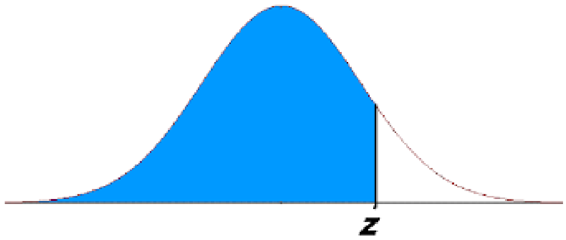
## 10.2 DISTRIBUCIÓN NORMAL

Describe aproximadamente muchos fenómenos que ocurren en la naturaleza, la industria y la investigación. La distribución normal presenta un valor de máxima frecuencia n, a partir del cual, decae hacia ambos lados con una **simetría perfecta**.

Esta simetría hace que a valores situados a igual distancia del valor modal por izquierda y por derecha de la distribución, les corresponda la misma probabilidad.



Se suele trabajar con la




## 10.3 DISTRIBUCIÓN NORMAL ESTÁNDAR

Su media es 0, y su desviación estándar es 1.

### 10.3.1 Estandarización de variables

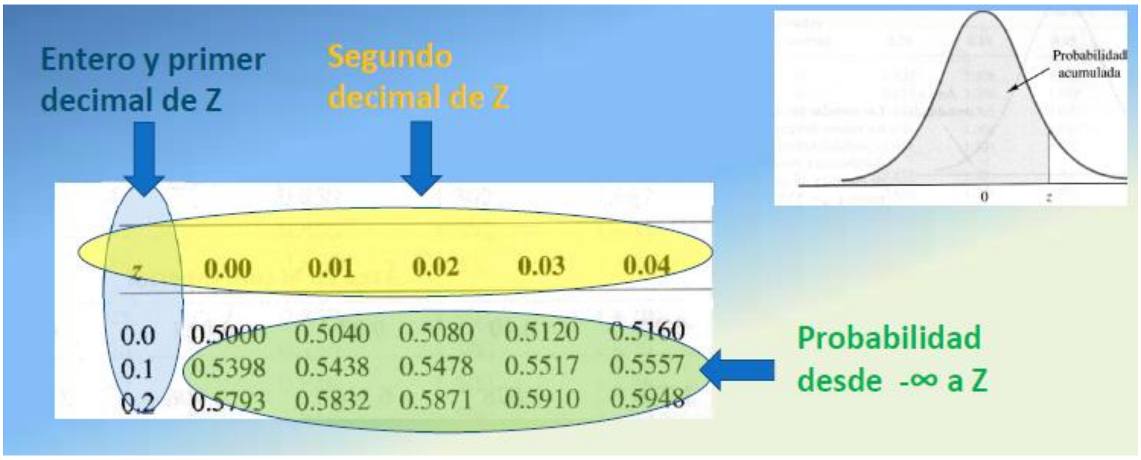
Para obtener una variable normal estandarizada:

$$Z_i = \frac{(X_I - \mu)}{\sigma}$$

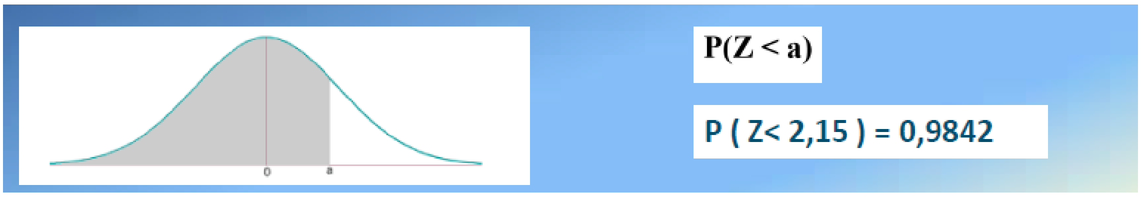
- $X$: valor del problema con el que estoy trabajando
- $\mu$: valor de la media de la distribución con la que estoy trabajando
- $\sigma$: Desviación estándar del problema con el que trabajo.

### 10.3.2 Búsqueda en tabla

Para obtener las probabilidades **desde menos infinito hasta Z**, usamos la tabla.

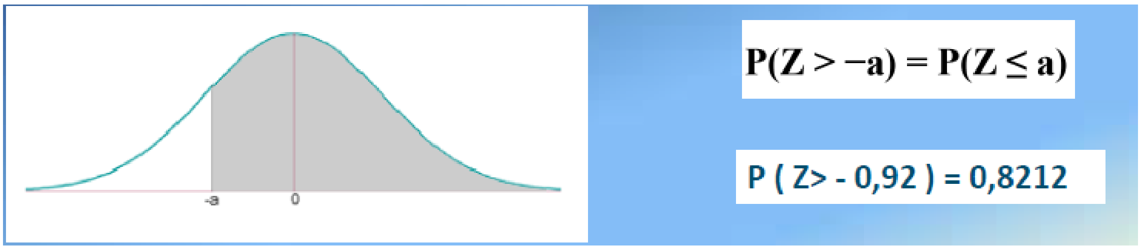



Si buscamos **desde donde z sea menor a 2,15**

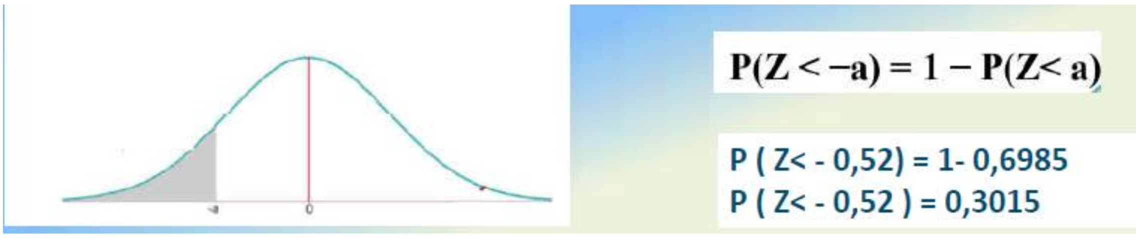



Pero, por otro lado, si busco la probabilidad **desde donde z sea mayor a 2,15**

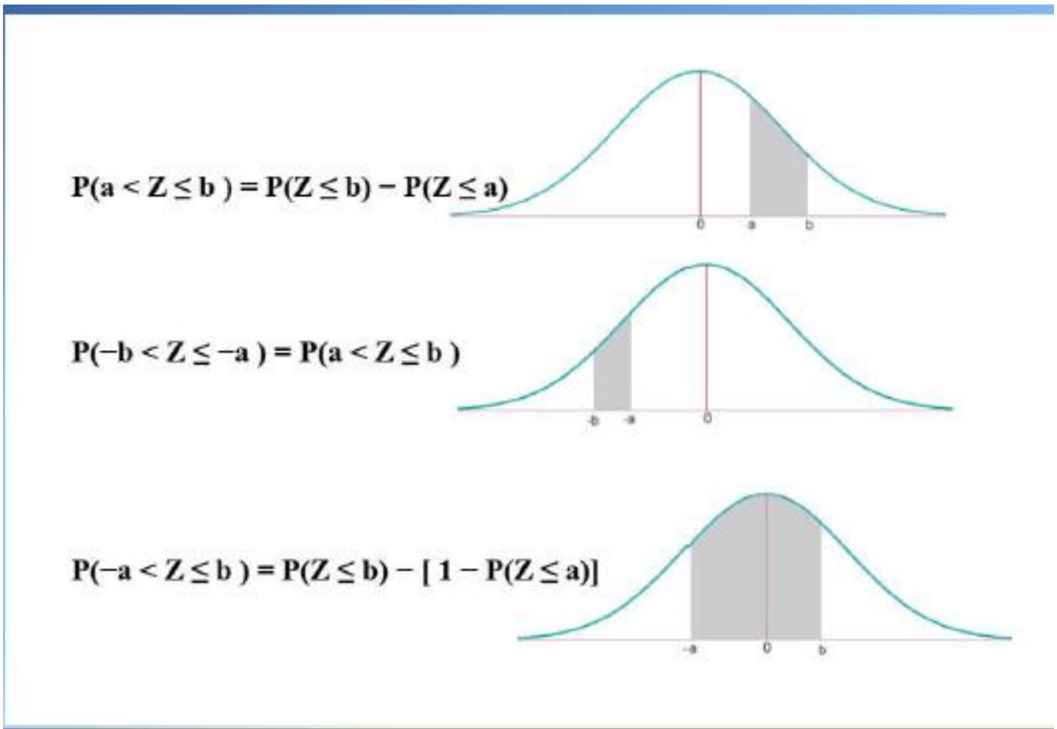
La parte a la derecha de un numero negativo, va a ser igual a la parte a la izquierda de un numero positivo.



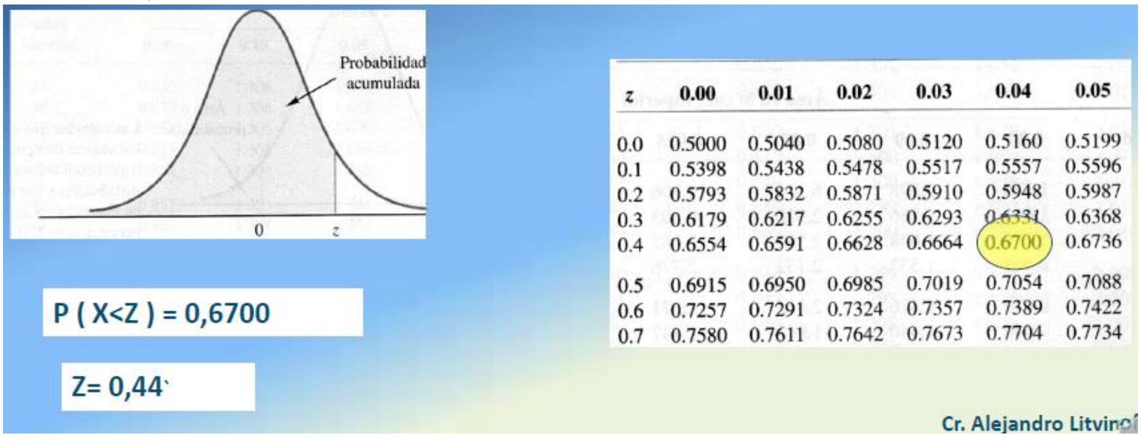

Y para un valor a la izquierda de un numero negativo, va a ser igual a la parte a la derecha de un numero positivo.




Y si quiero obtener la probabilidad entre dos números arbitrarios



### 10.3.3 Búsqueda inversa en tabla




| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 |

Si busco el Z que me de el valor de probabilidad de 0,67.

Arranco en la columna 0.00 hasta que me pase de ese valor. Es decir 1 fila antes de 0.6915

Luego, voy por la fila hasta llegar al valor 0.67

Si busco el Z que salga de una P=0,7, que no está en la tabla:


| | P | | Z |
|---|---|---|---|
| | 0,6985 | | 0,52 |
| | 0,7019 | | 0,53 |
| Diferencia | 0,0034 | | 0,01 |
| | 0,0015 | | 0,0044 |
| | 0,52+0,0044 | | 0,5244 |
| | 0,0019 | | 0,0056 |
| | 0,53-0,0056 | | 0,5244 |

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 |

Haremos una aproximación a ese valor. Usando las diferencias.
En la Z, siempre tengo una diferencia de 0,01.
A esta diferencia en Z, la diferencia en las probabilidades es de 0,0034 (P1 en Z=0.52, y P2 en Z=0.53)
Vamos a suponer que le voy a sumar 0,0015 al 0,6985 para obtener el 0,7.
Si sumo 0,0015, el Z para ello, lo saco con regla de 3. Z me queda 0,0044.
El valor de Z para 0,7 será 0,52 (correspondiente a P1) + (mas) 0,0044 (que salió de la regla de 3 simple….)

### 10.3.4 Otras aplicaciones

#### *10.3.4.1 Aproximación normal a la binomial*

$$Z_i = \frac{(X_I - \mu)}{\sigma} = \frac{x - n.p}{\sqrt{n.p.q}}$$

- $\mu$: Es reemplazado por la esperanza $(n.p)$
- $\sigma$: Es reemplazada por la raíz cuadrada de la varianza $(n.p.q)$

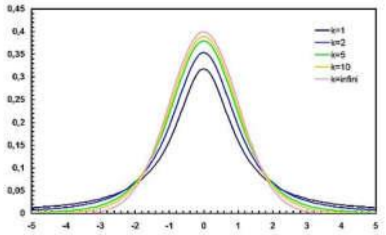
#### *10.3.4.2 Aproximación normal a Poisson*

$$Z_i = \frac{(X_I - \mu)}{\sigma} = \frac{x - \lambda}{\sqrt{\lambda}}$$

- $\mu$: Es reemplazado por la esperanza $(\lambda)$
- $\sigma$: Es reemplazada por la raíz cuadrada de la varianza $(\lambda)$

## 10.4 Distribución T de Student

Es una distribución de probabilidad que surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño.




Es **continua, simétrica**. Cambia para cada <u>grado de libertad</u> (las diferentes líneas), y a medida que aumentan los grados de libertad tiende a la distribución normal (Z).

