

UNIVERSIDAD PANAMERICANA

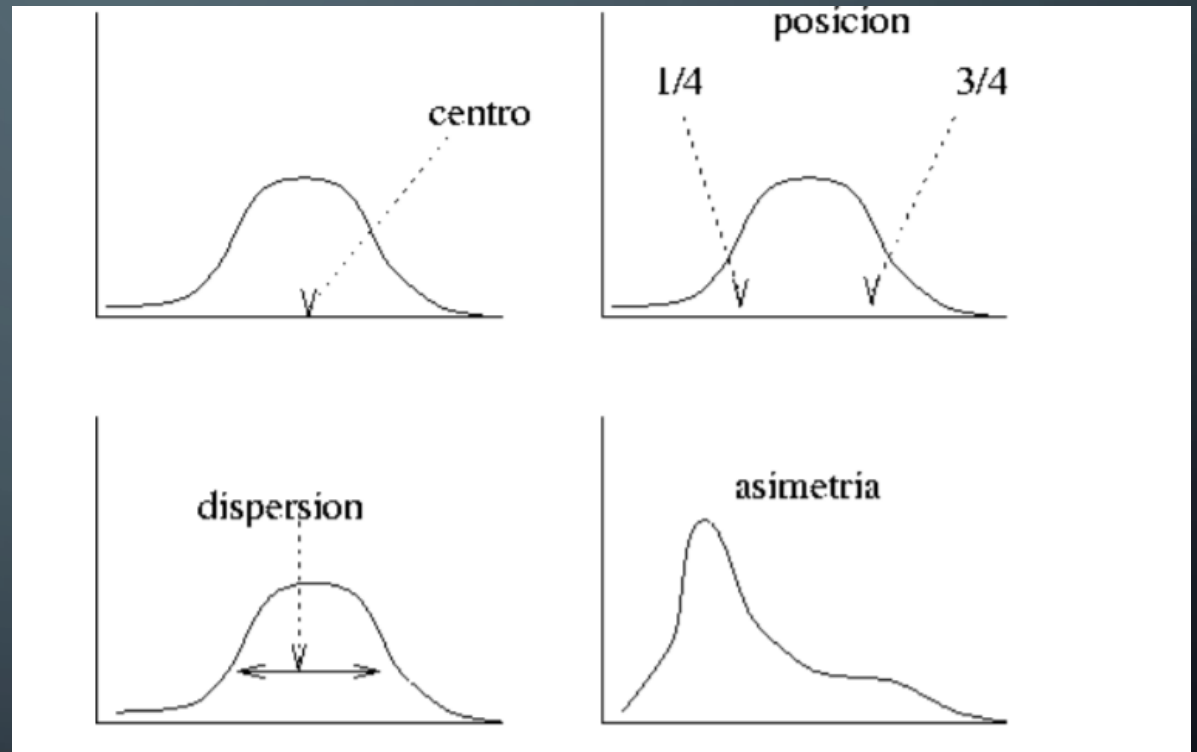
FACULTAD DE INGENIERÍA

MAESTRÍA EN CIENCIA DE DATOS

21 DE ABRIL DE 2020

MEDIDAS DESCRIPTIVAS

- La tendencia central de los datos,
- La dispersión o variación con respecto a este centro,
- Los datos que ocupan ciertas posiciones,
- La simetría de los datos y
- La forma en la que los datos se agrupan.



- Media aritmética
- Media geométrica
- Media armónica
- Media cuadrática

MEDIDAS DE CENTRALIZACIÓN

Media aritmética

La media aritmética de una variable estadística es la suma de todos sus posibles valores, ponderada por las frecuencias de los mismos.

Sea X una variable aleatoria con distribución de probabilidad $f(x)$. La **media** o **valor esperado** de X es

$$\mu = E(X) = \sum_x xf(x)$$

si X es discreta, y

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

si X es continua.

Si la variable es continua y se utilizan intervalos, se cambiarán los valores de x_i por las marcas de clase correspondientes c_i .

En general, la media aritmética obtenida a partir de las marcas de clase c_i , diferirá de la media obtenida con los valores reales, x_i

Habrà una pérdida de precisión que será tanto mayor cuanto mayor sea la diferencia entre los valores reales y las marcas de clase, esto es, cuanto mayores sean las longitudes de los intervalos.

La suma de las diferencias de la variable con respecto a la media es nula.

Ejemplo

Obtener las desviaciones con **respecto a la media** en la siguiente distribución y comprobar que su suma es cero.

lim_{i-1}	lim_{ii}	n_i
0-10		1
10-20		2
20-30		4
30-40		3
		n=10

Obtener las desviaciones con respecto a la media en la siguiente distribución y comprobar que su suma es cero.

$lim_{i-1} - lim_i$	n_i	x_i	$x_i n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})n_i$
0-10	1	5	5	-19	-19
10-20	2	15	30	-9	-18
20-30	4	25	100	1	4
30-40	3	35	105	11	33
	$n = 10$		$\sum_{i=1}^k x_i n_i = 240$		$\sum_{i=1}^k (x_i - \bar{x}) n_i = 0$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i = 24$$

No tiene sentido su cálculo en variables de tipo cualitativo o nominal

INCONVENIENTES DE LA MEDIA:

- Es muy **sensible a los valores extremos de la variable**, todas las observaciones intervienen en el cálculo de la media, de manera que la aparición de una observación extrema hará que la media se desplace en esa dirección.
- **No es recomendable usar la media como medida central en las distribuciones muy asimétricas.**
- Depende de la división en intervalos en el caso de utilizar tablas estadísticas.
- Si se considera **una variable discreta el valor de la media puede no pertenecer al conjunto de posibles valores que pueda tomar la variable.**

MEDIA GEOMÉTRICA

$$\bar{x} = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Ventajas

- Considera todos los valores de la distribución
- Es **menos sensible que la media aritmética a los valores extremos.**

Desventajas

- Es de significado estadístico menos intuitivo que la media aritmética.
- Su cálculo es más difícil.
- Si un valor de $x_i=0$, entonces la media geométrica se anula o no queda determinada.

