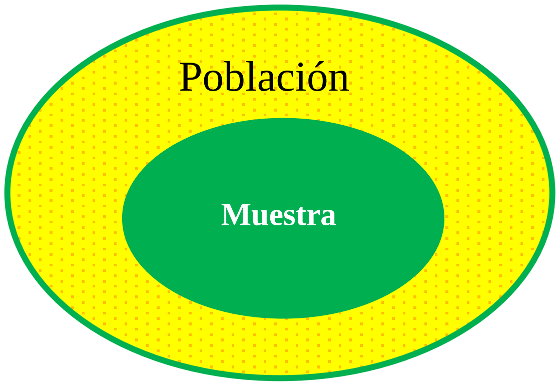
REPASO DE LAS HERRAMIENTAS ESTADÍSTICAS Y MATEMÁTICAS BÁSICAS

POBLACIÓN: conjunto de todos los individuos objeto de estudio estadístico. En teoría de probabilidad es el conjunto de todos los resultados posibles de un experimento aleatorio (al azar), también llamado ***ESPACIO MUESTRAL.***

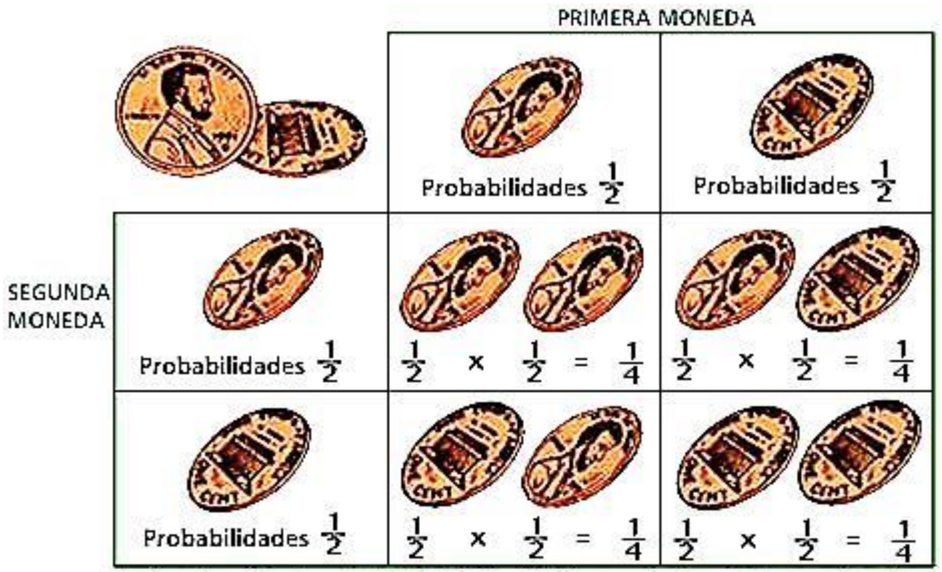
MUESTRA: es un conjunto de individuos parte de la población objeto de estudio, considerados una representación valedera y de interés para la investigación que se va a desarrollar, debe:

- ✓ Representar fielmente la totalidad de la población de donde se deriva.
- ✓ Ser una base válida posible a la aplicación de la teoría de distribución estadística.
- ✓ Permitir medir su grado de probabilidad.
- ✓ **NO DEBE SER SESGADA NI MANIPULADA POR EL INVESTIGADOR.**



PUNTO MUESTRAL: se define así a cada miembro del espacio muestral.

Cuando se realiza el experimento de lanzar dos monedas, el espacio muestral tiene cuatro resultados posibles: CC, CS, SC, SS.



Aplicación de la Ley de Probabilidad del lanzamiento de 2 monedas. La probalidad de obtener cara en el primer lanzamiento es de 1/2, y la probabilidad de obtener sello es tambien 1/2. La misma predicción se puede hacer para la segunda moneda.

EVENTO: es un subconjunto del espacio muestral. Ejemplo:

Si el evento A se define como la ocurrencia de una cara y un sello (CS) en el lanzamiento de dos monedas, entonces de los posibles resultados en la figura anterior solamente dos pertenecen al evento A. (CS o SC).