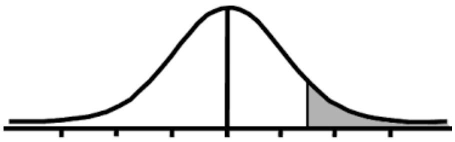
### 10.4.1 Grados de libertad (gl)

Los grados de libertad están dados por el número de valores que pueden ser asignados de forma arbitraria, antes de que el resto de las variables tomen un valor forzado para lograr un resultado estipulado previamente.

*En cualquier operación que yo tenga, son los valores que yo puedo dejar libres y de todas formas llegar al mismo resultado.*

El grado de libertad que debemos trabajar, deberá ser de $gl = n - 1$

| | α | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| v | 0,45 | 0,4 | 0,3 | 0,25 | 0,2 | 0,1 | 0,05 | 0,025 | 0,010 | 0,005 |
| 1 | 0,158 | 0,325 | 0,727 | 1,000 | 1,376 | 3,078 | 6,314 | 12,706 | 31,821 | 63,656 |
| 2 | 0,142 | 0,289 | 0,617 | 0,816 | 1,061 | 1,886 | 2,920 | 4,303 | 6,965 | 9,925 |
| 3 | 0,137 | 0,277 | 0,584 | 0,765 | 0,978 | 1,638 | 2,353 | 3,182 | 4,541 | 5,841 |
| 4 | 0,134 | 0,271 | 0,569 | 0,741 | 0,941 | 1,533 | 2,132 | 2,776 | 3,747 | 4,604 |
| 5 | 0,132 | 0,267 | 0,559 | 0,727 | 0,920 | 1,476 | 2,015 | 2,571 | 3,365 | 4,032 |
| 6 | 0,131 | 0,265 | 0,553 | 0,718 | 0,906 | 1,440 | 1,943 | 2,447 | 3,143 | 3,707 |

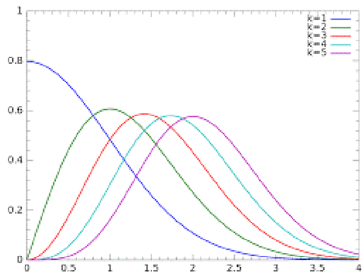


Al contrario de la tabla Z, esta tabla acumula desde $+\infty$ hasta $t$.

La probabilidad, esta arriba, y los t están dentro de la tabla. Exactamente al revés que la tabla de distribución normal (Z)

## 10.5 Distribución Chi cuadrado $\chi^2$

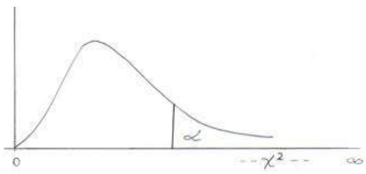
La distribución $\chi^2$ tiene muchas aplicaciones en inferencia estadística. La más conocida es la de la denominada prueba $\chi^2$ utilizada como prueba de independencia y como prueba de bondad de ajuste y en la estimación de varianzas.




Es **continua, asimétrica** (valores siempre positivos). Cambia para cada <u>grado de libertad</u> (las diferentes líneas).

El grado de libertad que debemos trabajar, deberá ser de $gl = n - 1$

| | α | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| v | 0,995 | 0,990 | 0,975 | 0,950 | 0,900 | 0,750 | 0,500 | 0,250 | 0,100 | 0,050 | 0,025 | 0,010 | 0,005 |
| 1 | 0,0000 | 0,0002 | 0,0010 | 0,0039 | 0,016 | 0,102 | 0,455 | 1,323 | 2,706 | 3,841 | 5,024 | 6,635 | 7,879 |
| 2 | 0,0100 | 0,0201 | 0,0506 | 0,1026 | 0,211 | 0,575 | 1,386 | 2,773 | 4,605 | 5,991 | 7,378 | 9,210 | 10,597 |
| 3 | 0,0717 | 0,1148 | 0,2158 | 0,3518 | 0,584 | 1,213 | 2,366 | 4,108 | 6,251 | 7,815 | 9,348 | 11,345 | 12,838 |
| 4 | 0,207 | 0,297 | 0,484 | 0,711 | 1,064 | 1,923 | 3,357 | 5,385 | 7,779 | 9,488 | 11,143 | 13,277 | 14,860 |
| 5 | 0,412 | 0,554 | 0,831 | 1,145 | 1,610 | 2,675 | 4,351 | 6,626 | 9,236 | 11,070 | 12,832 | 15,086 | 16,750 |




Se utiliza la tabla de la misma manera que en la distribución T, y acumula igual que la tabla desde $+\infty$ hasta $\chi^2$.

## 10.6 Distribución F

Usada en teoría de probabilidad y estadística, la distribución F es una distribución de probabilidad continua utilizada para análisis de Varianza. También se le conoce como distribución F de Snedecor (por George Snedecor) o como distribución F de Fisher-Snedecor.

## 10.7 Distribución exponencial

A pesar de que la distribución Normal puede utilizarse para resolver muchos problemas en ingeniería y ciencias, existen aún numerosas situaciones que requieren diferentes tipos de funciones de densidad, tales como la exponencial.

**Función de densidad:** $f(x) = \alpha * e^{-\alpha * x}$

**Función de distribución:** $F(x) = \int_0^X \alpha * e^{-\alpha * x} = [-e^{-\alpha * x}]_0^X = \mathbf{1 - e^{-\alpha * x}}$

# 11 Muestreo y selección de muestras

Muestreo: Herramienta de la investigación científica para obtener una muestra a partir de cierta población.

- Tamaño de muestra: **n**
- Tamaño de población: **N**

## 11.1 Tipos de muestreo

### 11.1.1 Muestreo no probabilístico

Solo para estudios exploratorios, cuyas conclusiones solo aplican a las muestras, no podemos representar una población.

### 11.1.2 Muestreo probabilístico

Muestreos aleatorios que permiten generalizaciones para toda la población

## 11.2 Métodos de selección de muestras

### 11.2.1 Muestreo irrestricto aleatorio, o aleatorio simple

Buscamos en una tabla de números aleatorios y elegimos los datos a muestrear.

### 11.2.2 Muestreo aleatorio sistemático

Es un tipo de muestreo donde se hace una selección aleatoria del primer elemento, pero los posteriores se selecciona utilizando intervalos fijos o sistemáticos hasta alcanzar el tamaño de la muestra deseado.

#### *11.2.2.1 Uso*

Calculamos el intervalo de muestreo (**k**) dividiendo el número de elementos en el marco de muestreo (**N**) por el tamaño de la muestra específica (**n**)

$$\boxed{k = \frac{N}{n}}$$

El numero aleatorio que utilizamos para elegir el primer elemento lo denominamos (**a**), y debe ser **aleatorio entre 1 y k.**

$$a = ran(1, k)$$

### 11.2.3 Muestreo por conglomerados

Dividimos a la población en conglomerados, donde son heterogéneos internamente, y homogéneos por fuera. Separamos en grupos que representan correctamente el total de la población.

### 11.2.4 Muestreo aleatorio estratificado

Separando los elementos de la población en grupos o estratos obtenemos nuestras muestras. Estos grupos son homogéneos por dentro, y heterogéneos por fuera.

Para representar cada estrato, usamos la afijación**.**

#### *11.2.4.1 Afijación igual*

$$\boxed{n_i = \frac{n}{r}} \left[\frac{tamaño\ de\ la\ muestra}{cantidad\ de\ estratos}\right]$$

#### *11.2.4.2 Afijación proporcional*

No lo dividimos según la cantidad de estratos, sino según la participación que tiene cada estrato en la población.

$$\boxed{n_i = \frac{N_i}{N} n} \left[\frac{tamaño\ del\ estrato}{tamaño\ de\ la\ poblacion} * tamaño\ de\ la\ muestra\right]$$

#### *11.2.4.3 Afijación optima*

No damos la proporcionalidad solo según el tamaño del estrato, sino por el tamaño del estrato y su desviación.

$$\boxed{n_i = \frac{N_i \sigma}{\sum N_i \sigma} n}$$

#### *11.2.4.4 Uso*

| Muestra | Afijación Igual | |
|---|---|---|
| $n_1$ | 600/3 | 200 |
| $n_2$ | 600/3 | 200 |
| $n_3$ | 600/3 | 200 |
| $n$ | | 600 |

| | |
|---|---|
| $n$ | 600 |
| $N_1$ | 2000 |
| $N_2$ | 3000 |
| $N_3$ | 1000 |

| Muestra | Afijación Proporcional | |
|---|---|---|
| $n_1$ | (2k/6k)*600 | 200 |
| $n_2$ | (3k/6k)*600 | 300 |
| $n_3$ | (1k/6k)*600 | 100 |
| $n$ | | 600 |

Y para hacer la Afijación Optima, necesito la desviación estándar de cada uno de los estratos.

| | |
|---|---|
| $\sigma_1$ | 2 |
| $\sigma_2$ | 3 |
| $\sigma_3$ | 5 |

Y armo el siguiente cuadro:

| $\sigma.N_i$ | |
|---|---|
| $\sigma_1.N_1$ | 4000 |
| $\sigma_2.N_2$ | 9000 |
| $\sigma_3.N_3$ | 5000 |
| $\sum$ | 18000 |

| Muestra | Afijación Proporcional | |
|---|---|---|
| $n_1$ | (4k/18k)*600 | 133 |
| $n_2$ | (9k/18k)*600 | 300 |
| $n_3$ | (5k/18k)*600 | 167 |
| $n$ | | 600 |

# 12 DISTRIBUCIONES DE MUESTREO

Una distribución de muestreo es la distribución de probabilidad de un estadístico dado, como la media.

## 12.1 DISTRIBUCIÓN DE LA MEDIA MUESTRAL

Distribución de probabilidades de todas las medias de las muestras de un determinado tamaño de una población.

Tomamos 4 elementos:

| Elem | $X_i$ |
|---|---|
| A | 3 |
| B | 3 |
| C | 4 |
| D | 6 |

Muestrearemos con 2 elementos, sin repetición, donde, tomando el valor de cada elemento, calculamos la media para esa muestra:

| Muestra | | Media |
|---|---|---|
| A | B | 3.0 |
| A | C | 3.5 |
| A | D | 4.5 |
| B | C | 3.5 |
| B | D | 4.5 |
| C | D | 5.0 |

Para calcular la distribución de media muestral, buscamos los diferentes tamaños de media que podemos tener, y ver la probabilidad de estos.

| Media ($X_i$) | $f_i$ | $p$ |
|---|---|---|
| 3.0 | 1 | 0.1667 |
| 3.5 | 2 | 0.3333 |
| 4.5 | 2 | 0.3333 |
| 5.0 | 1 | 0.1667 |

Partiendo de esto, se elabora el:

## 12.2 TEOREMA CENTRAL DEL LÍMITE

Dice: *La distribución de la media de una muestra aleatoria proveniente de una población que puede no ser normal, sigue aproximadamente una distribución normal, siempre que el tamaño de la muestra sea lo suficientemente grande (>30)*

Esto es importante, ya que muchas veces trabajo con poblaciones de las cuales no conozco su distribución, pero necesito estimar y tomar decisiones sobre esta, con este teorema, puedo usar la distribución normal para esos cálculos siempre y cuando la muestra sea lo suficientemente grande.

También este teorema nos dice lo siguiente:

1. $\bar{\mu}_{(\bar{X})} = \mu$
   La media de la distribución de la media muestral será igual a la media de la población.
2. $\sigma_{(\bar{X})} = \sigma / \sqrt{n}$
   La desviación estándar de la distribución de la media muestral, será igual a la desviación estándar de la población, dividido la raíz cuadrada de **n**

## 12.3 Distribución de la proporción muestral

### 12.3.1 Proporción *(p, P)*

Es el número de observaciones con una característica en particular entre la población o muestra de referencia. El numerador siempre esta incluido en el denominador.

***Ejemplo:*** *si hay 50 alumnos y 20 son mujeres, decimos que la proporción de mujeres es igual al 40%.*

Su calculo es similar a la distribución de la media muestral, pero calculando la proporción (p=x/n) de los elementos que **yo busco dentro de la población**.

Tomamos 4 elementos, **[A, B, C, D]**, de los cuales, me son favorables los elementos **[A, B]**

Muestrearemos con 2 elementos, sin repetición, donde, calculamos la proporción para esa muestra:

| **Muestra** | | **Proporción** |
|---|---|---|
| A | B | 1.0 |
| A | C | 0.5 |
| A | D | 0.5 |
| B | C | 0.5 |
| B | D | 0.5 |
| C | D | 0.0 |

Calculamos de misma manera:

| **Proporción ($p_i$)** | $f_i$ | $p$ |
|---|---|---|
| 0.0 | 1 | 0.1667 |
| 0.5 | 4 | 0.6667 |
| 1.0 | 1 | 0.1667 |

# 13 ESTIMACIÓN DE PARÁMETROS

Usando un estimador $\hat{\theta}$, que calculamos a partir de datos muestrales, obtenemos conclusiones, y nos proporciona información sobre el valor del parámetro $\theta$.