Sólo es relevante la media geométrica si todos los números son positivos.

Si hubiera un número negativo (o una cantidad impar de ellos) entonces la media geométrica sería negativa o inexistente en los números reales.

En muchas ocasiones se utiliza su transformación en el manejo estadístico de variables con distribución no normal.

La media geométrica es relevante cuando varias cantidades son multiplicadas para producir un total. Esto es, cuando una variable presenta variaciones acumulativas

$$\bar{x} = \sqrt[n]{x_1 x_2 x_3 \dots x_n}$$

$$\log(\bar{x}) = \log(x_1 x_2 x_3 \dots x_n)^{1/n}$$

$$\log(x) = \frac{\log(x_1 x_2 x_3 \dots x_n)}{n}$$

MEDIA ARMÓNICA

Se define como la inversa de la media de las inversas de las observaciones. Se usa cuando se promedian variables como productividades, velocidades o rendimientos

Ventaja

- Considera todos los valores de la distribución y en ciertos casos, es más representativa que la media aritmética.

Desventajas

- La influencia de los valores pequeños y el hecho de que no pueda ser determinada en distribuciones con valores iguales a cero.

Suele usarse para promediar velocidades, tiempos, rendimientos, etc.

$$\overline{x_A} = \left(\frac{\frac{1}{x_1} + \dots + \frac{1}{x_n}}{n} \right)^{-1}$$

$$\overline{x_A} = \left(\frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}} \right)$$

MEDIA CUADRÁTICA

Es la raíz cuadrada de la media aritmética de los cuadrados de las observaciones.

Es útil para calcular la **media** de variables **que toman valores negativos y positivos y se desea obtener un promedio que no recoja los efectos del signo.**

Se suele utilizar cuando el símbolo de la variable no es importante y lo **que** interesa es el valor absoluto del elemento.

$$\overline{x_c} = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

Puede calcularse **para una serie de valores discretos o para una función matemática de variable continua.**

Ejemplo de uso para calcular la media de errores de medida.

RELACIÓN ENTRE MEDIAS

Existe una relación de orden entre cuatro tipos de media.

Sean:

H la media armónica

MG la media geométrica

\bar{x} la media aritmética

RMS la media cuadrática

$$H \leq MG \leq \bar{x} \leq RMS$$

Sólo se cumple la **igualdad** cuando todos los datos son iguales

MEDIANA

Se considera una variable X cuyas observaciones han sido ordenadas de menor a mayor.

La mediana, Me , es el primer valor de la variable que deja por debajo de sí al 50 % de las observaciones.

Si n es el número de observaciones, la mediana corresponderá a la observación que ocupa la posición $[n/2] + 1$ (donde $[]$ es la parte entera de un número).

Si el número de datos es impar,

$$Me = n/2$$

Si es par, los dos datos que están en el centro de la muestra ocupan las posiciones $n/2$ y $(n/2)+1$

$$Me = \frac{\left(\frac{n}{2} + \left(\frac{n}{2} + 1\right)\right)}{2}$$

PROPIEDADES DE LA MEDIANA

- Como medida descriptiva, tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas.
- Es adecuado su uso en distribuciones asimétricas.
- Es de cálculo rápido y de interpretación sencilla, pero no tiene sentido su cálculo en variables de tipo cualitativo o nominal, al igual que la media.
- A diferencia de la media, la mediana de una variable es siempre un valor de la variable que se estudia.

MEDIANA PARA VARIABLES CONTINUAS

Si la mediana se encuentra en un intervalo dado $[l_{i-1}, l_i]$ y hay que determinar el punto que deja exactamente la mitad de observaciones a un lado y al otro.

$$\frac{N_i - N_{i-1}}{a_i - a_{i-1}} = \frac{\frac{n}{2} - N_{i-1}}{p} \Rightarrow p = \frac{\frac{n}{2} - N_{i-1}}{N_i - N_{i-1}} (a_i - a_{i-1})$$

N_{i-1} y N_i son las frecuencias acumuladas

$$N_{i-1} < \frac{n}{2} < N_i$$

$$Me = l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i$$

Donde:

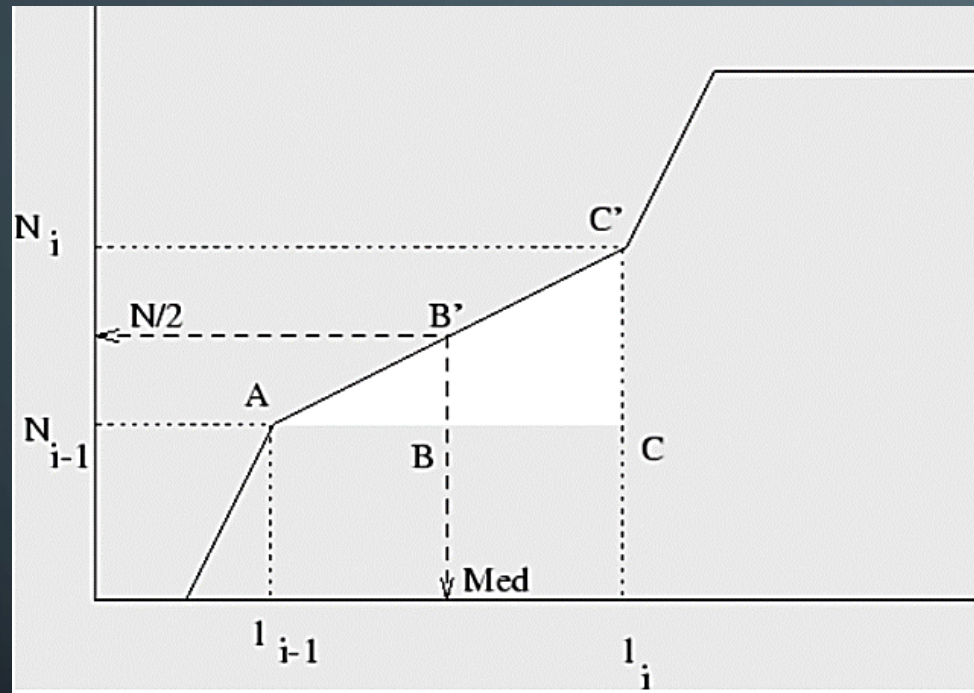
l_{i-1} es el extremo inferior del intervalo donde se encuentra el valor de la mediana,

n es el tamaño total de la muestra,

n_i es la frecuencia absoluta que aparece en el intervalo donde se encuentra el valor de la mediana,

a_i es la amplitud de dicho intervalo.

Si se utiliza interpolación lineal, aplicando el teorema de Thales:



$$\begin{aligned} \frac{CC'}{AC} &= \frac{BB'}{AB} \Rightarrow \\ \frac{n_i}{a_i} &= \frac{\frac{n}{2} - N_{i-1}}{Me - l_{i-1}} \Rightarrow \\ Me &= l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i \end{aligned}$$

MODA

Es el valor con mayor frecuencia en una de las distribuciones de datos.

Propiedades:

- Es muy fácil de calcular,
- Se puede considerar para variables tanto cuantitativas como cualitativas.
- Puede no ser única.

ESTADÍSTICOS DE POSICIÓN

Si las variables son continuas y están agrupadas en intervalos, el cálculo de estos estadísticos de posición se basa en que:

Se supone que el valor se encuentra en un intervalo dado $[l_{i-1}, l_i]$

Se busca determinar el punto que deja exactamente el porcentaje correspondiente de observaciones a un lado y al otro.

Percentil

Para una variable discreta, se define el percentil de orden k , como la observación, P_k , que deja por debajo de sí el k % de la muestra.

El percentil puede estar situado en cualquier lugar de la distribución, por lo que no se puede considerar como una medida de tendencia central, sino de posición.

Los cuartiles, Q_i , son un caso particular de los percentiles.

Existen 3:

Primer Cuartil: $Q_1 = P_{25}$

Segundo Cuartil: $Q_2 = P_{50}$ (equivalente a la Mediana)

Tercer Cuartil: $Q_3 = P_{75}$

- 1er cuartil (Q_1)
25% de los datos es menor o igual a este valor.
- 2do cuartil (Q_2)
50% de los datos es menor o igual a este valor.
- 3er cuartil (Q_3)
75% de los datos es menor o igual a este valor.

EJEMPLO

Dada la siguiente distribución en el número de hijos de cien familias, calcular sus cuartiles (en R).

x_i	n_i	N_i
0	14	14
1	10	24
2	15	39
3	26	65
4	20	85
5	15	100
	$n = 100$	

`quantile(x)`

`IQR(x)`

`barplot(quantile(x))`

MEDIDAS DE VARIABILIDAD O DISPERSIÓN

Los estadísticos de tendencia central o posición indican dónde se sitúa un grupo de puntuaciones.

Los de variabilidad o dispersión indican si esas puntuaciones o valores están próximos entre sí o si están dispersos.

Amplitud o rango

Se obtiene restando el valor más bajo de un conjunto de observaciones del valor más alto.

- Inconvenientes:
 - Sólo utiliza de dos observaciones.
 - Se puede ver muy afectada por alguna observación extrema.
 - El rango aumenta con el número de observaciones, o bien se queda igual.
 - En cualquier caso nunca disminuye.

DESVIACIÓN MEDIA

Es la media de las diferencias en valor absoluto de los valores de la variable a la media, es decir, si tenemos un conjunto de n observaciones, x_1, \dots, x_n , entonces

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

Esto es, es la media de las desviaciones absolutas y es un resumen de la dispersión estadística.

La desviación media conserva las mismas dimensiones que las observaciones.

Tiene la limitante de estar calculada con respecto a la media.

Es muy general, su uso es limitado

RANGO INTERCUARTÍLICO

Se define como la diferencia entre el tercer y el primer cuartil $IQR = Q3 - Q1$

Abarca el 50% central de los datos.

Debido a que no son afectados por observaciones extremas, la mediana y el rango intercuartil constituyen una buena medida de la tendencia central y la dispersión de conjuntos de datos altamente asimétricos, en comparación con la media y la desviación estándar.