Los eventos son ***mutuamente excluyentes*** cuando la ocurrencia de uno impide la ocurrencia de otro y ***exhaustivos*** (colectivamente) cuando todos los posibles resultados se agotan.

PROBABILIDAD (P): Sea A un evento en el espacio muestral, P(A) es la probabilidad del evento A, es decir, la proporción de veces que el evento A ocurrirá en ensayos repetidos de un experimento. En un total de *n* posibles resultados con igual probabilidad de ocurrencia en un evento, si *m* de ellos son favorables a A, la razón *m/n* se conoce como la *frecuencia relativa* de A.

Propiedades:

1) $0 \leq P(A) \leq 1$, para todo A.
2) Si A,B,C, …, son un conjunto de eventos exhaustivos, entonces $P(A+B+C+\ldots) = 1$.
3) Si A, B, C, …, son eventos mutuamente excluyentes,
   $P(A+B+C+\ldots) = P(A) + P(B) + P(C) + \ldots$

VARIABLES ALEATORIAS (VA): es una variable cuyo valor está determinado por el resultado de un experimento al azar o aleatorio. Ésta puede ser ***discreta (VAD) o continua (VAC).***

X = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 ó 12 $\longrightarrow$ VAD

Y = [95,72 ; 125,5] $\longrightarrow$ VAC

OPERADORES DE SUMATORIA Y DE PRODUCTO:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \cdots + X_n$$

1) Suma de una constante (k) sobre *n* observaciones:
   $\sum_{i=1}^{n} k = nk$
2) Suma de una constante (k) por una variable:
   $\sum_{i=1}^{n} kX_i = k\sum_{i=1}^{n} X_i$
3) Suma de una constante (a) independiente y una constante (b) por una variable:
   $\sum_{i=1}^{n}(a + bX_i) = na + b\sum_{i=1}^{n} X_i$
4) Suma de las observaciones de dos variables:
   $\sum_{i=1}^{n}(X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$

En el caso de la doble sumatoria:

1) $\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}$
2) $\sum_{i=1}^{n}\sum_{j=1}^{m} X_i Y_j = \sum_{i=1}^{n} X_i \sum_{j=1}^{m} Y_j$
3) $\sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij}+Y_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}$

4) $[\sum_{i=1}^{n} X_i]^2 = \sum_{i=1}^{n} {X_i}^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_i X_j$

El operador producto: $\prod_{i=1}^{n} X_i = X_1 * X_2 * \ldots * X_n$

Función de densidad de probabilidad de una VAD (fmp): Sea X una va discreta que toma valores diferentes x_1, x_2,..., x_n. Entonces, la función $f(x) = P(X = x)\ \forall\ i = 1,2, \ldots n.$ $f(x) = 0$ Para $x \neq x_i$

Se denomina función de densidad de probabilidad discreta (FDP) de X, donde $P(X=x_i)$ significa la probabilidad de que la VADX tome el valor de x_i.

Función de densidad de probabilidad de una VAC (fdp): Sea X una va continua. Entonces, se dice que *f (x)* es la FDP de X si se satisfacen las siguientes condiciones:

- $f(x) > 0$
- $f(x) = \int_{-\infty}^{\infty} f(x)dx = 1$
- $f(x) = \int_{a}^{b} f(x)dx = P(a \leq x \leq b)$

Función de densidad de probabilidad conjunta (FDP conjunta discreta):

Sean X y Y dos variables aleatorias discretas. Entonces, la función $f(x, y) = P(X = x, Y = y)$. Se conoce como función de densidad de probabilidad conjunta discreta y da la probabilidad (conjunta) de que X tome el valor de *x* y Y tome el valor de *y*.

Función de densidad de probabilidad marginal (fdpm): En relación con $f(x, y)$, $f(x)$ y $f(y)$ se denominan funciones de densidad de probabilidad **individuales** o **marginales**. Estas FDP marginales se obtienen de la siguiente manera:

$f(x) = \sum_y f(x, y)$ y $f(y) = \sum_x f(x, y)$

Función de densidad de probabilidad condicional (fdpc): se conoce como FDP condicional de X; da la probabilidad de que X tome el valor de x porque Y asumió el valor de y.