## 13.1 PROPIEDADES DE LOS ESTIMADORES

### 13.1.1 Insesgabilidad

Un estimador es insesgado cuando la esperanza matemática de su distribución en el muestreo coincide con el valor del parámetro. $E(\hat{\theta}) = \theta$

*Para estimar la varianza poblacional, usamos la varianza insesgada, calculada con $n - 1$ en el denominador.*

### 13.1.2 Eficiencia

Un estimador es eficiente u optimo cuando su **varianza es mínima**.

*El estimador mas eficiente, va a ser el que tenga la menor dispersión.*

### 13.1.3 Consistencia

El estimador $\hat{\theta}$ se acerca al parámetro a medida que aumenta el tamaño de la muestra. Si $n \to \infty$, el estimador es **insesgado** y de **varianza 0**.

### 13.1.4 Suficiencia

Un estimador es suficiente cuando incluye **toda la información relevante** de la muestra, de forma que ningún otro estimador puede considerar información adicional.

*Si tenemos dos estimadores sobre un mismo parámetro, no serán suficientes, debido a que necesitan información adicional.*

*Si tenemos un estimador que me da toda la información del parámetro que nos es relevante, lo podemos considerar un estimador **suficiente.***

### 13.1.5 Invariabilidad

Un estimador es **invariable** cuando si transformamos el parámetro a estimar mediante una función $f(\theta)$, dicha función puede ser estimada por la función del estimador: $f(\hat{\theta})$.

### 13.1.6 Robustez

Un estimador es **robusto** cuando si se vulnera alguno de los supuestos en los que se basa el proceso de estimación, la estimación no cambia significativamente y sigue ofreciendo resultados fiables.

*Si es muy sensible el resultado a las variaciones que pueda tener, no será robusto.*

## 13.2 ESTIMACIÓN PUNTUAL

Es el primer tipo de estimación utilizada para estimar los parámetros de la población:

| | | |
|---|---|---|
| $\mu \sim \bar{X}$<br>*Media de la muestra es un estimador puntual de la media de la población.* | $Me \sim \tilde{x}$<br>*Mediana de la muestra es un estimador puntual de la mediana de la población* | $Me \sim \hat{x}$<br>*Moda de la muestra es un estimador puntual de la moda de la población* |
| $\sigma^2 \sim s^2$<br>*Varianza de la muestra es un estimador puntual de la varianza de la población.* | $\sigma \sim s$<br>*Desviación estándar de la muestra es un estimador puntual de la desviación estándar de la población.* | $p \sim P$<br>*proporción poblacional de algún parametro de la muestra es un estimador puntual de la proporción de ese parámetro de la población.* |

# 13.3 ESTIMACIÓN POR INTERVALOS

*No solo nos dice el valor estimado, sino la probabilidad de ocurrencia de que ese valor se de en la realidad (que tan confiables son los estimadores)*

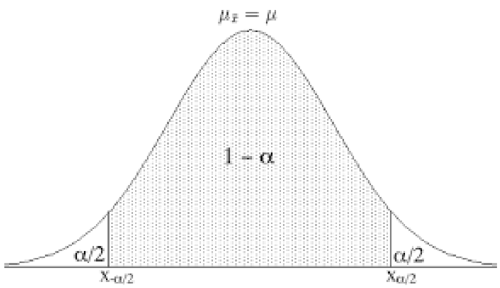
Es un rango de números llamado intervalo, construido alrededor de la estimación puntual. La estimación por intervalos consiste en establecer el intervalo de valores donde es más probable que se encuentre el parámetro.

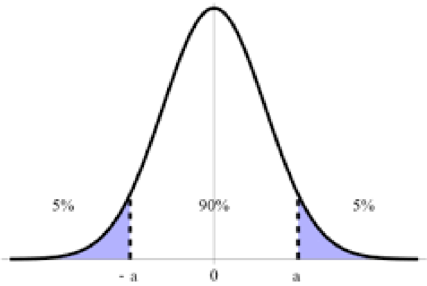
A este rango de números, le aplicamos una **confianza**, la cual es la probabilidad de que los números ciertamente se encuentren dentro de esos valores.

## 13.3.1 Intervalo de confianza

Se construye de manera que la probabilidad del parámetro de la población se localice en algún lugar dentro del intervalo conocido.

$$\boxed{confianza = 1 - \alpha}$$




## 13.3.2 Estimación de la media poblacional

### *13.3.2.1 Caso 1 (σ conocida)*

Estimación de la media poblacional con varianza conocida.

$$\underbrace{\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}_{Extremo\ Izquierdo} \leq \mu \leq \underbrace{\bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}_{Extremo\ Derecho} \quad o\ bien \quad \boxed{\mu \to \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}$$

- $\bar{X}$ media de la muestra.
- $\mu$ media poblacional, la cual estamos acotando / estimando.
- $Z$ sale de tabla
- $Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ sería el error de estimación ($e$)

Si nos dan el error de estimación como dato, puedo buscar el valor del tamaño de la muestra ($n$):

$$\boxed{n = Z_{\frac{\alpha}{2}}^{2} \frac{\sigma^2}{e^2}}$$

Ejemplo:

| |
|---|
| $\sigma = 2$ |
| $n = 36$ |
| $\bar{X} = 10$ |
| $\alpha = 0{,}05$ |
| $1 - \alpha = 0{,}95$ |

Debemos sacar el valor de $Z_{\frac{\alpha}{2}}$. Sabiendo que a la derecha y a la izquierda tendremos $\frac{\alpha}{2}$, visto en el gráfico de arriba a la izquierda, el valor que debemos buscar, sabiendo que la tabla acumula desde $-\infty$ hasta $Z$, será de $1 - \frac{\alpha}{2}$.

Este valor (0,9750 en este caso), lo buscamos con la **tabla de distribución normal estándar.** Teniendo el valor de Z, es simplemente calcular:

$$\mu \to \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \to 10 \pm 1{,}96\ \frac{2}{\sqrt{36}} \to 10 \pm 0{,}6533$$

#### *13.3.2.2 Caso 2 ($\sigma$ desconocida, muestra $\geq 30$)*

Cuando tengo la varianza desconocida, y una muestra grande

$$\underbrace{\bar{X} - Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}}_{Extremo\ Izquierdo} \leq \mu \leq \underbrace{\bar{X} + Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}}_{Extremo\ Derecho} \quad o\ bien \quad \boxed{\mu \to \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}}$$

La formula es exactamente igual, pero como no tengo sigma ($\sigma$), uso la desviación estándar ($s$)

$$\boxed{n = Z_{\frac{\alpha}{2}}^{\ 2} \frac{s^2}{e^2}}$$

Ejemplo:

| |
|---|
| $s = 3$ |
| $n = 36$ |
| $\bar{X} = 10$ |
| $\alpha = 0{,}05$ |
| $1 - \alpha = 0{,}95$ |

$$\mu \to \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \to 10 \pm 1{,}96 \frac{3}{\sqrt{36}} \to 10 \pm 0{,}98$$

#### *13.3.2.3 Caso 3 ($\sigma$ desconocida, muestra $< 30$)*

En este caso, no se utiliza la distribución normal, sino la **Distribución t de Student**.

$$\underbrace{\bar{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}}_{Extremo\ Izquierdo} \leq \mu \leq \underbrace{\bar{X} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}}_{Extremo\ Derecho} \quad o\ bien \quad \boxed{\mu \to \bar{X} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}}$$

Las fórmulas no cambian, solo la distribución utilizada.

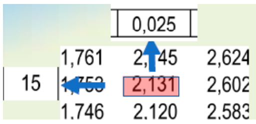
$$\boxed{n = t_{\frac{\alpha}{2}}^{\ 2} \frac{s^2}{e^2}}$$

Ejemplo:

| |
|---|
| $s = 3$ |
| $n = 16$ |
| $\bar{X} = 10$ |
| $\alpha = 0{,}05$ |
| $1 - \alpha = 0{,}95$ |

Primero debemos buscar $t_{\frac{\alpha}{2}}$.

$$\frac{\alpha}{2} = 0{,}025\ ; \quad gl = n - 1 = 15 \ \to \ t_{\frac{\alpha}{2}} = 2{,}131$$




Calculando:

$$\mu \to \bar{X} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \to 10 \pm 2{,}131 \frac{3}{\sqrt{16}} \to 10 \pm 1{,}5983$$

### 13.3.3 Estimación de la proporción poblacional

Las formulas son iguales que en el caso 1, pero con un ligero cambio:

$$\underbrace{p - Z_{\frac{\alpha}{2}} \sqrt{\frac{p\ .\ q}{n}}}_{Extremo\ Izquierdo} \leq P \leq \underbrace{p + Z_{\frac{\alpha}{2}} \sqrt{\frac{p\ .\ q}{n}}}_{Extremo\ Derecho} \quad o\ bien \quad \boxed{P \to p \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p\ .\ q}{n}}}$$

$$\boxed{n = Z_{\frac{\alpha}{2}}^{\ 2} \frac{p\ .\ q}{e^2}}$$

### 13.3.4 Estimación de la varianza poblacional

Acá si cambia el estimador, y la distribución que utilizamos es **Chi Cuadrado ($\chi^2$)**

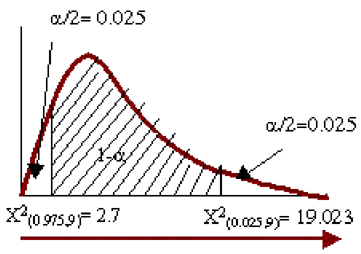
$$\underbrace{\frac{(n-1)\,.\,s^2}{{\chi^2}_{\left(\frac{\alpha}{2}\right)}}}_{Extremo\ Izquierdo} < \sigma^2 < \underbrace{\frac{(n-1)\,.\,s^2}{{\chi^2}_{\left(1-\frac{\alpha}{2}\right)}}}_{Extremo\ Derecho}$$

Ejemplo:

| |
|---|
| $s = 3$ |
| $n = 10$ |
| $\alpha = 0{,}05$ |
| $\alpha/2 = 0{,}025$ |

1. Busco los valores en la tabla.

$${\chi^2}_{\left(\frac{\alpha}{2}\right)} = 19{,}023 \quad ; \quad {\chi^2}_{\left(1-\frac{\alpha}{2}\right)} = 2{,}7$$




2. Calculo

$$\frac{(n-1)\,.\,s^2}{{\chi^2}_{\left(\frac{\alpha}{2}\right)}} < \sigma^2 < \frac{(n-1)\,.\,s^2}{{\chi^2}_{\left(1-\frac{\alpha}{2}\right)}}$$

$$\frac{(10-1)\,.\,3^2}{19{,}023} < \sigma^2 < \frac{9\,.\,9}{2{,}7}$$

$$4{,}258 < \sigma^2 < 30$$

### 13.3.5 Estimación de la desviación estándar poblacional

También uso la distribución **Chi Cuadrado ($\chi^2$)**

**Como:** $\boldsymbol{Desviación\ Estandar\ (\sigma) = \sqrt{(\sigma^2)}\ \ (raiz\ cuadrada\ de\ la\ varianza)}$

$$\underbrace{\sqrt{\frac{(n-1)\,.\,s^2}{{\chi^2}_{\left(\frac{\alpha}{2}\right)}}}}_{Extremo\ Izquierdo} < \sigma < \underbrace{\sqrt{\frac{(n-1)\,.\,s^2}{{\chi^2}_{\left(1-\frac{\alpha}{2}\right)}}}}_{Extremo\ Derecho}$$

# 14 PRUEBAS DE HIPÓTESIS

Son un procedimiento estadístico que nos permite aceptar o rechazar una afirmación hecha con respecto a un fenómeno o suceso.

Son lo opuesto a la estimación por intervalos.

En la estimación por intervalos, partimos de los datos de una muestra, y estimamos los valores de la población.

En la prueba de hipótesis, tomamos la decisión de aceptar o rechazar una hipótesis nula cuantificando la probabilidad de cometer un error al tomar esta decisión. **Tenemos los datos de la población, tan solo probamos si esa muestra es o no de la población.**

### 14.1.1 Hipótesis estadísticas

- **$H_0$: Hipótesis nula**
- **$H_1$: Hipótesis alternativa**
  Es contraria a la hipótesis nula.

### 14.1.2 Etapas

1. **Formular** la hipótesis nula y alternativa
2. **Especificar** el nivel de significación ($\alpha$)
   *Es lo contrario del nivel de confianza. Es la probabilidad de cometer el Error tipo I*
3. **Obtener** el resto de los datos y con eso determinar el estadístico
   *Importante, ya que, si tenemos los datos y el estadístico, solo reemplazamos el estadístico y resolvemos.*
4. **Establecer** los valores críticos que dividen las regiones de rechazo y de no rechazo
5. **Calcular** el valor del estadístico apropiado y determinar si el estadístico ha caído en la región de rechazo o en la región de no rechazo y tomar la decisión estadística
6. **Expresar** la decisión estadística en términos del problema

## 14.2 DECISIÓN ESTADÍSTICA

*Es la acción de aceptar o rechazar la prueba de hipótesis.*

| | $H_0 = TRUE$ | $H_0 = FALSE$ |
|---|---|---|
| **Acepto $H_0$** | Decisión Correcta | Error tipo II |
| **Rechazo $H_0$** | Error tipo I | Decisión Correcta |

### 14.2.1 Alfa

*$\alpha$ Es la probabilidad de cometer el error tipo I.*

$$\alpha = P(Error\ tipo\ I) = P\left(\frac{Rechazar\ H_0}{H_0\ es\ verdadera}\right)$$

En otras palabras, es la probabilidad de Rechazar una Hipótesis Nula ($H_0$) que en realidad es verdadera.

