



Nada en esta maravillosa vida es 100% seguro. En todo lo que hacemos, siempre estamos estimando los chances de resultados exitosos: en los negocios, en la medicina, en el clima y principalmente en los juegos de azar.

Así el estudio de las leyes formales de lo aleatorio es la *Probabilidad*



**Estadística es la ciencia,
pura y aplicada, de creación,
desarrollo y de aplicación de
técnicas tales que la
incertidumbre de la
inferencia inductiva pueda
ser evaluada**

Tomar decisiones es una gran **responsabilidad**.

Para tomar decisiones se requiere **INFORMACIÓN** disponible, esperanzadamente **confiable** y **útil**.

Generalmente se necesita una porción de la base de datos o **muestra** para revelar un **patrón lógico** o realizar un **análisis estadístico**.

Estadística

Población

Muestra

Probabilidad

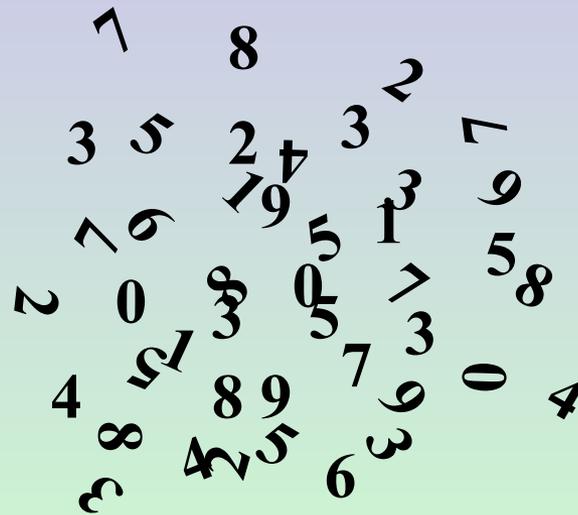
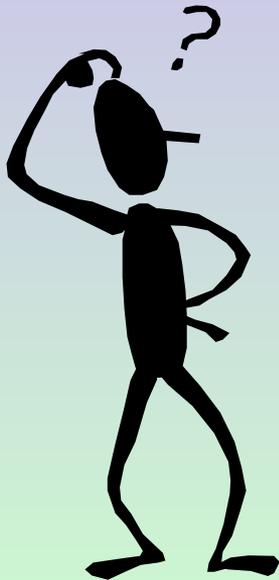
Objetivo de la Estadística Descriptiva