VARIANZA

La varianza, σ^2 , se define como la media de las diferencias cuadráticas de n puntuaciones con respecto a su media aritmética

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 .$$

$$\begin{aligned} \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \underbrace{\frac{1}{n} \sum_{i=1}^n x_i}_1 + \frac{1}{n} n\bar{x}^2 = \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 . \end{aligned}$$

La varianza no tiene la misma magnitud que las observaciones

PARA DATOS AGRUPADOS

$$\sigma^2 = \frac{\sum_{i=1}^n x_i^2 f_i}{N} - \bar{x}^2 = \frac{x_1^2 f_1 + x_2^2 f_2 + \dots + x_n^2 f_n}{N} - \bar{x}^2$$

Si se pretende calcular de **modo aproximado la varianza de una población** a partir de la varianza de una muestra, el error cometido es generalmente más pequeño, si en vez de considerar como estimación de la varianza de la población, a la varianza muestral se considera la **cuasivarianza muestral**, s^2 :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n\sigma^2}{n-1}.$$

Para valores de n grandes, pueden considerarse válidas ambas

DESVIACIÓN ESTÁNDAR

Si se busca que la **medida de dispersión tenga las mismas dimensiones que las observaciones** bastará con tomar su raíz cuadrada. Por ello se define la desviación estándar, σ , como:

$$\sigma = +\sqrt{\sigma^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2}.$$

Para datos agrupados

$$\sigma = \sqrt{\frac{\sigma_1^2 k_1 + \sigma_2^2 k_2 + \dots + \sigma_n^2 k_n}{k_1 + k_2 + \dots + k_n}}$$

Tipificación

Es el proceso de restar la media y dividir entre su desviación típica a una variable X .

De este modo se obtiene una nueva variable $z = \frac{X - \bar{x}}{\sigma}$, de media 0 y desviación estándar $\sigma_z = 1$, que se denomina variable tipificada.

$$z = \frac{X - \bar{x}}{\sigma}$$

Esta nueva variable carece de unidades y permite hacer comparables dos medidas que en un principio no lo son, por aludir a conceptos diferentes.

También es aplicable al caso en que se quieran comparar individuos semejantes de poblaciones diferentes.

Ejemplo

Si se desea comparar el nivel académico de dos estudiantes de diferentes Universidades para otorgar una beca de estudios, en principio sería injusto concederla directamente al que posea una nota media más elevada, ya que la dificultad para conseguir una buena calificación puede ser mucho mayor en un centro que en el otro, lo que limita las posibilidades de uno de los estudiante y favorece al otro.

En este caso, lo más correcto es comparar las calificaciones de ambos estudiantes, pero tipificadas cada una de ellas por las medias y desviaciones típicas respectivas de las notas de los alumnos de cada Universidad

PROPIEDADES DE LA VARIANZA Y LA DESVIACIÓN ESTÁNDAR

Ambas son sensibles a la variación de cada una de las puntuaciones, es decir, si una puntuación cambia, cambia con ella la varianza ya que es función de cada una de las puntuaciones.

La desviación estándar tiene la propiedad de que en el intervalo $(x - 2\sigma; x + 2\sigma)$ se encuentran por lo menos el 75 % de las observaciones (es el llamado teorema de Chebyshev).

Incluso si se tienen muchos datos y estos provienen de una distribución normal, podremos llegar al 95 % de las observaciones.

No es recomendable el uso de ellas, cuando tampoco lo sea el de la media como medida de tendencia central, por ejemplo, en datos nominales.

COEFICIENTE DE VARIACIÓN

En el caso de que se busque la relación entre el tamaño de la media y la variabilidad de la variable, se aplica en coeficiente adimensional CV.

$$CV = \frac{\sigma}{\bar{x}} \cdot 100.$$

expresa la desviación estándar como porcentaje de la media aritmética, mostrando una interpretación *relativa* del grado de variabilidad, independiente de la escala de la variable, a diferencia de la desviación típica o estándar.

Es muy sensible ante cambios de origen en la variable,

Es importante que todos los valores sean positivos y su media sea un valor positivo.

A mayor valor del coeficiente de variación mayor heterogeneidad de los valores de la variable; y a menor CV, mayor homogeneidad en los valores de la variable.

Por ejemplo, si el CV es menor o igual al 80%, significa que la media aritmética es representativa del conjunto de datos, por ende el conjunto de datos es **"Homogéneo"**.

Depende de la desviación estándar, y en mayor medida de la media aritmética, dado que cuando ésta es 0 o muy próxima a este valor el CV pierde significado, ya que puede dar valores muy grandes, que no necesariamente implican una gran dispersión de datos.

Dada la distribución de edades (medidas en años) en un colectivo de 100 personas, obtener:

La variable tipificada Z , los valores de la media y varianza de Z , el coeficiente de variación de Z .

x_i	n_i
2	47
7	32
15	17
30	4
	$n = 100$

xi	ni	Xi*ni	xi ² ni		
2	47	94	188		
7	32	224	1568		
15	17	255	3825		
30	4	120	3600		
	100	693	9181		
x=		6.93			
Varianza=		43.79 años ²			
Desviación st=		6.62			
Valores tipificados		ni	Zi*ni	zi ² ni	
z1=	-0.745	47	-35.017	26.089	
z2=	0.011	32	0.339	0.004	
z3=	1.220	17	20.733	25.285	
z4=	3.486	4	13.946	48.622	
z media		100	0.000	100	
		z media=	0		
		Varianza=	1		
		Desviación=	1		

A	B	C	D	E	F
xi	ni	xini	xi ² ni		
2	47	=A2*B2	=A2^2*B2		
7	32	=A3*B3	=A3^2*B3		
15	17	=A4*B4	=A4^2*B4		
30	4	=A5*B5	=A5^2*B5		
	=SUMA(B2:B5)	=SUMA(C2:C5)	=SUMA(D2:D5)		
x=		=C6/B6			
sigma cuad		=(D6/B6)-C9^2	años ^2		
sigma		=RAIZ(C10)			
Valores tipificados			ni	zini	zi ² ni
z1=	=(A2-\$C\$9)/\$C\$11		47	=C14*D14	=C14^2*D14
z2=	=(A3-\$C\$9)/\$C\$11		32	=C15*D15	=C15^2*D15
z3=	=(A4-\$C\$9)/\$C\$11		17	=C16*D16	=C16^2*D16
z4=	=(A5-\$C\$9)/\$C\$11		4	=C17*D17	=C17^2*D17
z media			=SUMA(D14:D17)	=SUMA(E14:E17)	=SUMA(F14:F17)
			z promedio	=E18/D18	33
			varianza	=F18/D18	
			desviación	=RAIZ(E22)	