### 14.2.2 Beta

*$\beta$ Es la probabilidad de cometer el error tipo II.*

$$\beta = P(Error\ tipo\ II) = P\left(\frac{Aceptar\ H_0}{H_0\ es\ falsa}\right)$$

En otras palabras, es la probabilidad de Aceptar una Hipótesis Nula ($H_0$) que en realidad es falsa.

### 14.2.3 Potencia de una prueba

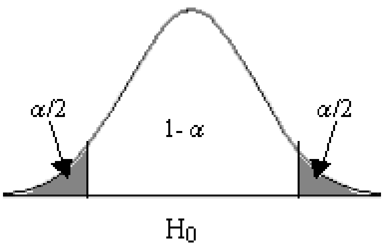
Es la **probabilidad de <u>*no*</u> *cometer* el error tipo II.**

$$Potencia\ de\ una\ prueba = 1 - \beta$$

## 14.3 PRUEBAS LATERALES Y BILATERALES

Cuando me interesa que el valor que tengo que probar sea exacto, uso

### 14.3.1 Prueba bilateral




Una prueba es bilateral (dos zonas de rechazo, una a la izquierda y la otra a la derecha), cuando la hipótesis nula, indica un valor específico del parámetro una serie de valores igual al valor del parámetro propuesto.

$$H_0 = \theta = \theta_0 \quad H_1 = \theta \neq \theta_0$$

*Planteo que la $H_0$ es igual a un valor, y la $H_1$ es distinta a ese valor.*

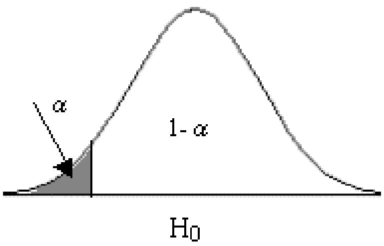
Como el $\alpha$ (significación) está dividido en 2, mitad a la izquierda y a la derecha,

### 14.3.2 Prueba lateral izquierda

Una prueba es lateral izquierda, (zona de rechazo a la izquierda) cuando la hipótesis nula, plantea un valor del parámetro o una serie de valores mayores al valor del parámetro propuesto.

$$H_0 = \theta \geq \theta_0 \quad H_1 = \theta < \theta_0$$

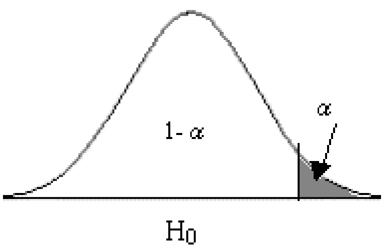
*Planteo que la $H_0$ es mayor o igual a un valor, y la $H_1$ es menor a ese valor.*




*La significación (zona de rechazo) de esta prueba se encuentra a la izquierda de la campana.*

*Cuando busco el valor de $\alpha$ en la tabla, no debo dividirlo por 2.*

### 14.3.3 Prueba lateral derecha




Una prueba es lateral derecha (zona de rechazo a la derecha), cuando la hipótesis nula, plantea un valor del parámetro o una serie de valores menores al valor del parámetro propuesto.

$$H_0 = \theta \leq \theta_0 \quad H_1 = \theta > \theta_0$$

*Planteo que la $H_0$ es menor o igual a un valor, y la $H_1$ es mayor a ese valor.*

## 14.4 APLICADO A LA MEDIA POBLACIONAL

### 14.4.1 Caso 1: ($\boldsymbol{\sigma}$ conocida)

- $Z$: estadístico a utilizar
- $\bar{X}$: Media muestral
- $\mu$: Media poblacional
- $\sigma$: Desviación estándar poblacional
- $n$: Cantidad de datos de la muestra

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

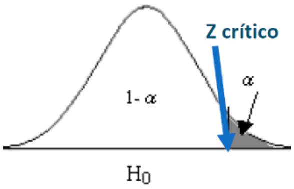
#### *14.4.1.1 Ejemplo*

Un camionero transporta bolsas de Harina de promedio de 50 Kg con una varianza de 4 Kg ². Para no pasarse del peso necesita probar con una significación del 5% que las bolsas no superan ese peso

Para eso toma una muestra de 25 bolsas y le da una media de 51 kg con una desviación estándar de 3 Kg.

##### 14.4.1.1.1 Formular la hipótesis

Lo que nos interesa acá es las bolsas **no se pasen en peso**. Utilizaremos una **Prueba lateral derecha**. Mi valor crítico esta a la derecha. Para este lado rechazo.




$$H_0 = \mu \leq 50kg \qquad H_1 = \mu > 50kg$$

##### 14.4.1.1.2 Especifico $\alpha$

El mismo ejercicio nos indica: $\alpha = 5\% = 0{,}05$

##### 14.4.1.1.3 Determinar el estadístico

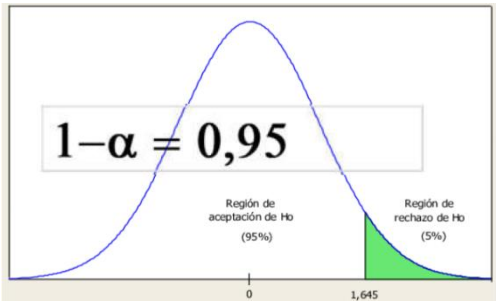
$$\bar{X} = 51kg \;\; ; \;\; \mu = 50kg \;\; ; \;\; \sigma^2 = 4kg^2 \;\rightarrow\; \sigma = 2kg \; ; \;\; n = 25 \;\; ; \;\; s = 3kg \;\; ; \;\; \alpha = 0{,}05$$

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{51kg - 50kg}{\frac{2kg}{\sqrt{25}}} = 2{,}5$$

##### 14.4.1.1.4 Establezco valores críticos

Hago una búsqueda inversa en la tabla para el valor de $1 - \alpha = 0.95$.

Si no lo encuentro en la tabla Z, lo busco en la tabla t, con infinitos grados de libertad. Si uso esta tabla, busco para el valor de $\alpha = 0{,}05$, ya que acumula al revés.




$$\boldsymbol{Z(1-\alpha) = t(\infty, \alpha) = 1{,}645}$$

Entonces podemos plantear las **regiones de rechazo:**

$$Si\ Z_c \leq 1{,}645\ \boldsymbol{Acepto}\ H_0$$

$$Si\ Z_c > 1{,}645\ \boldsymbol{Rechazo}\ H_0$$

##### 14.4.1.1.5 Tomo la decisión estadística

En el paso 3, habíamos calculado $Z = 2{,}5$.

Como $Z > Z_c$ rechazo la hipótesis.

##### 14.4.1.1.6 Expreso la decisión estadística

"Rechazamos con una **significación** del **5%** que las bolsas de harina no superan un **promedio de 50kg**"

#### 14.4.2 Caso 2: ($\sigma$ desconocida, muestra $\geq 30$)

Solo cambio la fórmula:

- $Z$: estadístico a utilizar
- $\bar{X}$: Media muestral
- $\mu$: Media poblacional
- $s$: Desviación estándar muestral
- $n$: Cantidad de datos de la muestra

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

#### 14.4.3 Caso 3: ($\sigma$ desconocida, muestra $< 30$)

Cambio el estadístico a utilizar.

- $t$: Estadístico a utilizar
- $\bar{X}$: Media muestral
- $\mu$: Media poblacional
- $s$: Desviación estándar muestral
- $n$: Cantidad de datos de la muestra

$$\boxed{t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}}$$

## 14.5 Aplicado a la proporción poblacional

- $Z$: Estadístico a utilizar
- $p$: Proporción muestral favorable
- $P$: Proporción poblacional favorable
- $Q$: Proporción poblacional **desfavorable**
- $n$: Cantidad de datos de la muestra

$$\boxed{Z = \frac{p - P}{\sqrt{\frac{P.Q}{n}}}}$$

## 14.6 Aplicado a la varianza y desviación estándar poblacional

- $\chi^2$: Estadístico a utilizar
- $\sigma^2$: Varianza poblacional
- $s^2$: Varianza muestral
- $n$: Cantidad de datos de la muestra

$$\boxed{\chi^2 = \frac{(n-1).s^2}{\sigma^2}}$$

**Siempre trabajo con los valores de la varianza, y si tengo que probar una desviación estándar, solo elevo la desviación estándar muestral y poblacional al cuadrado y uso el mismo estadístico.**

#### 14.6.1 Ejemplo

Un camionero transporta bolsas de Harina de promedio de 50 Kg. Para no tener problemas con la **variabilidad** del peso necesita probar con una significación del 5% que las bolsas **tengan** una varianza de 4,5 Kg²

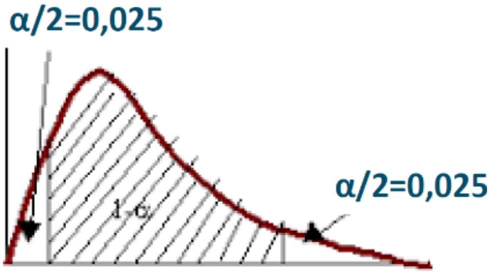
Para eso toma una muestra de 25 bolsas y le da una media de 51 kg con una desviación estándar de 2 Kg.

##### *14.6.1.1 Formular la hipótesis*

Al camionero le interesa probar que las bolsas **tengan (=)** una varianza de 4,5 Kg²

$$H_0 = \sigma^2 = 4{,}5kg^2 \qquad H_1 = \sigma^2 \neq 4{,}5kg^2$$

Esta es una prueba BILATERAL.




##### *14.6.1.2 Especifico $\alpha$*

El mismo ejercicio nos indica: $\alpha = 5\% = 0{,}05$

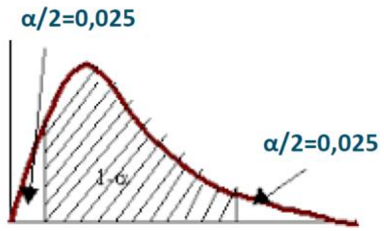
#### *14.6.1.3 Determinar el estadístico*

$$\bar{X} = 51kg \ ; \ \mu = 50kg \ ; \ \sigma^2 = 4,5kg^2 \ ; \ n = 25 \ ; \ s = 2kg \to s^2 = 4kg^2 \ ; \ \alpha = 0{,}05$$

$$\chi^2 = \frac{(n-1).s^2}{\sigma^2} = \frac{(25-1).4kg^2}{4{,}5\ kg^2} = 21{,}3333$$

#### *14.6.1.4 Establezco valores críticos*

Como es **bilateral**, establezco dos regiones de rechazo, una a la izquierda y otra a la derecha, cuya probabilidad será de $\alpha/2$




Para el valor de la derecha, solo busco en la tabla el valor de $\alpha/2$:

$$\chi^2(24, \alpha/2) = \chi^2(24, 0{,}025) = 39{,}364$$

Para el valor de la izquierda, busco para $1-\alpha/2$:

$$\chi^2(24{,}1-\alpha/2) = \chi^2(24, 0{,}975) = 12{,}401$$

Entonces nos quedan los intervalos.

$$Si\ 12{,}401 \leq \chi^2 \leq 39{,}364\ \boldsymbol{Acepto}\ H_0$$

$$Si\ \chi^2 > 39{,}364\ \boldsymbol{Rechazo}\ H_0$$

$$Si\ \chi^2 < 12{,}401\ \boldsymbol{Rechazo}\ H_0$$

#### *14.6.1.5 Tomo la decisión estadística*

En el paso 3, habíamos calculado $\chi^2 = 21{,}3333$

Como $12{,}401 \leq \chi^2 \leq 39{,}364$ Acepto la hipótesis.