$f(x|y) = P\big((X = x)\big|Y = y\big)$

Independencia: Dos variables aleatorias X y Y son estadísticamente independientes si y sólo si $f(x,y) = f\ (x)f\ (y)$ es decir, si la FDP conjunta se expresa como el producto de las FDP marginales.

CARACTERÍSTICAS DE LAS DISTRIBUCIONES DE PROBABLIDAD

Cada distribución de probabilidad[1] se puede resumir en términos de sus ***momentos.*** Los dos primeros momentos de una distribución son la media o valor esperado y la varianza.

VALOR ESPERADO: es la medida de tendencia central de una VA, constituye el primer momento estadístico de la distribución de probabilidad. Es también conocido como ***media poblacional.***

$$E(X) = \sum_{i=1}^{n} X_i f(X_i) = \mu_X$$

[1] Un distribución de probabilidad es el conjunto de los distintos valores que puede tomar una variable aleatoria.

Propiedades:

1) $E(b) = b$, b es una constante.
2) $E(aX) = aE(X)$, a es una constante.
3) $E(X \pm Y) = E(X) \pm E(Y)$
4) $E(aX + b) = aE(X) + b$, a y b son constantes.
5) $E(XY) = E(X)E(Y)$, (X,Y) son VA.
6) $E(X/Y) \neq E(X)/E(Y)$

VARIANZA: es el segundo momento estadístico de una variable. Mide la distribución o dispersión de los valores de la variable alrededor de la media.

$$
\begin{aligned}
Var(X) = \sigma^2 &= E(X - \mu)^2 \\
&= E(X^2) - \mu^2 \\
&= E(X^2) - [E(X)]^2
\end{aligned}
$$

La raíz cuadrada de la varianza es ***desviación estándar***, quien indica que tan cercanos o dispersos están los valores individuales de la variable respecto a su media.

Propiedades:

1) $Var(a)=0$, siendo a una constante.
2) $Var(X + b) = Var(X)$ para b constante.

3) $Var\,(aX + b) = a^2 Var(X)$, a y b son constantes.

4) $Var(X+Y) = Var\,(X) + Var(Y)$.

5) $Var(X\text{-}Y) = Var(X) + Var(Y)$.

6) $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$

Para X *y* Y VA independientes.

COVARIANZA: mide la dirección del movimiento conjunto de dos VA X y Y

$$Cov(X,Y) = E\{(X - \mu_X)(Y - \mu_Y)\} = E(XY) - \mu_X\mu_Y$$

<u>Propiedades:</u>

1) $Cov(X,Y) = E(XY) - \mu_X\mu_Y$
$= \mu_X\mu_Y - \mu_X\mu_Y$
$= 0$ cuando X *y* Y son independientes.

2) $Cov(a + bX, c + dY) = bdCov(X,Y)$. donde a, b, c y d son constantes.

COEFICIENTE DE CORRELACIÓN: mide el grado de asociación lineal entre dos variables aleatorias.

$$\rho = \frac{Cov(X,Y)}{\sqrt{\{Var(X)Var(Y)\}}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- ✓ Se encuentra entre $[-1\ ;1]$
- ✓ Tiene el mismo signo de la covarianza.
- ✓ No depende de las unidades de medida de las variables.

VARIANZA DE VARIABLES CORRELACIONADAS:

- ✓ $Var\ (X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
- ✓ $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X,Y)$

MEDIDAS ESTADISTICAS BÁSICAS

1) MEDIDAS DE TENDENCIA CENTRAL:

- ✓ ***Moda:*** es el valor de la variable que más se repite en una población o muestra.

$$Mo = Max\{f_i\}$$

- ✓ ***Media aritmética:*** es una medida afectada por los valores extremos de la distribución. Representa el centro de gravedad de la distribución de frecuencias, considerado el valor característico en estudio que tendrían todos los elementos de la población o muestra si no existiese diferencia entre ellos.

$$\bar{X} = \frac{\sum X_i}{n}$$

2) MEDIDAS DE DISPERSIÓN O VARIABILIDAD:

- ✓ ***Rango:*** determina la dispersión de los datos de una distribución de frecuencias y corresponde a la diferencia entre el mayor valor de los datos (máximo) y el menor (mínimo).

$$R = Max(X_i) - Min(X_i)$$

- ✓ ***Rango intercuartil:*** proporciona una indicación de la variabilidad de un conjunto de datos, es rango contenido entre el primer y tercer cuartil de la distribución.