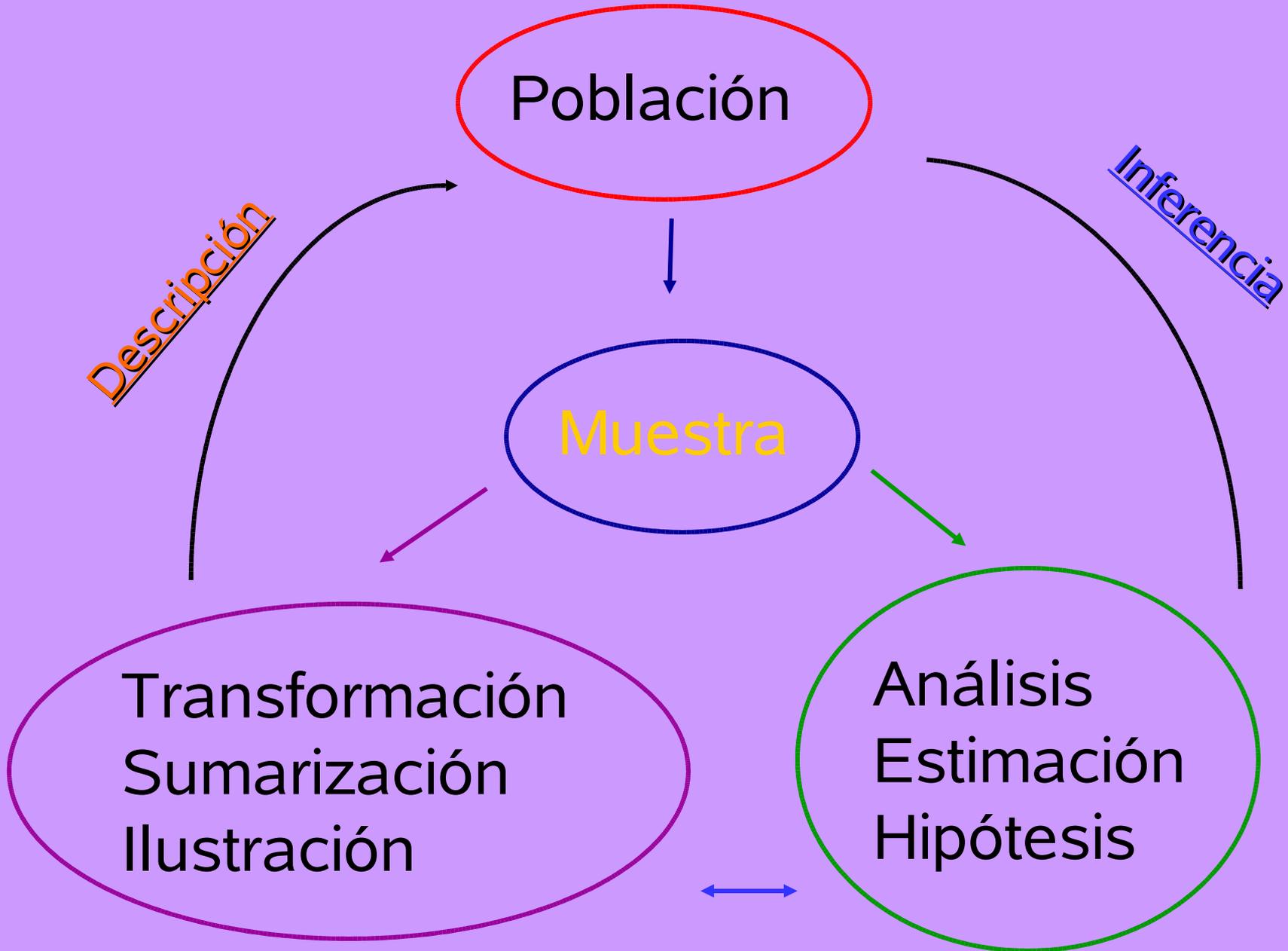
Conocer la información que se tiene para poder identificar e interpretar aspectos relevantes de una muestra.

Utilizar esta información para obtener resultados, planear o hacer inferencia acerca de la población bajo estudio.

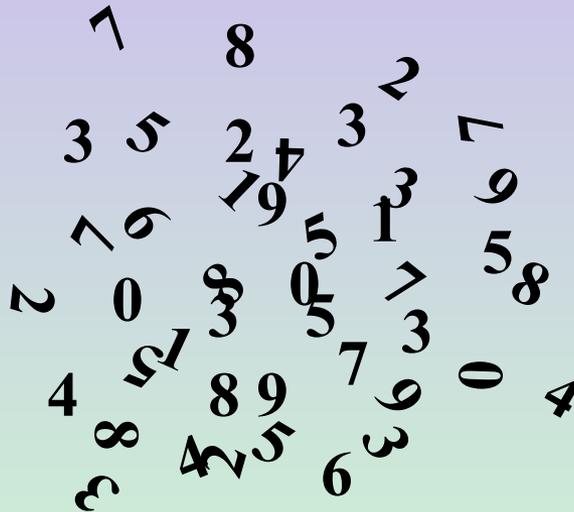
Los ***datos*** son la materia prima del estadístico. Usa los números para interpretar la realidad.

Todos los problemas estadísticos involucran o la recolecta, la descripción y el análisis de los datos, o pensar cómo recolectar, describir y hacer el análisis de los datos.





**Tengo un 98% de probabilidad
de hacer algo que tenga sentido
con estos números.**



El conjunto de datos que describen un fenómeno (nuestro objetivo) constituyen lo que se llama **Población**

Una **Muestra** es un subconjunto de la población sobre la cual vamos a realizar las medidas sobre una o mas características de interés

muestreo ...

¿Por qué muestreamos?