#### *14.6.1.6 Expreso la decisión estadística*

"Aceptamos con una **significación** del **5%** que las bolsas de harina pueden tener una **varianza de** $4{,}5kg^2$"

## 14.7 APLICADO A LA DIFERENCIA DE MEDIAS

*En este caso, tengo en la hipótesis nula, que una media poblacional, menos otra media poblacional, me puede dar igual, mayor o menor a tal valor. Se puede plantear de dos formas.*

| | | |
|---|---|---|
| $H_0 \to \mu_1 - \mu_2 = a$<br>$H_1 \to \mu_1 - \mu_2 \neq a$ | $H_0 \to \mu_1 - \mu_2 \leq a$<br>$H_1 \to \mu_1 - \mu_2 > a$ | $H_0 \to \mu_1 - \mu_2 \geq a$<br>$H_1 \to \mu_1 - \mu_2 < a$ |
| $\begin{cases} H_0:\ \mu_1 = \mu_2 \\ H_1:\ \mu_1 \neq \mu_2 \end{cases}$ | $\begin{cases} H_0:\ \mu_1 \leq \mu_2 \\ H_1:\ \mu_1 > \mu_2 \end{cases}$ | $\begin{cases} H_0:\ \mu_1 \geq \mu_2 \\ H_1:\ \mu_1 < \mu_2 \end{cases}$ |

### 14.7.1 Caso 1: ($\sigma$ conocida)

- $Z$: estadístico a utilizar
- $\bar{X}$: Media muestral
- $\mu$: Media poblacional
- $\sigma^2$: Varianza poblacional
- $n$: Cantidad de datos de la muestra
- *Subindice 1: Dato perteneciente a población 1*
- *Subindice 2: Dato perteneciente a población 2*

$$\boxed{Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

**Para los siguientes dos casos, determinamos el tamaño de la muestra sumando $n_1 + n_2$**

### 14.7.2 Caso 2: ($\boldsymbol{\sigma}$ desconocida, muestra $\geq \mathbf{30}$)

- $Z$: estadístico a utilizar
- $\bar{X}$: Media muestral
- $\mu$: Media poblacional
- $s^2$: Varianza muestral
- $n$: Cantidad de datos de la muestra
- *Subindice 1: Dato perteneciente a población 1*
- *Subindice 2: Dato perteneciente a población 2*

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

### 14.7.3 Caso 3: ($\boldsymbol{\sigma}$ desconocida, muestra $< \mathbf{30}$)

En vez de trabajar con las varianzas muestrales, trabajo con un promedio de estas.

- $t$: estadístico a utilizar
- $\bar{X}$: Media muestral
- $\mu$: Media poblacional
- $s_p^2$: Varianza muestral promedio
- $n$: Cantidad de datos de la muestra
- *Subindice 1: Dato perteneciente a población 1*
- *Subindice 2: Dato perteneciente a población 2*

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

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Los grados de libertad a utilizar se dan por la fórmula:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n+1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n+1}\right)} - 2$$

Pero como esta es muy complicada, se suele utilizar la simplificación:

$$\boxed{gl = n_1 + n_2 - 2}$$

## 14.8 APLICADO A LA DIFERENCIA DE PROPORCIONES

- $Z$: Estadístico a utilizar
- $p$: Proporción muestral favorable
- $\bar{p}$: Proporción promedio favorable
- $P$: Proporción poblacional favorable
- $x$: Valores favorables de cad amuestra.
- $n$: Cantidad de datos de la muestra
- *Subindice 1: Dato perteneciente a población 1*
- *Subindice 2: Dato perteneciente a población 2*

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

## 14.9 CÁLCULO ERROR TIPO II

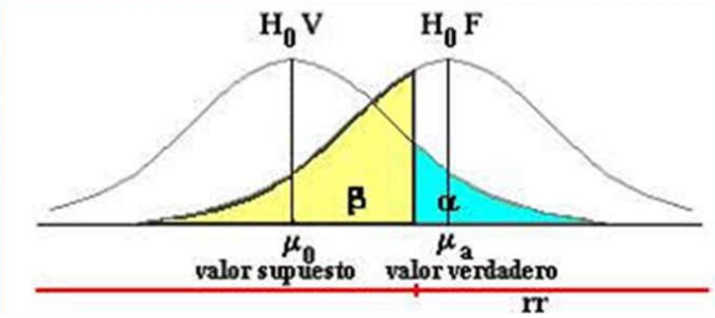
Visto en Decisión Estadística:

*$\beta$ Es la probabilidad de cometer el error tipo II.*

$$\beta = P(Error\ tipo\ II) = P\left(\frac{Aceptar\ H_0}{H_0\ es\ falsa}\right)$$

Lo que nosotros **vamos a calcular,** es la **probabilidad de cometer** el error tipo 2.

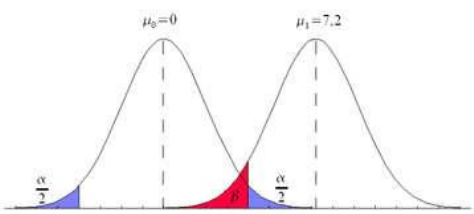
Para representar $\beta$, recurrimos al siguiente grafico

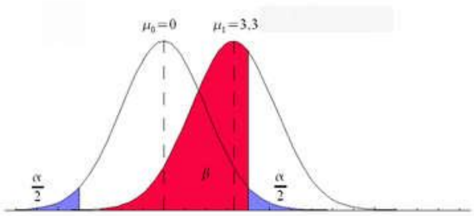



*Tenemos la prueba lateral derecha de hipótesis, del lado izquierdo.*

Para plantear el cálculo del error tipo II, establecemos un valor verdadero $\mu_a$ para la población arbitrario. Planteamos una nueva distribución centrada allí, cuya parte $\beta$ será la que recaiga dentro de mi distribución original, para los valores que aceptaría en mi hipótesis.

También aplica lo mismo para las pruebas bilaterales.

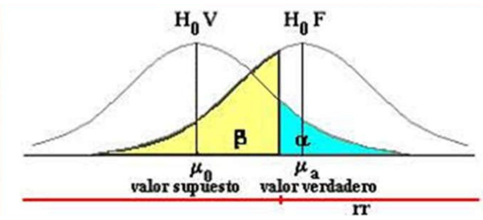



Tomando la prueba original (izquierda), calculamos el punto crítico donde se unen $\alpha$ con $\beta$.

Esto lo calculo con:

$$\boxed{\bar{X} = \mu \pm Z_\alpha \frac{\sigma}{\sqrt{n}}}$$




- Si buscamos un valor a la derecha de la media, uso $\boxed{+}$.
- Si busco un valor a la izquierda de la media, uso $\boxed{-}$.

Una vez calculado $\bar{X}$, lo usamos en la misma ecuación del problema original para encontrar el nuevo valor de Z que corresponde a la distribución de $\beta$. Con ese Z, podemos calcular la probabilidad de cometer el error tipo II, ósea $\beta$.

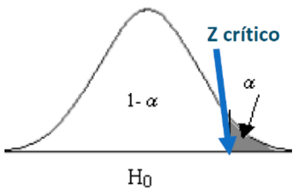
### 14.9.1 Ejemplo

Un camionero transporta bolsas de Harina de promedio de 50kg con una varianza de 4kg ². Para no pasarse del peso necesita probar con una significación del 5% que las bolsas no superan ese peso

Para eso toma una muestra de 25 bolsas y le da una media de 51 kg con una desviación estándar de 3kg.

*Cuál es la **probabilidad de cometer el error tipo II** y cuál es la potencia de la Prueba si en realidad la muestra pertenece a una población con media de 51,5kg*

Lo que nos interesa acá es las bolsas **no se pasen en peso**. Utilizaremos una **Prueba lateral derecha**. Mi valor crítico está a la derecha. Para este lado rechazo.




$$H_0 = \mu \leq 50kg \qquad H_1 = \mu > 50kg$$

Tomando el valor de Z previamente calculado aquí, tenemos:

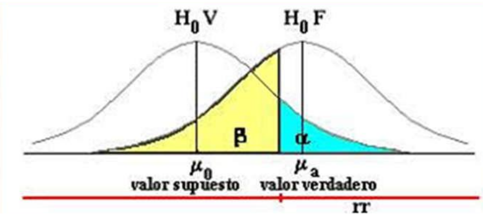
$$\boldsymbol{Z(1-\alpha) = t(\infty, \alpha) = 1,645}$$

Ahora, buscamos el valor de $\bar{X}$ con el valor de $Z$, usando los datos de la prueba:

$$\bar{X} = \mu \pm Z_\alpha \frac{\sigma}{\sqrt{n}} = \underbrace{50kg + 1,645 \frac{2kg}{\sqrt{25}}}_{\substack{Uso\ \boxed{+}\ ya\ que\ busco \\ un\ valor\ a\ la\ derecha \\ de\ la\ media\ original.}} = 50,568kg$$

Ahora con este nuevo $\bar{X}$, podemos pasar al siguiente gráfico, donde vamos a sacar el $Z$

Volvemos a la ecuación original, pero el $\mu$ con el que trabajamos, no es más el del problema original, sino el que nos indica como la media real de la muestra (ver resaltado amarillo arriba)




$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{50,568kg - 51,5kg}{\frac{2kg}{\sqrt{25}}} = -2,11$$

Ahora, por último, buscamos en la misma tabla de distribución, este valor de Z:

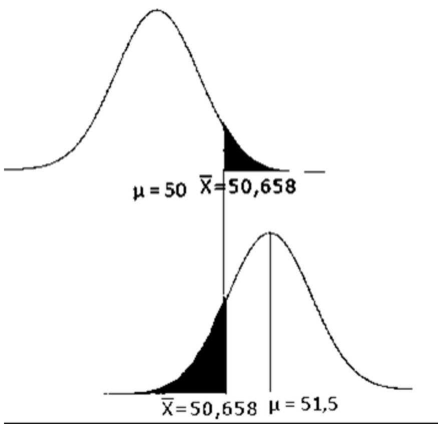
$$P(Z < -a) = 1 - P(Z < a) \quad , \quad siendo\ a = 2,11$$

Entonces la probabilidad de cometer el error tipo II será de:

$$\beta = 1 - P(Z < a) = 1 - P(Z < 2,11) = 1 - 0,9826 = 0,0174$$

$$\beta = P(error\ tipo\ II) = 0,0174$$

$$Potencia\ de\ una\ prueba = 1 - \beta = 0,9826$$



# 15 CONTROL ESTADÍSTICO DE LA CALIDAD

La **calidad** es la totalidad de particularidades y características de un producto o servicio que inciden sobre su adecuación para satisfacer determinadas necesidades.

Un **proceso esta sometido a una serie de factores de carácter aleatorio** que **hacen imposible** que **dos productos** o servicios **sean iguales.**

Las **características** no son uniformes, por lo que presentan una **variabilidad** que siempre va a existir.

## 15.1 VARIABILIDAD

Esta **Variabilidad** es **indeseable**, el objetivo es **reducirla**, o **acotarla a ciertos límites**.

### 15.1.1 Causas de variación

- **Causas aleatorias o normales:** Donde se producen variaciones pequeñas, que provienen de factores que intervienen en el proceso. Ejemplo:
  *Desgaste de bujes.*
  *Cambio de material.*
  *Vibración de una máquina.*
- **Causas asignables o anormales:** Donde se producen variaciones **grandes** e **inusuales**. Ejemplo:
  *Mal puesta a punto.*
  *Material de mala calidad.*
  *Error de operario.*

## 15.2 MUESTREO DE ACEPTACIÓN

Tomamos una **muestra**, y realizamos una **prueba** para ver si **cumple las especificaciones**. Con esta información podemos **aprobar** un lote.

Esta prueba permite **realizar estudios**, **detectar defectos** y **medir la calidad**, a un grado especifico de certeza **sin tener que probar cada uno de los productos** o servicios.

Consiste en realizar una prueba de hipótesis:

| | $H_0\ verdadera$ | $H_0\ falsa$ |
|---|---|---|
| Continuar el proceso | **Decisión correcta** | **Error tipo II**<br>Continua el proceso, **fuera de control.** |
| Ajustar el proceso | **Error tipo I**<br>Ajusta el proceso **que está bajo control**. | **Decisión correcta** |

### 15.2.1 Ventajas

- **Costo**: es menos costoso a otro tipo de estudio. Las pruebas se realizan con menos productos.
- **Daños**: Los productos **no se dañan** debido al muestreo.
- **Reduce el riesgo de errores** durante la inspección.
- Al identificar los errores, los empleados se sienten motivados para realizar las mejoras.

### 15.2.2 Desventajas

- Existe el riesgo de encontrar lotes de productos que se encuentren dañados **en su mayoría**. Lo que lleva a **rechazar los buenos.**
- Ofrece **poca información** sobre la calidad de los productos y del proceso de creación.
- Requiere **mucho tiempo** de planificación y documentación.

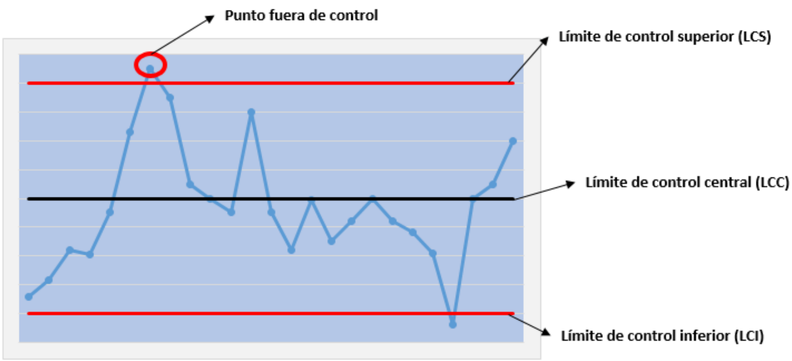
## 15.3 CONTROL ESTADÍSTICO DEL PROCESO

El Control Estadístico de Procesos se basa en **analizar la información aportada** por el proceso para **detectar** la presencia de **causas asignables** y habitualmente **se realiza mediante** una construcción gráfica denominada **Gráfico de Control.**

Para que tenga sentido la aplicación de los gráficos de control, **el proceso ha de tener una estabilidad suficiente** que, aun siendo aleatorio, **permita un cierto grado de predicción**

### 15.3.1 Gráfico de control

Es una de las herramientas de análisis y solución de problemas. Es un diagrama que muestra los valores producto de **la medición de una característica** de calidad, ubicados en una **serie cronológica.**




En él establecemos una **línea central** o valor nominal, que **suele ser el objetivo del proceso o el promedio histórico**, junto a **uno o más límites de control**, tanto superior como inferior, **usados para determinar cuándo es necesario analizar una eventualidad**

Los clasificamos en 2 tipos:

#### *15.3.1.1 Por atributo (discretos)*

Se utiliza cuando una característica de calidad no se puede medir en forma continua. Es decir, es defectuoso o no.