RIC= (Q3 – Q1)

- ✓ ***Varianza:*** es la suma de los cuadrados de las desviaciones de los valores de la variable respecto a su media aritmética. Es mayor o igual que cero. Es la medida de dispersión absoluta.

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

- ✓ ***Desviación estándar:*** definida como la raíz cuadrada positiva de la varianza, se expresa en la misma unidad de medida de la variable.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

- ✓ ***Coeficiente de variación:*** es una medida de dispersión relativa que no depende de la unidad de medida de las variables.

$$\mathrm{CV} = \frac{\mathrm{S}}{\bar{\bar{\mathrm{X}}}}$$

3) MEDIDAS DE LOCALIZACIÓN:

- ✓ ***Mínimo:*** menor valor que toma la variable en la distribución.

$$X_1 = Min(X_i)$$

- ✓ ***Máximo:*** mayor valor que toma la variable en la distribución.

$$X_n = Max(X_i)$$

- ✓ ***Mediana:*** es aquel dato central que divide la muestra en dos partes iguales. Identifica la tendencia central de la muestra sin que se vea afectado por los valores extremos.

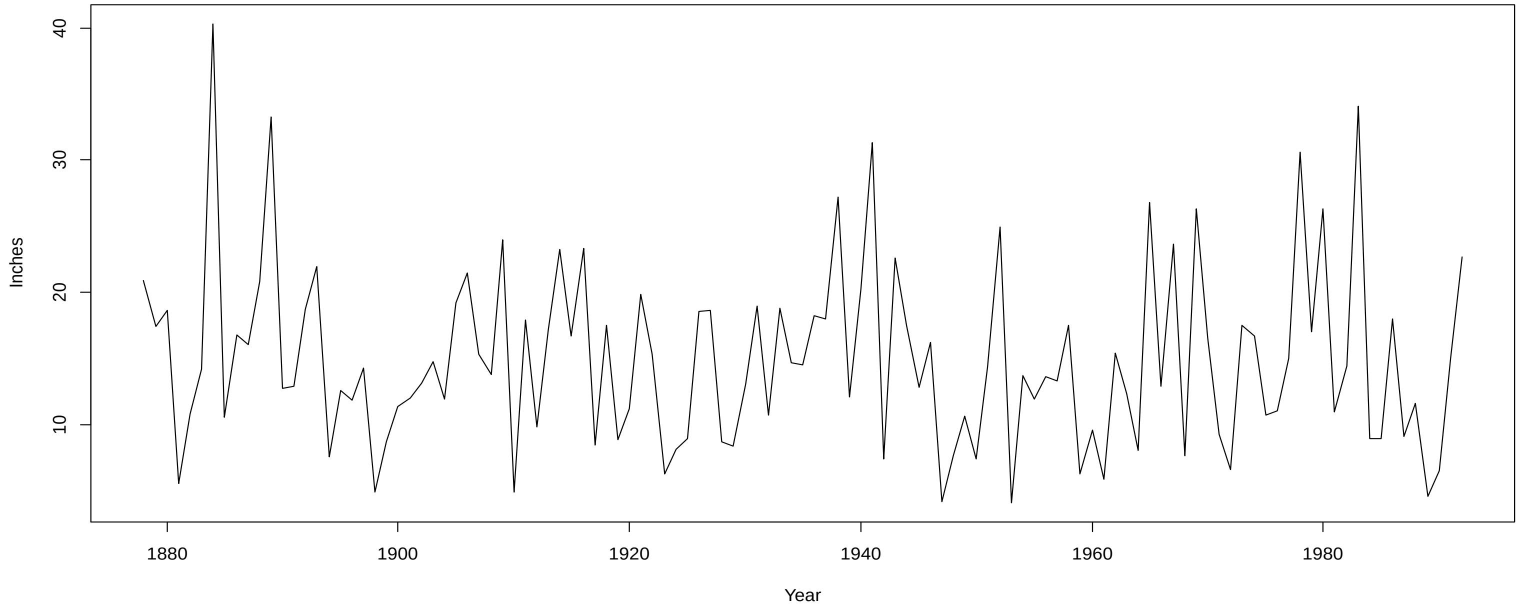
$Md = X_{(m)}$ $\qquad m = \frac{(n+1)}{2};\ \ para\ n\ impar$