- Poblaciones muy grandes
- Respuesta rápida
- Destrucción de la muestra

muestreo ...

Una característica importante de una muestra es que sea **Representativa** de la población de interés.

Cualquiera que sea nuestro objetivo: describir a la población, analizar o pronosticar el comportamiento de la población, la muestra, al ser representativa, será **Confiable**

- Cada unidad tiene la misma oportunidad de ser elegida

- La selección de una unidad no tiene influencia sobre la elección de otra unidad

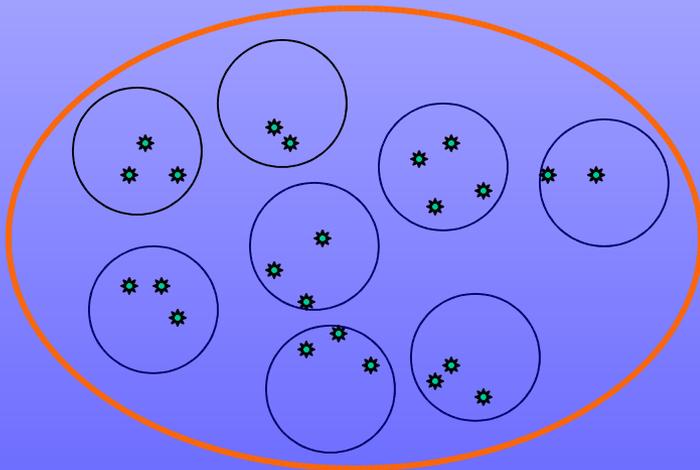
Muestreo Aleatorio



muestreo ...

Muestreo Estratificado

- Divide a la población en grupos homogéneos
- Se extrae una muestra aleatoria simple de cada grupo o estrato



Muestreo por Conglomerado

- Divide a la población en grupos
- Se extrae una muestra aleatoria simple de los grupos
- Se muestrean todos los elementos del grupo seleccionado

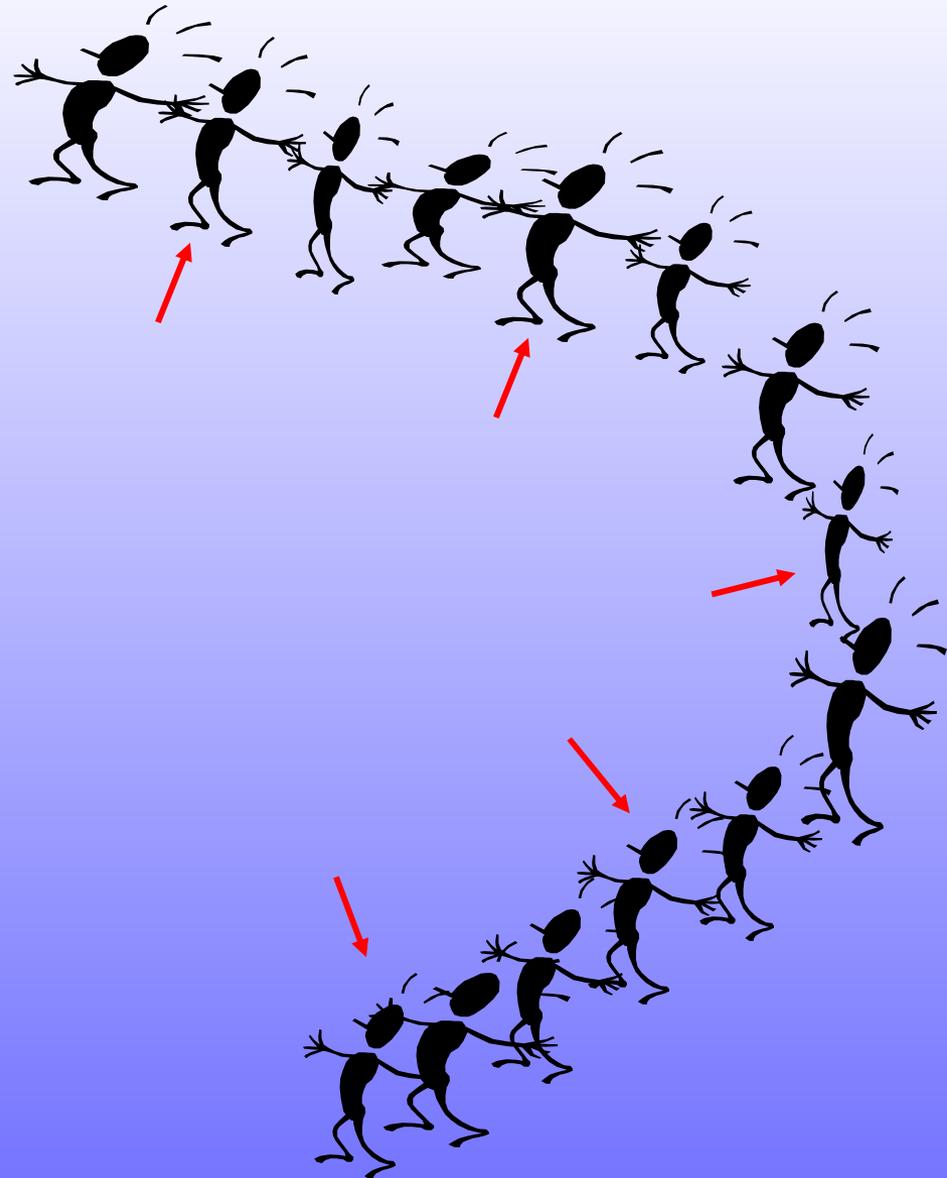
muestreo ...