Permite establecer controles respecto al número de unidades defectuosas producidas.

Si la característica de la calidad es medible, cabe la posibilidad de darle un tratamiento como variable o como atributo.

##### 15.3.1.1.1 Tipos de gráficos:

- **pn**: número de unidades defectuosas.
- **p:** Proporción de piezas defectuosas.
- **c:** Número de defectos.
- **u**: Número de defectos **por unidad**.

#### *15.3.1.2 Por variables (continuos)*

Nos permite estudiar la calidad de características **numéricas**. Estos gráficos permiten detectar **pequeños cambios en la media** del proceso, teniendo capacidad de controlar **pequeñas variaciones**.

##### 15.3.1.2.1 Tipos de gráficos

- $\boldsymbol{\bar{x} - R}$**:** Calidad en función de una variable *(longitud, peso, temperatura)*
- **X:** Valor medido

# 15.4 GRÁFICO $\bar{x} - R$

Es un gráfico de **medias** y de **rangos** que se interpretan conjuntamente. Es el mas completo de todos y el que vamos a ver en esta materia.

A diferencia de otros gráficos son dos gráficos que nosotros interpretamos de forma conjunta.

## 15.4.1 Construcción

### *15.4.1.1 Caso 1: (μ y σ conocidos)*

$$LCC = \mu \qquad LSC = \mu + 3\frac{\sigma}{\sqrt{n}} \qquad LIC = \mu - 3\frac{\sigma}{\sqrt{n}}$$

- σ: Varianza poblacional
- μ: Media poblacional
- $n$: **Tamaño de cada muestra**
- $LCC$: Límite central de control
- $LSC$: Límite superior de control
- $LIC$: Límite inferior de control

**Coeficientes para estimaciones de σ en Gráficos $\bar{X}$ - R**

| Tamaño Subgrupo | Gráfico X A2 | Gráfico R D3 | D4 | d2 |
|---|---|---|---|---|
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |

n es el tamaño de cada muestra
No confundir n con la cantidad de muestras

### *15.4.1.2 Caso 2: (μ conocido, σ no)*

Si no conocemos σ solo lo estimamos con:

$$\sigma = \frac{\bar{R}}{d2}$$

- $\bar{R}$: Media del rango
- $d2$: Sale de la tabla de arriba:

### *15.4.1.3 Caso 3: (μ y σ desconocidos)*

Uso la tabla de la derecha.

- $\bar{\bar{x}}$: **Media de las medias muestrales.**

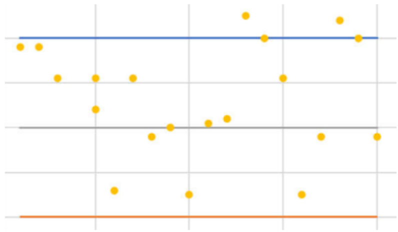
| | Gráfico $\bar{x}$ | Gráfico $R$ |
|---|---|---|
| $LCC$ | $\bar{\bar{x}}$ | $\bar{R}$ |
| $LSC$ | $\bar{\bar{x}} + A2\,.\,\bar{R}$ | $D4\,.\,\bar{R}$ |
| $LIC$ | $\bar{\bar{x}} - A2\,.\,\bar{R}$ | $D3\,.\,\bar{R}$ |

### 15.4.2 Análisis

Para determinar si debemos parar o no el proceso, analizaremos diferentes aspectos:

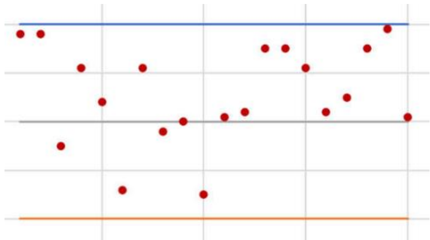
#### *15.4.2.1 Puntos fuera de los límites de control*

Si hay un solo punto, puede ser un error de medición. Si tengo **mas de uno**, puedo descartar que sea un error de medición, y necesariamente debo parar el proceso.



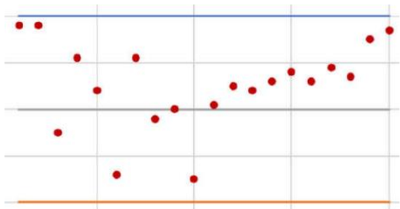
#### *15.4.2.2 Corridas o desplazamiento*

Si tengo **8 puntos del mismo lado**



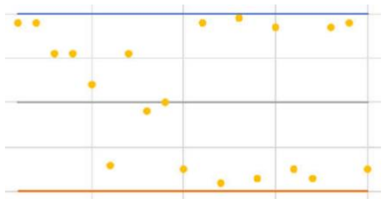
#### *15.4.2.3 Tendencia*

Si tengo **6 o 7 puntos <u>sucesivos</u> en igual tendencia.**



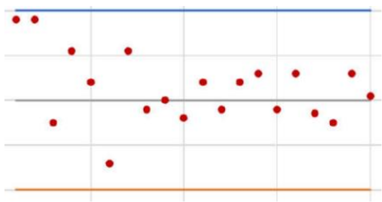
#### *15.4.2.4 Acercamiento a los límites*

Si tengo **2 de 3 puntos fuera de las líneas $2\sigma$.** Es decir, pegados a los límites.



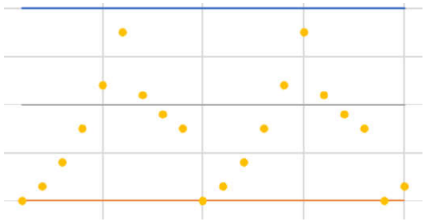
#### *15.4.2.5 Acercamiento a línea central*

Si tengo **todos los puntos dentro de las líneas $1.5\sigma$**. Es este caso, de repente el proceso tiene una menor variación. **NO HACE FALTA PARAR EL PROCESO**. Hace falta, con estos datos nuevos **CALCULAR NUEVAMENTE EL GRAFICO**.



#### *15.4.2.6 Periodicidad o Ciclos*

Si observo que se asimilan dos partes del grafico distintas.

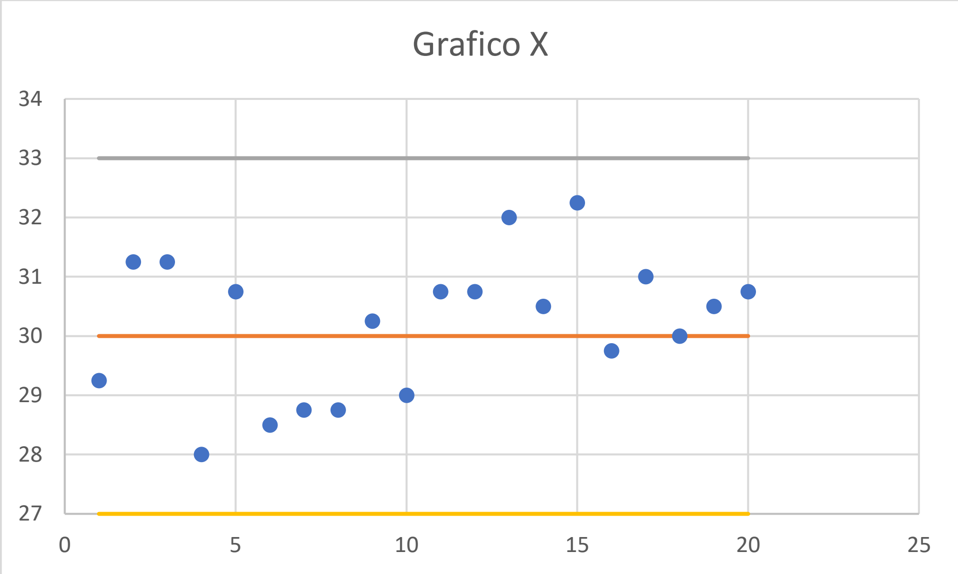


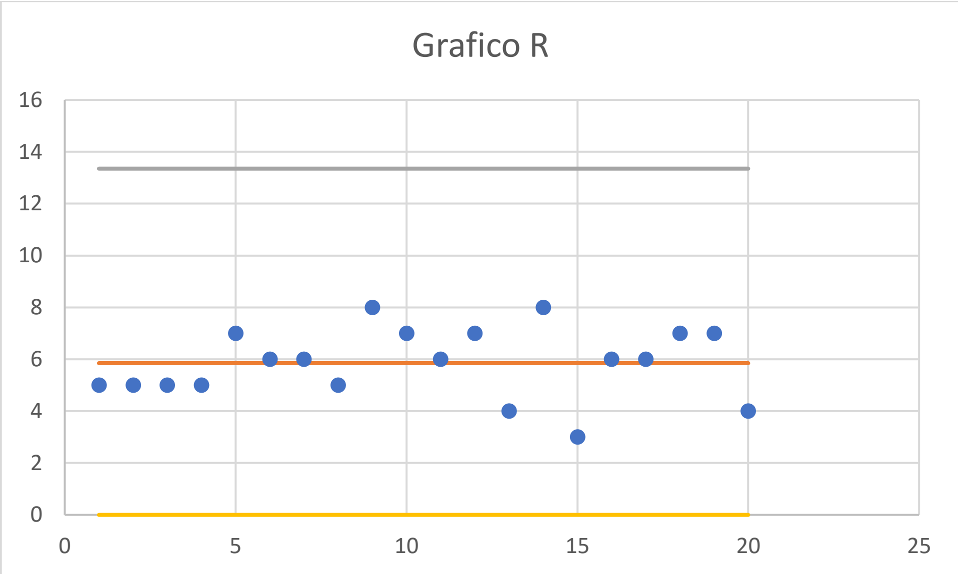
### 15.4.3 Ejemplo

Se dispone de un cierto proceso que se desea verificar si permanece bajo control. Se conoce que la **media** aritmética es **30 cm** y la **varianza 4 cm²**. Para su control se desea construir un **gráfico X-R** con los siguientes datos obtenidos del proceso y determinar si el mismo está bajo control

**Este es del tipo** Caso 1

| A | B | C | D | $\bar{x}$ | R |
|---|---|---|---|---|---|
| 30 | 28 | 27 | 32 | **29,25** | **5** |
| 28 | 32 | 32 | 33 | **31,25** | **5** |
| 34 | 33 | 29 | 29 | **31,25** | **5** |
| 31 | 29 | 26 | 26 | **28** | **5** |
| 28 | 34 | 27 | 34 | **30,75** | **7** |
| 26 | 32 | 29 | 27 | **28,5** | **6** |
| 26 | 27 | 32 | 30 | **28,75** | **6** |
| 28 | 31 | 30 | 26 | **28,75** | **5** |
| 33 | 28 | 34 | 26 | **30,25** | **8** |
| 29 | 28 | 26 | 33 | **29** | **7** |
| 31 | 33 | 27 | 32 | **30,75** | **6** |
| 32 | 33 | 32 | 26 | **30,75** | **7** |
| 33 | 31 | 30 | 34 | **32** | **4** |
| 26 | 29 | 33 | 34 | **30,5** | **8** |
| 31 | 31 | 33 | 34 | **32,25** | **3** |
| 29 | 30 | 27 | 33 | **29,75** | **6** |
| 34 | 31 | 31 | 28 | **31** | **6** |
| 27 | 34 | 31 | 28 | **30** | **7** |
| 31 | 32 | 33 | 26 | **30,5** | **7** |
| 33 | 29 | 29 | 32 | **30,75** | **4** |




Como puedo observar, en el grafico X al final tengo 8 de los últimos 10 puntos sobre la línea central. Por lo que detengo el proceso.

# 16 ANÁLISIS DE RELACIÓN ENTRE VARIABLES

Hay 2 tipos: Análisis de regresión y Análisis de correlación.

## 16.1 ANÁLISIS DE REGRESIÓN

Tenemos 2 partes fundamentales:

1. Elaborar una línea que represente a una ecuación derivada de los puntos de la relación entre las dos variables.
2. Efectuar predicciones basándose en ecuaciones matemáticas.