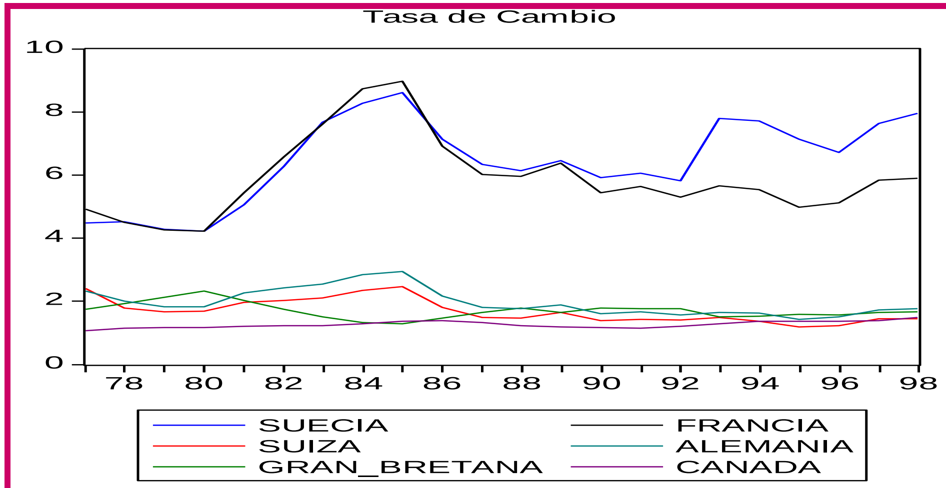
$$Md = \frac{\left[X_{(m1)} + X_{(m2)}\right]}{2}\ ;\ m1 = \frac{n}{2}\ y\, m2 = m1 + 1;\ \ para\, n\, par$$

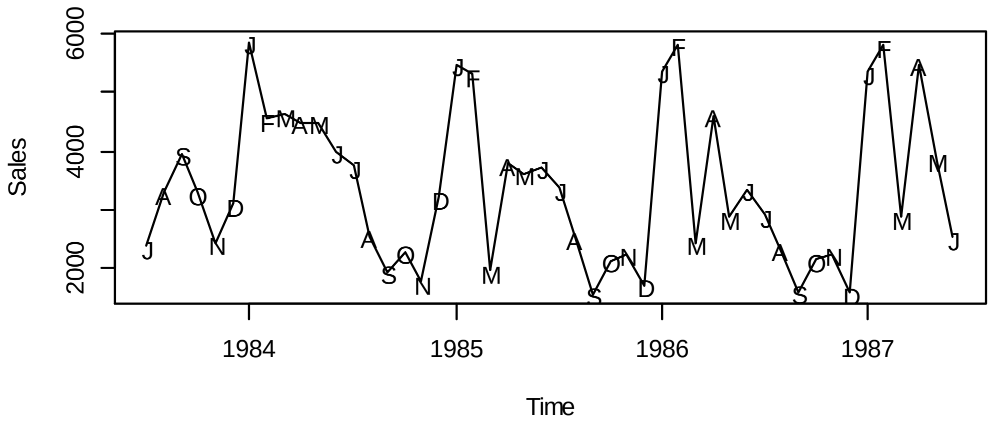
- ✓ ***Cuartil:*** al igual que los deciles y los percentiles es una medida de localización similar a la mediana, que subdivide un conjunto de mediciones de acuerdo con la proporción de frecuencias. Los ***cuartiles (Qi***) dividen la distribución en cuatro partes iguales; los ***deciles (Di)*** lo hacen en diez partes iguales y los ***percentiles (Pi)*** dividen la distribución en cien partes iguales.