EJERCICIO PARA LA CLASE

Para el conjunto de datos CEP.sav

```
install.packages("haven")
```

```
library(haven)
```

```
CEP <- read_spss("c:/Users/anton/Documents/ClasesUP/ESTADISTICA_MCD/CEP2.sav")
```

EJEMPLO: `mean(CEP$DS_P2_EXACTA)`

1. Calcular los estadísticos descriptivos para las variables SV_1 y SV_2 (satisfacción de vida 1 y 2) y DS_P2_EXACTA (edad).
2. Determinar si las variables SV_1 y SV_2 tienen comportamiento normal.



¿Y SI LOS DATOS NO SIGUEN UNA DISTRIBUCIÓN NORMAL?...

- TRANSFORMACIÓN DE VARIABLES
- PRUEBAS NO PARAMÉTRICAS

- Las inferencias en cuanto a las medias son en general robustas, por lo que si el tamaño de muestra es grande, los intervalos de confianza y contrastes basados en la t de *Student* son aproximadamente válidos, con independencia de la verdadera distribución de probabilidad de los datos; pero si ésta distribución no es normal, los resultados de la estimación serán poco precisos.

Pruebas para verificar el ajuste de los datos a una distribución de probabilidad

- Contraste de χ^2 de Pearson, es una prueba no paramétrica que mide la discrepancia entre una distribución observada y otra teórica (bondad de ajuste), indicando en qué medida las diferencias existentes entre ambas, de haberlas, se deben al azar en el contraste de hipótesis.
- Prueba de Shapiro-Wilks, **se recomienda** para contrastar el ajuste de nuestros datos a una distribución normal, sobre todo cuando la muestra es pequeña ($n < 30$).
- ...

- Prueba de Kolmogorov-Smirnov, este contraste es válido únicamente para variables continuas. Compara la función de distribución (probabilidad acumulada) teórica con la observada, y calcula un valor de discrepancia, que corresponde a la diferencia máxima en valor absoluto entre la distribución observada y la distribución teórica, proporcionando un valor de probabilidad P .

1. Si la distribución es más puntiaguda que la normal (mayor parte de los valores agrupados en torno de la media y colas más largas en los extremos), se debe revisar la presencia de heterogeneidad en los datos y de posibles valores atípicos o errores en los datos. La solución puede ser emplear pruebas no paramétricas.
2. Si la distribución es unimodal y asimétrica, la solución más simple y efectiva suele ser utilizar una transformación para convertir los datos en normales.
3. Cuando la distribución **no** es unimodal hay que investigar la presencia de heterogeneidad, ya que en estos casos la utilización de transformaciones no es adecuada y los métodos no paramétricos pueden también no serlo.

- Para hacer más simétrica una distribución se deben hacer transformaciones no lineales:
 - Escoger una transformación que conduzca a una distribución simétrica, y más cercana a la distribución normal

Distribuciones de frecuencias con asimetría negativa (frecuencias altas hacia el lado derecho de la distribución)

- Aplicar la transformación $y = x^2$

Esta transformación comprime la escala para valores pequeños y la expande para valores altos

Distribuciones asimétricas positivas (frecuencias altas hacia el lado izquierdo de la distribución)

- Se usan las transformaciones \sqrt{x} , $\ln(x)$, $\frac{1}{x}$

Comprimen los valores altos y expanden los pequeños.

El efecto de estas transformaciones está en orden creciente:


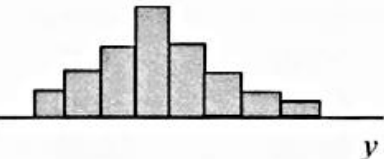
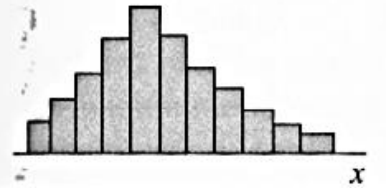

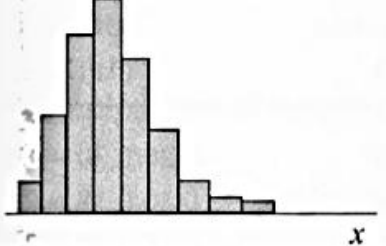
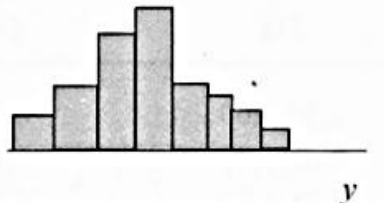
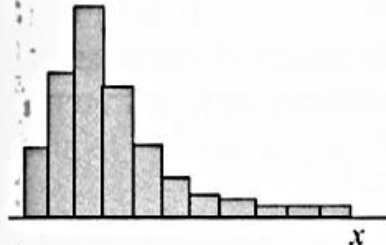
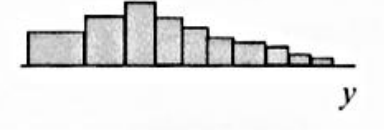
menos efecto \sqrt{x} , mayor efecto $\ln(x)$, más aún $1/x$

Se denominan pruebas no paramétricas las que no presuponen una distribución de probabilidad para los datos, por ello se conocen también como de distribución libre (*distribution free*).