### 16.1.1 Diagrama de dispersión

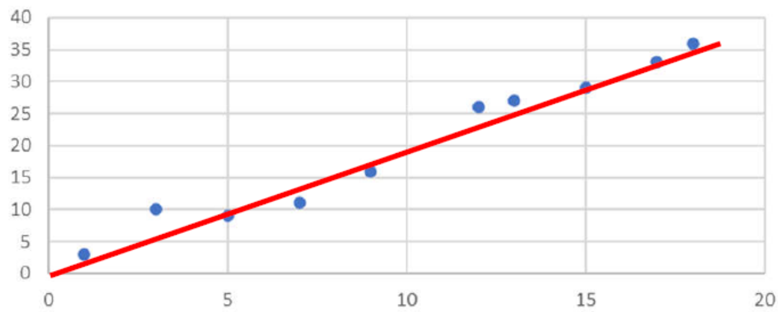
Es una gráfica bidimensional, en donde se trazan los valores observados o individuales de cada una de las variables. Cada valor se traza en sus coordenadas X e Y.

#### *16.1.1.1 Ecuación de la regresión lineal*

$$y = a + bx$$

### 16.1.2 Métodos de regresión

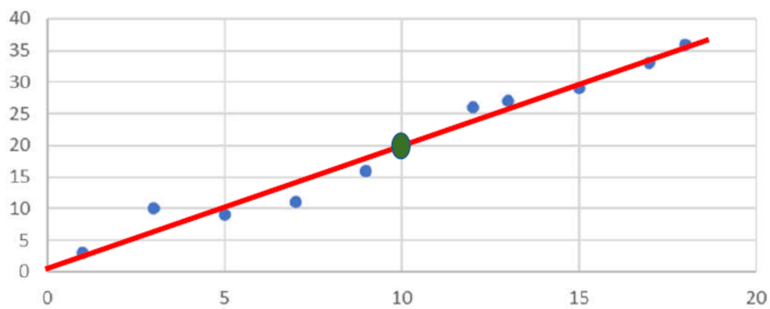
#### *16.1.2.1 Mano alzada (puro)*




| $X$ | $Y$ |
|---|---|
| 1 | 3 |
| 3 | 10 |
| 5 | 9 |
| 7 | 11 |
| 9 | 16 |
| 12 | 26 |
| 13 | 27 |
| 15 | 29 |
| 17 | 33 |
| 18 | 36 |

Simplemente, dado el diagrama de dispersión, trazamos una línea donde me parezca que represente mejor a los puntos.

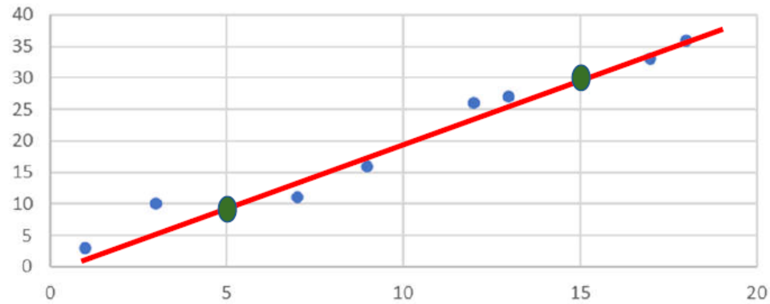
#### *16.1.2.2 Mano alzada (corregido)*




| $\bar{X}$ | $\bar{Y}$ |
|---|---|
| 10 | 20 |

Lo mismo que el método anterior, pero ajustándola para que pase por la media de los valores X y la media de los valores Y

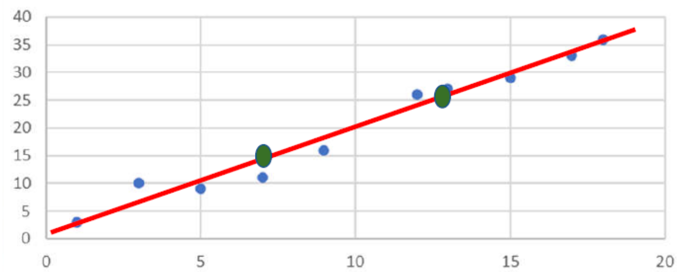
### *16.1.2.3 Semipromedios*




Divido los datos en 2 grupos **iguales**, calculo las medias en X e Y, y hago una recta que pase por los dos puntos.

| $X$ | $\bar{X}$ | $Y$ | $\bar{Y}$ |
|---|---|---|---|
| 1 | | 3 | |
| 3 | | 10 | |
| 5 | 5 | 9 | 9,8 |
| 7 | | 11 | |
| 9 | | 16 | |
| 12 | | 26 | |
| 13 | | 27 | |
| 15 | 15 | 29 | 30,2 |
| 17 | | 33 | |
| 18 | | 36 | |

### *16.1.2.4 Promedios móviles*




Divido los datos en 2 grupos **no iguales**, calculo las medias en X e Y, y hago una recta que pase por los dos puntos.

| $X$ | $\bar{X}$ | | $Y$ | $\bar{Y}$ | |
|---|---|---|---|---|---|
| 1 | | | 3 | | |
| 3 | | | 10 | | |
| 5 | | | 9 | | |
| 7 | 7,1 | | 11 | 14,6 | |
| 9 | | | 16 | | |
| 12 | | | 26 | | |
| 13 | | 13,0 | 27 | | 25,4 |
| 15 | | | 29 | | |
| 17 | | | 33 | | |
| 18 | | | 36 | | |

### *16.1.2.5 Mínimos Cuadrados*

$$y = a + bx$$

$$a = \frac{\sum X_i^2 . \sum Y_i - \sum X_i . \sum X_i . Y_i}{n . \sum X_i^2 - (\sum X_i)^2}$$

$$b = \frac{n . \sum X_i . Y_i - \sum X_i . \sum Y_i}{n . \sum X_i^2 - (\sum X_i)^2}$$

| $X_i$ | $Y_i$ | $X_i^2$ | $X_i . Y_i$ |
|---|---|---|---|
| 1 | 3 | 1 | 3 |
| 3 | 10 | 9 | 30 |
| 5 | 9 | 25 | 45 |
| 7 | 11 | 49 | 77 |
| 9 | 16 | 81 | 144 |
| 12 | 26 | 144 | 312 |
| 13 | 27 | 169 | 351 |
| 15 | 29 | 225 | 435 |
| 17 | 33 | 289 | 561 |
| 18 | 36 | 324 | 648 |
| **100** | **200** | **1316** | **2606** |

| | |
|---|---|
| n | 10 |
| a | 0,8228 |
| b | 1,9177 |

**y=0,8228+1,9177x**

## 16.2 ANÁLISIS DE CORRELACIÓN

Es el conjunto de técnicas estadísticas empleado para medir la intensidad de la asociación entre dos variables. El principal **objetivo del análisis de correlación** consiste en determinar qué tan intensa es la **relación** entre dos variables.

### Análisis de correlación lineal

Indica que tan precisa es la recta para representar la relación y predecir valores, es decir, que tan confiable son las proyecciones que realice para la toma de decisiones.

### 16.2.1 Error estándar de la estimación

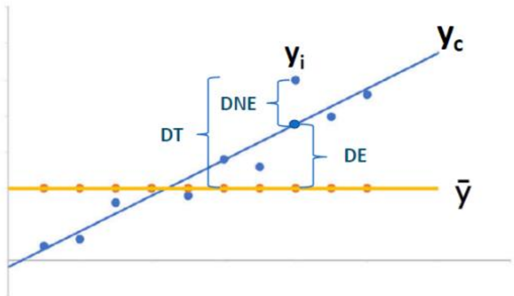
Indica la dispersión de los puntos con respecto a la recta de regresión

$$S_{yx} = \sqrt{\frac{\sum(y_i - y_c)^2}{n-2}} = \sqrt{\frac{\sum y_i^2 - a\sum y_i - b\sum x_i . y_i}{n-2}}$$

- $y_c$ Y calculado. Tomando el punto $x_i$, usamos la función de regresión calculada previamente. $\boldsymbol{y_c = a + b . x_i}$
- Si no calculo la función de regresión, utilizo la parte derecha de la ecuación.

### 16.2.2 Coeficiente de determinación $\boldsymbol{r^2}$

Es la proporción de la varianza total de la variable explicada por la regresión




$$\underbrace{y_i - \bar{y}}_{\substack{Desv.\\Total}} = \underbrace{(y_i - y_c)}_{\substack{Desv.\ No\\Explicada}} + \underbrace{(y_c - \bar{y})}_{\substack{Desv.\\Explicada}}$$

$$\underbrace{\sum (y_i - \bar{y})^2}_{\substack{Variación\\Total}} = \underbrace{\sum (y_i - y_c)^2}_{\substack{Variación\ No\\Explicada}} + \underbrace{\sum (y_c - \bar{y})^2}_{\substack{Variación\\Explicada}}$$

$$\boxed{r^2 = \frac{Var.Explicada}{Var.Total} = \frac{\sum(y_c - \bar{y})^2}{\sum(y_i - \bar{y})^2}}$$

$$0 \le r^2 \le 1$$

- **DT: Desviación total.**
  Desviación entre $\boldsymbol{y_i}$ y el $\bar{\boldsymbol{y}}$
- **DNE: Desviación no explicada.**
  Diferencia entre $\boldsymbol{y_i}$ y el $\boldsymbol{y_c}$
- **DE: Desviación explicada.**
  Desviación entre $\boldsymbol{y_c}$ y el $\bar{\boldsymbol{y}}$

### 16.2.3 Coeficiente de correlación lineal $r$

Es una medida estadística que cuantifica la dependencia lineal entre dos variables, es decir, indica lo bien o mal que los datos están representados en la recta de regresión.

Esto indica *que tan bien la recta representa los puntos y nos sirve para la toma de decisiones.*

$$\boxed{r = \sqrt{r^2} = \frac{n.\sum x_i . y_i - \sum x_i . \sum y_i}{\sqrt{[n.\sum x_i^2 - (\sum x_i)^2].[n.\sum y_i^2 - (\sum y_i)^2]}}}$$

- Si no calculo la función de regresión, utilizo la parte derecha de la ecuación.

- Si $\boldsymbol{r = 1}$ Existe una **correlación positiva perfecta.**
  El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en proporción constante.
- Si $\boldsymbol{0 < r < 1}$ Existe una **correlación positiva.**
- Si $\boldsymbol{r = 0}$ no existe una **relación lineal**.
  pueden existir todavía relaciones no lineales entre las dos variables.
- Si $\boldsymbol{-1 < r < 0}$ existe una **correlación negativa.**
- Si $\boldsymbol{r = -1}$ Existe una **correlación negativa perfecta.**
  existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en proporción constante.

- ✓ $\boldsymbol{+1}$ o $\boldsymbol{-1}$ **Correlación Perfecta.**
- ✓ $\boldsymbol{0.95}$ **Correlación fuerte.**
- ✓ $\boldsymbol{0.80}$ **Correlación significativa.**
- ✓ $\boldsymbol{0.70}$ **Correlación moderada.**
- ✓ $\boldsymbol{0.50}$ **Correlación parcial.**

# 17 ANÁLISIS DE VARIANZA

## 17.1 PRINCIPIOS GENERALES

- **Variable de respuesta:** Una variable de respuesta es un resultado medido dentro de un ensayo que puede ser influenciado por otros factores **(serian nuestros datos)**
- **Tratamientos:** grupo de las variables de respuesta **(serian nuestras muestras)**
- **Factores:** variaciones provocadas en los tratamientos que condicionan las variables de respuesta **(Serían nuestras variaciones)**

## 17.2 DEFINICIÓN

Es un método experimental que se utiliza para comparar tres o más medias y determinar si pueden pertenecen a la misma población.

El objetivo es determinar si existe una diferencia significativa entre los tratamientos.

$$\mu_1 = \mu_2 = \mu_3 = \mu_i$$

*Lo que vamos a realizar es una <u>prueba de hipótesis</u>. Donde comparamos 3 o mas medias, a través de los experimentos, definimos si existe o no diferencias significativas entre los tratamientos.*

*Sí $H_0 = Verdadera$ entendemos que todos los tratamientos pueden pertenecer a la misma población.*

*Sí $H_0 = Falsa$ entendemos que existen diferencias significativas entre los tratamientos, por lo que no pueden pertenecer a la misma población.*

## 17.3 SUPUESTOS

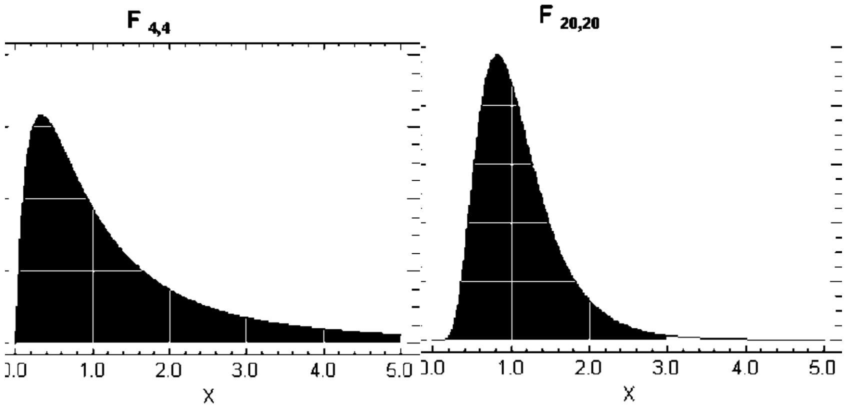
Estos 3 supuestos se deben cumplir **Sí o sí** para poder trabajar con **Análisis de Varianza**.