GRÁFICOS:

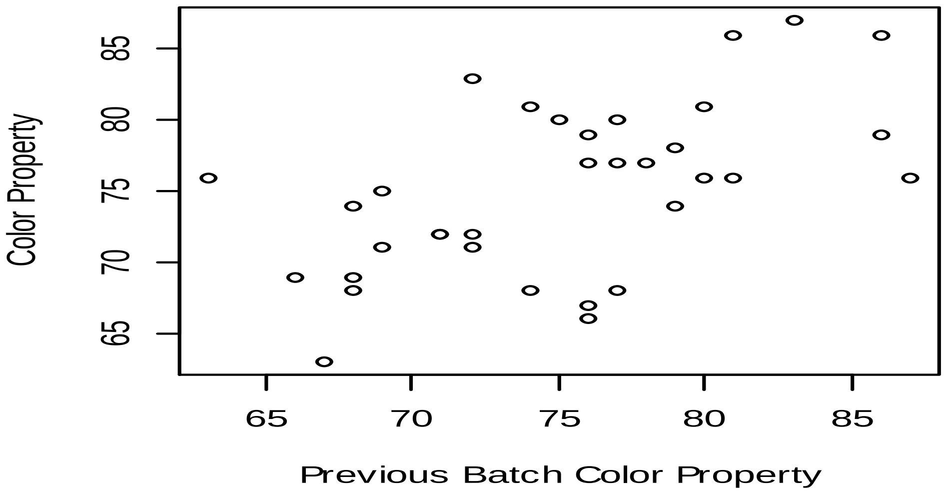
- ✓ De línea:



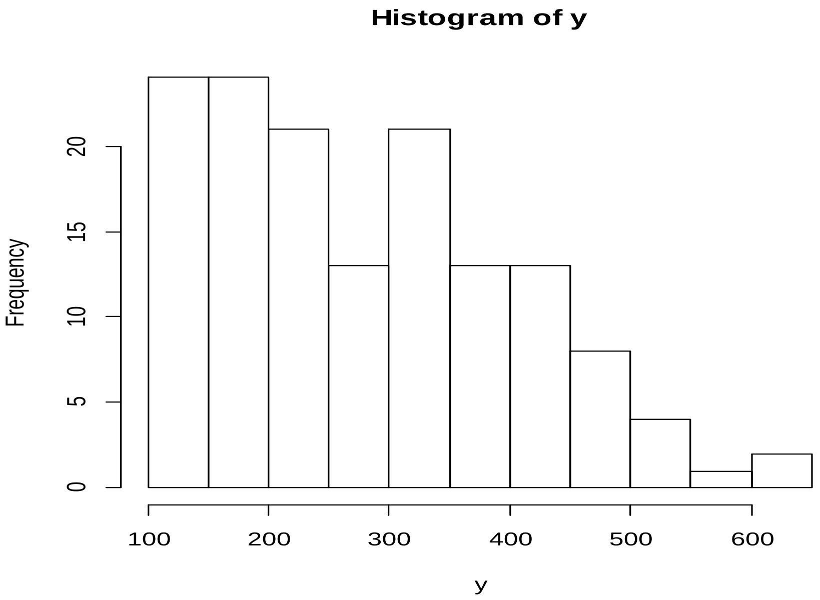


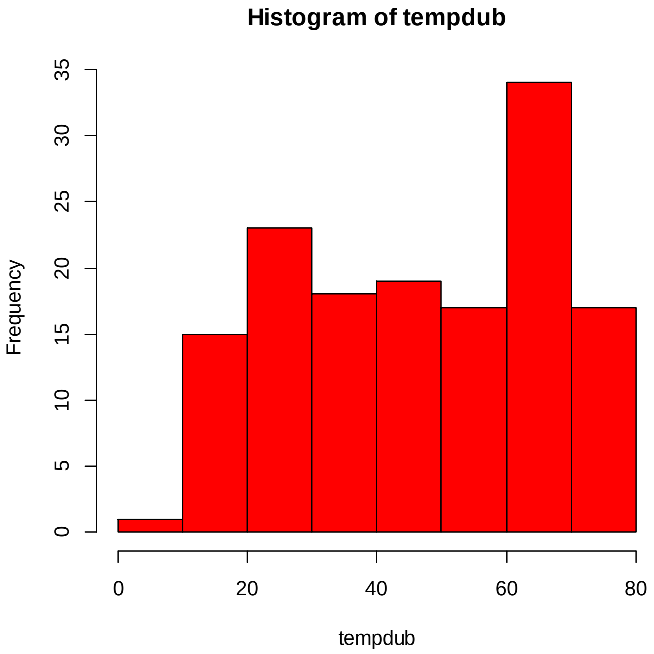


✓ De dispersión:



✓ Histograma: condensa los datos al agrupar valores similares en clases.





ELEMENTOS DE INFERENCIA ESTADÍSTICA

PRUEBA DE HIPÓTESIS

Una prueba de hipótesis no es otra cosa que un procedimiento mediante el cual se rechaza o no la hipótesis nula. Lo que se busca es saber si es compatible o incompatible una observación con alguna o algunas hipótesis planteadas.

¿QUÉ ES UNA HIPÓTESIS?

R: ES UN ESTAMENTO O DECLARACIÓN QUE SE REALIZA ACERCA DE UN PARÁMETRO QUE SE DESEA PROBAR

En estadística la hipótesis planteada se denomina **HIPÓTESIS NULA** $(\boldsymbol{H_0})$ y frente a ésta se plantea la **HIPÓTESIS ALTERNATIVA o MANTENIDA** $(\boldsymbol{H_1})$**.** La prueba de hipótesis puede ser simple o compuesta. La teoría estadística en materia de hipótesis se encarga de diseñar reglas que permitan decidir si rechazar o no una hipótesis nula, para ello existen dos métodos mutuamente complementarios: el intervalo de confianza y la prueba de significancia.

$\boldsymbol{H_0}: \beta_i = 1$

$\boldsymbol{H_1}: \beta_i > (<) 1 \longrightarrow$ una cola

$\beta_i \neq 1 \longrightarrow$ dos colas

PASOS PARA REALIZAR UNA PRUEBA DE HIPÓTESIS

PASO 1: Formular la hipótesis que se ha de probar, $\boldsymbol{H_0}$ y establecer la hipótesis alternativa $\boldsymbol{H_1}$.

PASO 2: recopilar una muestra aleatoria de artículos de la población, medirlos y calcular la estadística adecuada de la prueba para la muestra.

PASO 3: suponer que la $\boldsymbol{H_0}$ es verdadera y determinar la distribución muestral de la estadística de la prueba.

PASO 4: calcular la probabilidad de que el valor de la estadística de la muestra sea, por lo menos, tan grande como el que se observó.

PASO 5: si esta probabilidad es alta no se rechaza la hipótesis nula; si esta probabilidad es baja, la hipótesis nula se desacredita y puede rechazarse con una pequeña probalidad de error.

	ACCIÓN	
	No rechazar $\boldsymbol{H_0}$	Rechazar $\boldsymbol{H_0}$
$\boldsymbol{H_0}$ verdadera	DECISIÓN CORRECTA	ERROR TIPO I: PROBABILIDAD α
$\boldsymbol{H_0}$ falsa	ERROR TIPO II. PROBABILIDAD β	DECISIÓN CORRECTA

ERROR TIPO I: CONSISTE EN RECHAZAR UNA HIPÓTESIS VERDADERA.

ERROR TIPO II: CONSISTE EN NO RECHAZAR UNA HIPÓTESIS FALSA.