En la mayor parte de ellas los resultados estadísticos se derivan únicamente a partir de procedimientos de ordenación y recuento, de forma que su lógica es de fácil comprensión.

Si se tienen muestras pequeñas ($n \leq 10$) en las que se desconoce si es válido suponer la normalidad de los datos, conviene utilizar pruebas no paramétricas, al menos para corroborar los resultados obtenidos a partir de la utilización de la teoría basada en la normal.

En estos casos se emplea como parámetro de centralización la **mediana**, que es aquel punto para el que el valor de X está el 50% de las veces por debajo y el 50% por encima.

Histograma inicial	Transformación	Histograma transformado
	$y = x^2$	
	$y = \sqrt{x}$	
	$y = \ln x$	
	$y = \frac{1}{x}$	

Las medidas basadas en el orden de los datos, como la mediana o los cuartiles se mantienen iguales cuando se hace una transformación monótona:

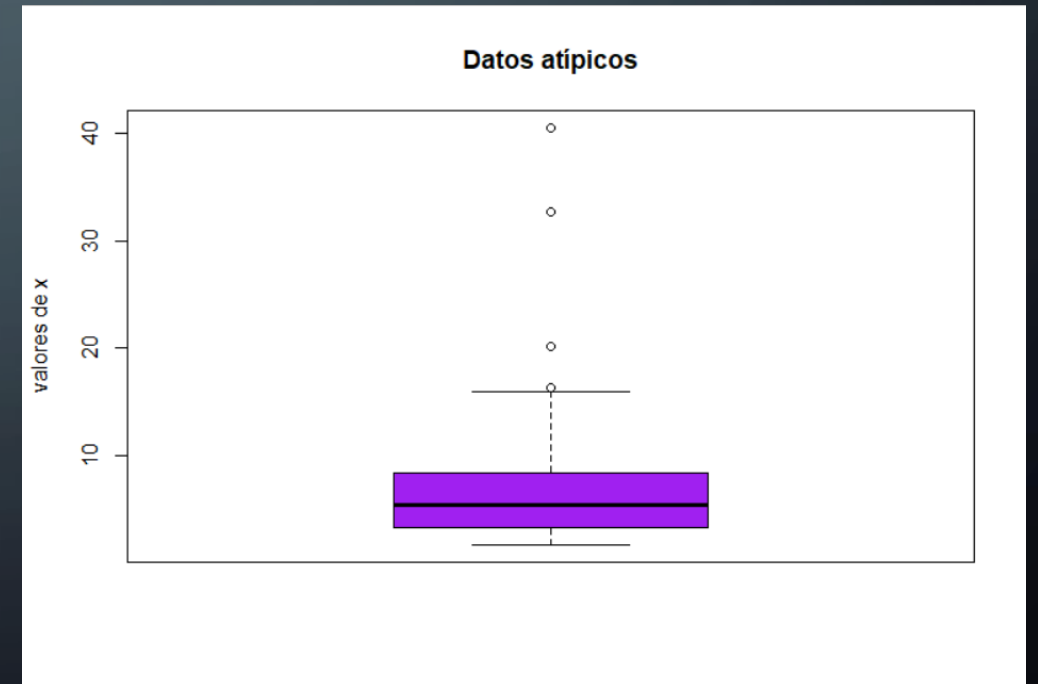
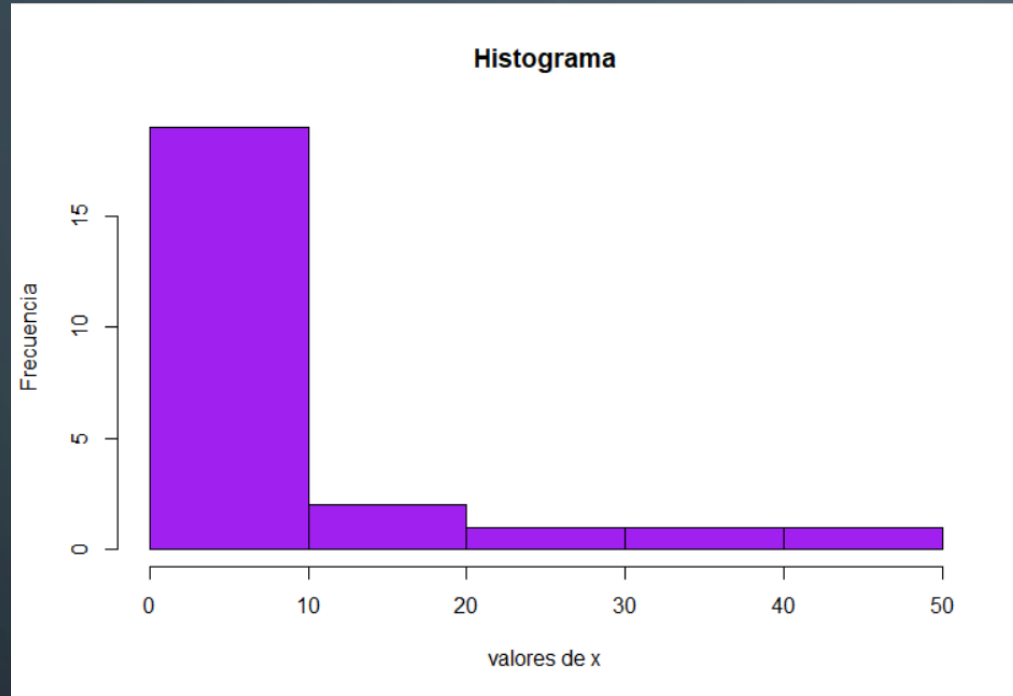
$$x_1 > x_2 \Rightarrow f(x_1) > f(x_2)$$