Muestreo Sistemático

- Se elige aleatoriamente a una unidad
- A partir de ésta se selecciona cada k -ésima unidad que se encuentra después de la elegida

Muestreo Oportunista

- Se muestrean los n primeros elementos que se presentan



Un **parámetro** es una medida numérica de un aspecto de la población μ, σ, ν, θ

Una **estadística** es una medida numérica de un aspecto de la muestra \bar{X}, S, n, \tilde{X}

Una estadística consiste de un conjunto de mediciones de dicha característica que varía de una observación (**unidad experimental**) a otra, y a estas mediciones las llamaremos **variable**

No todas las variables son numéricas entonces podemos clasificarlas de acuerdo a su tipo en:

Cualitativas: Son variables que denotan una cualidad o atributo y solo pueden ser clasificadas en categorías o clases mutuamente excluyentes y exhaustivas

Cuantitativas: Son aquellas variables que se obtuvieron de un proceso de conteo (discretas) o medición (contínuas)

Clasificación de las variables Cualitativas de acuerdo a su escala de medición:

Nominal: Son clasificadas en categorías, sin importar el orden. No tiene sentido hacer operaciones aritméticas con ellas (género, grupo sanguíneo, Fuma (si/no))

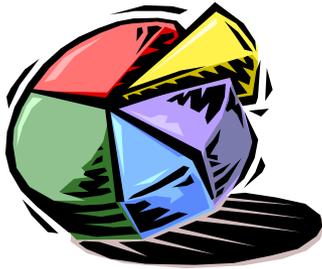
Ordinal: Las categorías se pueden arreglar en orden, pero las distancias entre las clases no son iguales (intensidad del dolor, escolaridad, nivel socioeconómico)

Clasificación de las variables Cuantitativas de acuerdo a su escala de medición:

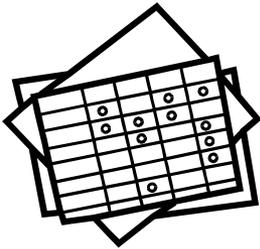
Intervalo: Son medidas en las que las distancias entre los valores es significativa pero no existe un cero absoluto (el cero no es ausencia de atributo) . No tiene sentido hacer cociente o producto (temperatura, usos horarios)

Razón: Las proporciones y razones tienen sentido al determinar cuánto mas tiene una unidad que otra de alguna característica. (peso, ingreso, rendimiento)

El análisis de cada variable se hace de acuerdo a su escala de medición



**Podemos
hacer
diagramas,
tablas y
resúmenes
numéricos
de los datos
recopilados**



¿Cómo presentar los datos?