VALOR CRÍTICO:

•El valor crítico es un valor tabulado determinado por la distribución de probabilidad y el riesgo de error tipo I: α=P(E_I).

•El riesgo de error α indica la probabilidad de que ese valor ocurra naturalmente. Un α=5%, significa que el valor crítico va a ser excedido por un evento aleatorio, menos del 5% de las veces. Se interpreta como la probabilidad asumida de error tipo I. El p-valor es el verdadero nivel de significación.

Ejemplo:

Suponga que se desea probar la hipótesis que el puntaje medio de los estudiantes en un examen nacional es de 500, contra la hipótesis alternativa de que es menor de 500. Se toma una muestra aleatoria de 15 estudiantes de entre la población y produce un puntaje medio de 475 en tal muestra. La desviación estándar de la población se estima mediante la desviación estándar de la muestra, *S*=35. Suponga que la población de los exámenes se distribuye normalmente:

$$\boldsymbol{H_0}: \boldsymbol{\mu} = 500$$
$$\boldsymbol{H_1}: \boldsymbol{\mu} < 500$$

n= 15
df= n-1 = 14
α = 5%

Regla de decisión: SI tc > $t_{tabulado}$ se rechaza $\boldsymbol{H_0}$

$t_{tabulado} = -1,76$ con un Pvalue de 0.0035

¿Qué significa el Pvalue y cómo se calcula?

$$t_c = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{475-500}{35/\sqrt{15}} = -2.77$$

(Más adelante veremos cómo decidir que distribución de probabilidad utilizar)

Conclusión: se rechaza la hipótesis nula, por tanto el promedio de notas a nivel nacional es menor a 500.

¿CÓMO SE CALCULA EL P-VALUE?

En un sentido amplio el p-value es una medida de la “credibilidad” de la hipótesis nula. Cuanto más pequeño es el valor p, menos probable es que H_0 sea verdadera y por ello, si es menor que el nivel de significación, H_0 se rechaza.

Suponiendo H_0 verdadera:

a) En un contraste unilateral derecho, el p-value corresponde a la probabilidad que el estadístico del contraste tome valores iguales o mayores que su valor calculado con datos experimentales.

b) En un contraste unilateral izquierdo, el p-value corresponde a la probabilidad que el estadístico del contraste tome valores iguales o menores que su valor calculado con datos experimentales.

c) En un contraste bilateral, el p-value se calcula como la suma de los p valores suponiendo una prueba unilateral derecha y una izquierda.

La regla de decisión para rechazar H_0, con el enfoque del valor p, es:

- Si el p-value > α, no se rechaza H_0.
- Si el p-value < α, se rechaza H_0.

Ejemplo:

Usando el enfoque del p-value, para una prueba bilateral, se desea encontrar la probabilidad de obtener un estadístico Z igual o más grande que 2,66 unidades de desviación estándar respecto del centro de la Distribución Normal Estándar. Esto significa que se debe calcular la probabilidad de obtener un valor Z > 2,66, como así también la probabilidad de obtener un valor menor que -2,66. De la tabla de los Cuantiles de la Distribución Normal Estándar, se obtienen las siguientes probabilidades: P (Z < −2,66) = P (Z > 2,66) = 0,0039, el p-value es 0,0039 + 0,0039= 0,0078.

En el ejemplo de referente a las notas promedio, se tiene:

$$t_c = -2.77$$

Pvalue: P(t > -2.77) equivalente a P(t > 2.77)

Pvalue = 1- P(t > 2.77) = 1- 0.98= 0.0200

Lecturas obligatorias:

- ✓ Gujarati, D. y Porter, D. (2010). Econometría. 5ta. Edición McGraw Hill. Apéndice s A y B.

- ✓ Wooldridge, J. (2010). Introducción a la econometría. Un enfoque moderno. Apéndices A, B y C.

REQUISITO OBLIGATORIO:

REPASO FUNDAMENTOS DE INFERENCIA ESTADÍSTICA (ESTADÍSTICA I Y II)

REPASAR LOS FUNDAMENTSO BÁSICOS DE ALGEBRA MATRICIAL (MATEMÁTICAS 31)