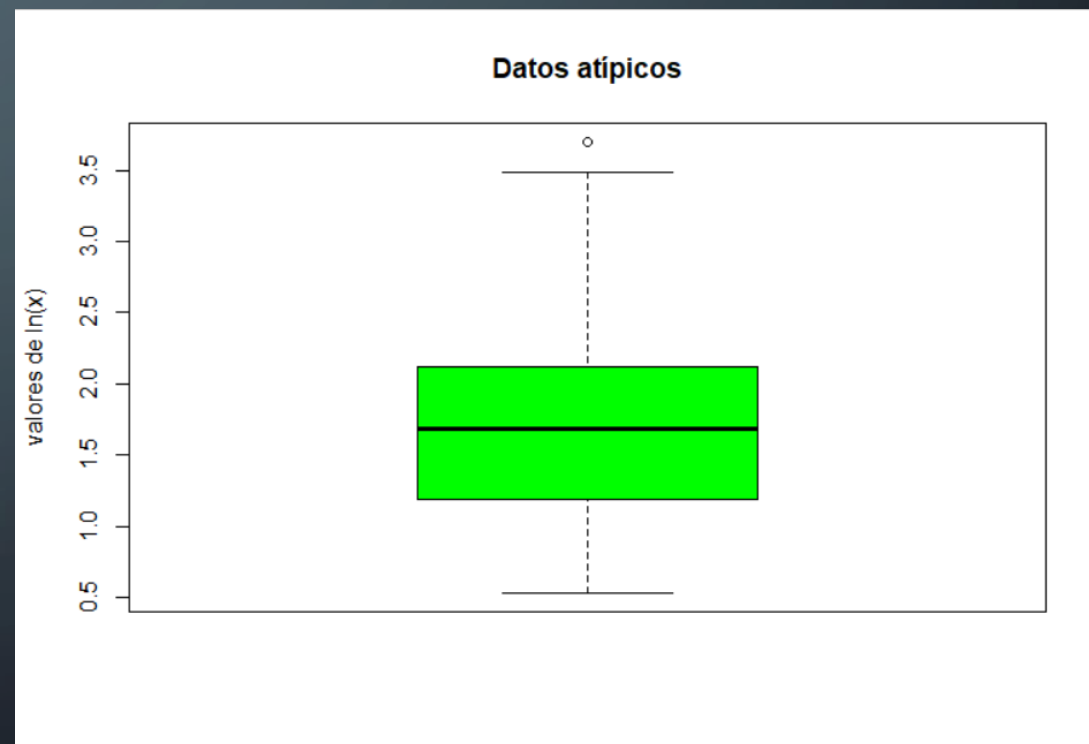
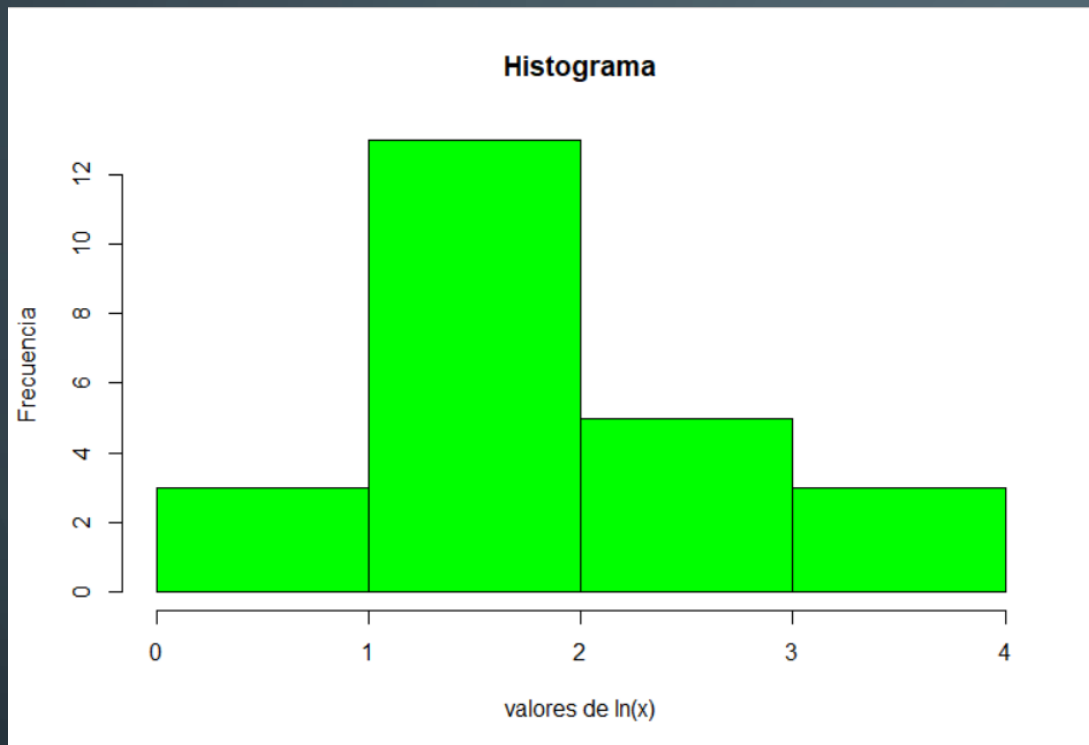
El resto de estadísticos cambia.

Los **cuartiles** son valores que dividen una muestra de datos en cuatro partes iguales.