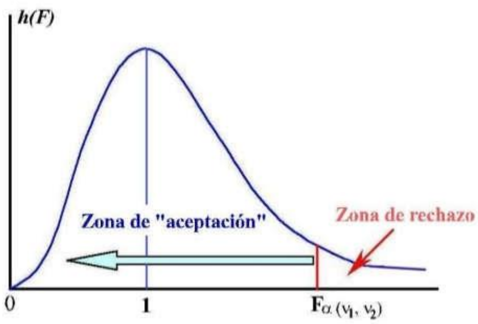
1. Para cada población, la variable de respuesta está totalmente distribuida.
   *Nosotros vamos a trabajar con <u>poblaciones normales.</u>*
2. La varianza de la variable de respuesta es la misma para todas las poblaciones.
3. Las observaciones deben ser independientes
   *Como trabajo con experimentos, los debo plantear de tal forma que las observaciones que me va a dar la variable de respuesta sean **independientes entre sí.***

## 17.4 DISTRIBUCIÓN F

Es una distribución de probabilidad continua, también conocida como distribución F de Snedecor (por George Snedecor) Se usa en la contrastación de la igualdad de varianzas de dos poblaciones normales, y , fundamentalmente en el análisis de la varianza

Tiene 2 grados de libertad: $\boldsymbol{m}$ y $\boldsymbol{n}$, se representa $F_{m;n}$ , $\boldsymbol{m}$ grados de libertad en el numerador y $\boldsymbol{n}$ grados de libertad en el denominadorabe




En la tabla, vamos a tener, en la parte superior los **grados de libertad del numerador.** Y en la parte inferior los **grados de libertad del denominador**.

Buscando estos datos, vamos a tener el **valor crítico de F**.

Hay 2 tablas, una con $\boldsymbol{\alpha = 0,05}$ y otra con $\boldsymbol{\alpha = 0,01}$

Nuestra zona de rechazo esta a la derecha, y la zona de aceptación a la izquierda.

## 17.5 DISEÑO DE EXPERIMENTOS

### 17.5.1 Completamente Aleatorio.

Este diseño consiste en la asignación de los tratamientos en forma completamente aleatoria a las variables de respuestas ya que no existe un factor que lo condicione

### 17.5.2 Bloques aleatorizados

Este diseño consiste en la asignación de los tratamientos en forma de bloques a las variables de respuestas ya que existe un factor que lo condicione

Es apropiado cuando se desea investigar las diferencias entre los promedios de k tratamientos en condiciones homogéneas, vale decir, eliminando las diferencias iniciales entre variables de respuesta

### 17.5.3 Multifactorial

Este procedimiento está diseñado para construir un modelo estadístico describiendo el impacto de dos o más factores que condicionan la variable de respuesta.

### 17.5.4 Diseños en cuadrados latinos

Los diseños en cuadrados latinos son apropiados cuando es necesario controlar dos fuentes de variabilidad. En dichos diseños el número de niveles del factor principal tiene que coincidir con el número de niveles de las dos variables de bloque o factores secundarios y además hay que suponer que no existe interacción entre ninguna pareja de factores.

## 17.6 EXPERIMENTOS COMPLETAMENTE ALEATORIOS

$$\boxed{\underbrace{SST}_{\substack{Suma\ de\ los\\ Cuadrados\\ Totales}} = \underbrace{SSTr}_{\substack{Suma\ de\ los\\ Cuadrados\\ debido\ al\\ tratamiento}} + \underbrace{SSE}_{\substack{Suma\ de\ los\\ Cuadrados\\ debido\ al\\ error}}}$$

$$\boxed{SSTr = \sum n_j\left(\bar{X}_j - \bar{\bar{X}}\right)^2}$$

$$\boxed{SSE = \sum \left(n_j - 1\right)s_j^2}$$

- $\boldsymbol{n_j}$ Tamaño del tratamiento
- $\boldsymbol{s_j^2}$ Varianza del tratamiento
- $\boldsymbol{\bar{X}_j}$ Media del tratamiento
- $\boldsymbol{\bar{X}_{ij}}$ Variables de respuesta
- $\boldsymbol{\bar{\bar{X}}}$ Media de medias de los tratamientos

También podemos calcular $SST$ con:

$$\boxed{SST = \sum\sum\left(\bar{X}_{ij} - \bar{\bar{X}}\right)^2}$$

Con esto calculado, calcularemos nuestro $\boldsymbol{F}$

$$\boxed{F = \frac{MSTr}{MSE} = \frac{\boldsymbol{Media}\ de\ la\ \boldsymbol{Suma\ de\ los\ Cuadrados}\ debidos\ al\ \boldsymbol{Tratamiento}}{\boldsymbol{Media}\ de\ la\ \boldsymbol{Suma\ de\ los\ Cuadrados}\ debido\ al\ \boldsymbol{Error}}}$$

Este valor nos dice si aceptaremos o rechazaremos la $\boldsymbol{H_0}$

$$\boxed{MSTr = \frac{SSTr}{(k-1)}} \quad ; \quad \boxed{MSE = \frac{SSE}{(n-k)}}$$

Estos denominadores se denominan $\boldsymbol{grados\ \ de\ libertad}$ y salen de la ecuación

$$n - 1 = (k - 1) + (n - k)$$

Donde $\boldsymbol{k}$ es la **cantidad de tratamientos** y $\boldsymbol{n}$ es la **cantidad total de variables de respuesta**

$$GL_{MSTr} = (k - 1) \quad ; \quad GL_{MSE} = (n - k)$$

### 17.6.1 Ejemplo

$$\mu_1 = \mu_2 = \mu_3 \qquad \boldsymbol{\alpha = 0,01}$$

$\bar{\bar{X}} = 6 \quad n_j = 6$

$$SSTr = \sum n_j \left(\bar{X}_j - \bar{\bar{X}}\right)^2$$
$$SSTr = 6 * (6.5 - 6)^2 + 6 * (6 - 6)^2 + 6 * (5.5 - 6)^2$$
$$SSTr = 1.5 + 0 + 1.5 = 3$$

$$SSE = \sum (n_j - 1) s_j^2$$
$$SSE = (6 - 1) * 1.9 + (6 - 1) * 1.2 + (6 - 1) * 1.1$$
$$SSE = 9.5 + 6 + 5.5 = 21$$

$$SST = SSTr + SSE = 3 + 21 = 24$$

| | Trat. 1 | Trat. 2 | Trat. 3 |
|---|---|---|---|
| **DATOS** | 5 | 6 | 7 |
| | 6 | 4 | 5 |
| | 7 | 6 | 5 |
| | 6 | 7 | 6 |
| | 9 | 6 | 6 |
| | 6 | 7 | 4 |
| $\bar{X}$ | **6.5** | **6** | **5.5** |
| $s^2$ | **1.9** | **1.2** | **1.1** |

$$MSTr = \frac{SSTr}{(k-1)} \qquad MSE = \frac{SSE}{(n-k)}$$
$$MSTr = \frac{3}{(3-1)} = 1.5 \qquad MSE = \frac{21}{(18-3)} = 1.4$$

$$\underbrace{F = \frac{MSTr}{MSE} = \frac{1.5}{1.4} = 1.0714}_{\textbf{F calculado}}$$

$$\underbrace{Fc = \frac{GL_{MSTr}}{GL_{MSE}} = \boxed{\overbrace{2}^{GL.\ del\ numerador} \Big/ \underbrace{15}_{GL.\ denominador}} = \overbrace{6.36}^{Sale\ de\ la\ tab}}_{\textbf{F calculado}}$$

Como el $\boxed{F < F_c}$ apruebo $H_0$: No existen diferencias significativas en los tratamientos.

## 17.7 Experimentos por bloques aleatorizados

$$\boxed{\underbrace{SST}_{\substack{Suma\ de\ los\\ Cuadrados\\ Totales}} = \underbrace{SSTr}_{\substack{Suma\ de\ los\\ Cuadrados\\ debido\ al\\ tratamiento}} + \underbrace{SSE}_{\substack{Suma\ de\ los\\ Cuadrados\\ debido\ al\\ error}} + \underbrace{SSBI}_{\substack{Suma\ de\ los\\ Cuadrados\\ debido\ al\\ bloque}}}$$

$$\boxed{SSTr = \sum n_j\left(\bar{X}_j - \bar{\bar{X}}\right)^2}$$

$$\boxed{SSBI = \sum n_i\left(\bar{X}_i - \bar{\bar{X}}\right)^2}$$

$$\boxed{SSE = SST - SSTr - SSTBI}$$

- $\boldsymbol{n_j}$ Tamaño del tratamiento
- $\boldsymbol{n_i}$ Tamaño del bloque
- $\boldsymbol{s_j^2}$ Varianza del tratamiento
- $\boldsymbol{\bar{X}_j}$ Media del tratamiento
- $\boldsymbol{\bar{X}_i}$ Media del bloque
- $\boldsymbol{\bar{X}_{ij}}$ Variables de respuesta
- $\boldsymbol{\bar{\bar{X}}}$ Media de medias de los tratamientos

También podemos calcular $SST$ con:

$$\boxed{SST = \sum\sum\left(\bar{X}_{ij} - \bar{\bar{X}}\right)^2}$$

Con esto calculado, calcularemos nuestro $\boldsymbol{F}$

$$\boxed{F = \frac{MSTr}{MSE} = \frac{\boldsymbol{Media}\ de\ la\ \boldsymbol{Suma\ de\ los\ Cuadrados}\ debidos\ al\ \boldsymbol{Tratamiento}}{\boldsymbol{Media}\ de\ la\ \boldsymbol{Suma\ de\ los\ Cuadrados}\ debido\ al\ \boldsymbol{Error}}}$$

Este valor nos dice si aceptaremos o rechazaremos la $\boldsymbol{H_0}$

$$\boxed{MSTr = \frac{SSTr}{(k-1)}} \quad ; \quad \boxed{MSE = \frac{SSE}{(k-1)*(b-1)}}$$

Estos denominadores se denominan $\boldsymbol{grados\ \ de\ libertad}$ y salen de la ecuación

$$n - 1 = (k-1) + (k-1)*(b-1) + (b-1)$$

Donde $\boldsymbol{k}$ es la **cantidad de tratamientos**, $\boldsymbol{b}$ es la **cantidad de bloques** y $\boldsymbol{n}$ es la **cantidad total de variables de respuesta**

$$GL_{MSTr} = (k-1) \quad ; \quad GL_{MSE} = (k-1)*(b-1)$$

### 17.7.1 Ejemplo

$$\mu_1 = \mu_2 = \mu_3 \qquad \boldsymbol{\alpha = 0,01}$$

$\bar{\bar{X}} = 6 \quad n_j = 6$

| | OP (bloque) | Trat. 1 | Trat. 2 | Trat. 3 |
|---|---|---|---|---|
| **DATOS** | 1 | 5 | 6 | 7 |
| | 2 | 6 | 4 | 5 |
| | 3 | 7 | 6 | 5 |
| | 4 | 6 | 7 | 6 |
| | 5 | 9 | 6 | 6 |
| | 6 | 6 | 7 | 4 |
| $\bar{X}$ | | **6.5** | **6** | **5.5** |
| $s^2$ | | **1.9** | **1.2** | **1.1** |

$$SSTr = \sum n_j \left(\bar{X}_j - \bar{\bar{X}}\right)^2$$
$$SSTr = 6 * (6.5 - 6)^2 + 6 * (6 - 6)^2 + 6 * (5.5 - 6)^2$$
$$SSTr = 1.5 + 0 + 1.5 = 3$$

$$SST = \sum \sum \left(\bar{X}_{ij} - \bar{\bar{X}}\right)^2 = 24$$

$$SSBI = \sum n_i \left(\bar{X}_i - \bar{\bar{X}}\right)^2$$
$$SSBI = 6.67$$
$$SSBI = 9.5 + 6 + 5.5 = 21$$

$$SSE = 24 - 1.5 - 6.67 = 15.83$$

$$MSTr = \frac{SSTr}{(k-1)} \qquad MSE = \frac{SSE}{(k-1)*(b-1)}$$
$$MSTr = \frac{3}{(3-1)} = 1.5 \qquad MSE = \frac{21}{(3-1)*(6-1)} = 1.583$$

$$\underbrace{F = \frac{MSTr}{MSE} = \frac{1.5}{1.583} = 0.9474}_{F\ calculado} \qquad \underbrace{Fc = \frac{GL_{MSTr}}{GL_{MSE}} = \boxed{\overbrace{2}^{GL.\ del\ numerador} \Big/ \underbrace{10}_{GL.\ denominador}} = \overbrace{7.56}^{Sale\ de\ la\ tab}}_{F\ calculado}$$

Como el $\boxed{F < F_c}$ apruebo $H_0$: No existen diferencias significativas en los tratamientos.