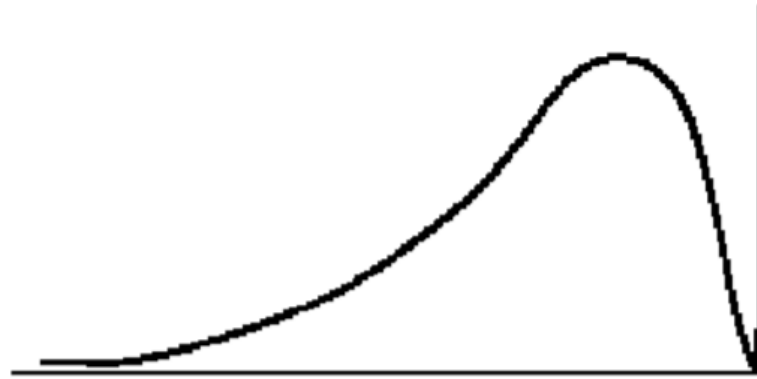
Ejemplo 1:

$x = \{2.2, 7.6, 2.9, 4.6, 4.1, 3.9, 7.4, 3.2, 5.1, 5.3, 20.1, 2.3, 5.5, 32.7, 9.1, 1.7, 3.2, 5.8, 16.3, 15.9, 5.9, 6.7, 3.4, 40.5\}$



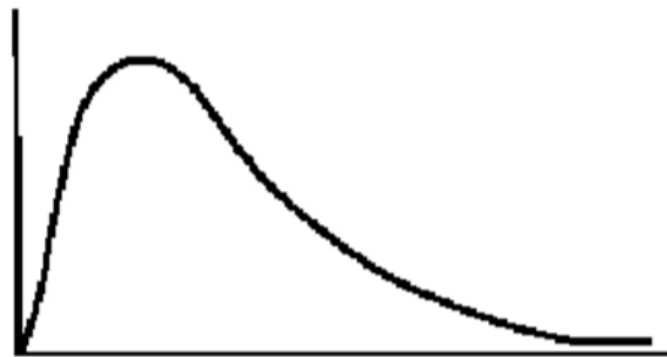


Asimétrica a la izquierda. Es el caso en que $Mo \geq Me \geq \bar{x}$



Curva Asimétrica a la izquierda

Asimétrica a la derecha. Es el caso en que $Mo \leq Me \leq \bar{x}$



Curva Asimétrica a la derecha

$$m_r = \sum_{i=1}^n (x_i - \bar{x})^r \frac{n_i}{N}$$

Los momentos de una muestra forman una sucesión de números, para cada número natural r se puede definir el momento r —ésimo.

Los momentos de una distribución son medidas obtenidas a partir de todos sus datos y de sus frecuencias absolutas.

Estas medidas caracterizan de tal forma a las distribuciones que si los momentos de dos distribuciones son iguales, diremos que las distribuciones son iguales

$$\begin{aligned} m_0 &= \sum_{i=1}^n (x_i - \bar{x})^0 \frac{n_i}{N} = \frac{N}{N} = 1 \\ m_1 &= \sum_{i=1}^n (x_i - \bar{x})^1 \frac{n_i}{N} = \bar{x} - \bar{x} = 0 \\ m_2 &= \sum_{i=1}^n (x_i - \bar{x})^2 \frac{n_i}{N} = S^2 \end{aligned}$$

m_3 y m_4 se usan para calcular asimetría y curtosis

Coeficiente de asimetría de Fischer

$$g_1 = \frac{\frac{\sum_{i=1}^r (x_i - \bar{x})^3 n_i}{N}}{s^3} = \frac{m_3}{s^3}$$

- Si $g_1 > 0$, la distribución es asimétrica positiva o a la derecha.
- Si $g_1 = 0$, la distribución es simétrica.
- Si $g_1 < 0$, la distribución es asimétrica negativa o a la izquierda.

Si una variable es continua simétrica y unimodal, coinciden la media, la mediana y la moda.

En una distribución simétrica los valores se sitúan en torno a la media aritmética de forma simétrica.

El coeficiente de asimetría de Fisher se basa en la relación entre las distancias a la media y la desviación típica.

Coeficiente de asimetría de Pearson

Se basa en que en una distribución simétrica, la media coincide con la moda.

Sólo se puede utilizar en distribuciones uniformes, unimodales y moderadamente asimétricas.

$$A_p = \frac{\bar{x} - Mo}{s}$$

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

Coeficiente de Curtosis de Fischer

El coeficiente de curtosis o apuntamiento de Fischer comparara la curva de una distribución con la curva de la variable Normal, en función de la cantidad de valores extremos de la distribución.

En una distribución normal se verifica que:

$$m_4 = 3s^4;$$

$$\frac{m_4}{s^4} - 3 = 0$$

Por lo que el coeficiente de Curtosis de Fischer, se calcula:

$$K = g_2 = \frac{\sum_{i=1}^r (x_i - \bar{x})^4 n_i}{N s^4} - 3 = \frac{m_4}{s^4} - 3$$

- Si $g_2 = 0$, la distribución es **Mesocúrtica**: Al igual que en la asimetría es bastante difícil encontrar un coeficiente de curtosis de cero, por lo que se suelen aceptar los valores cercanos (≈ 0.5 aprox.).
- Si $g_2 > 0$, la distribución es **Leptocúrtica**
- Si $g_2 < 0$, la distribución es **Platicúrtica**

- Si $g_2=0$, la distribución es normal.
- Si $g_2>0$, la distribución es más apuntada y con colas más gruesas que la normal.
- Si $g_2<0$, la distribución es menos apuntada y con colas menos gruesas que la normal.

Algunos investigadores sostienen que la curtosis poco tiene que ver con el centro de la distribución y su apuntamiento y mucho con las colas y la posible existencia de outliers.

Prueba de Wilcoxon de los rangos con signo

Esta prueba nos permite comparar los datos con una mediana teórica.