La **frecuencia absoluta** f_i para una clase particular es el número de observaciones que caen en cada clase.

La **frecuencia relativa** o **porcentaje** para una clase particular es su frecuencia absoluta entre el número total de observaciones

$$p_i = \frac{f_i}{n}$$

Esta frecuencia ayuda a sumarizar en forma ordenada la información contenida en la muestra tanto en tablas como en gráficas.

<i>género</i>	<i>frecuencia</i>	<i>porcentaje</i>
0	19	0.63
1	11	0.37
Total	30	1

tabla de distribución de
frecuencias

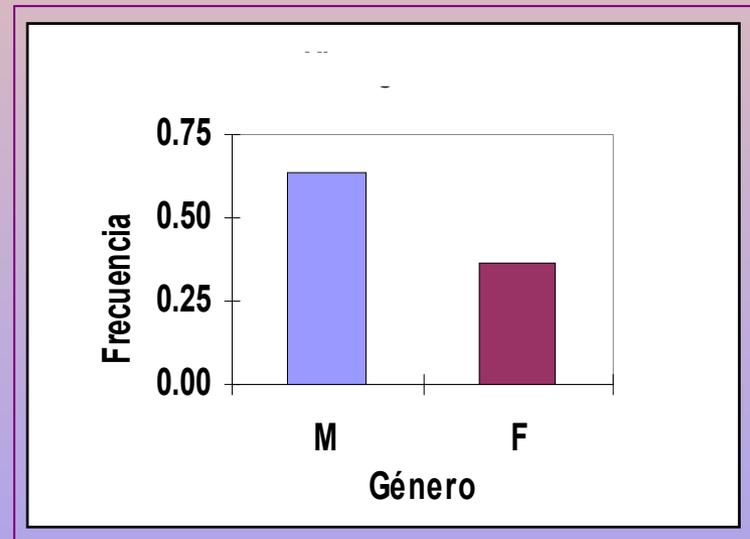
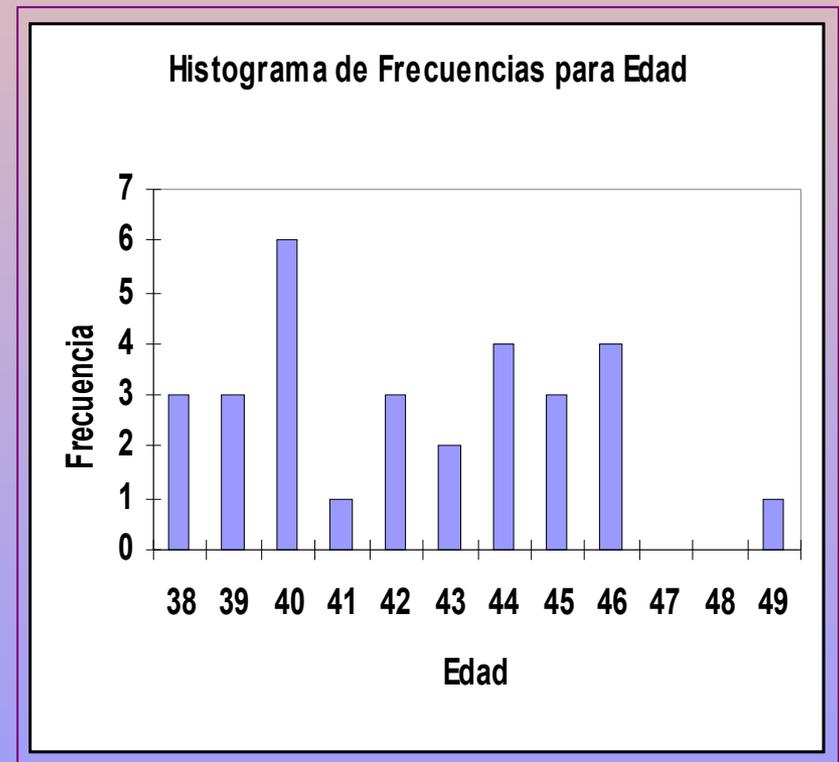


diagrama de barras

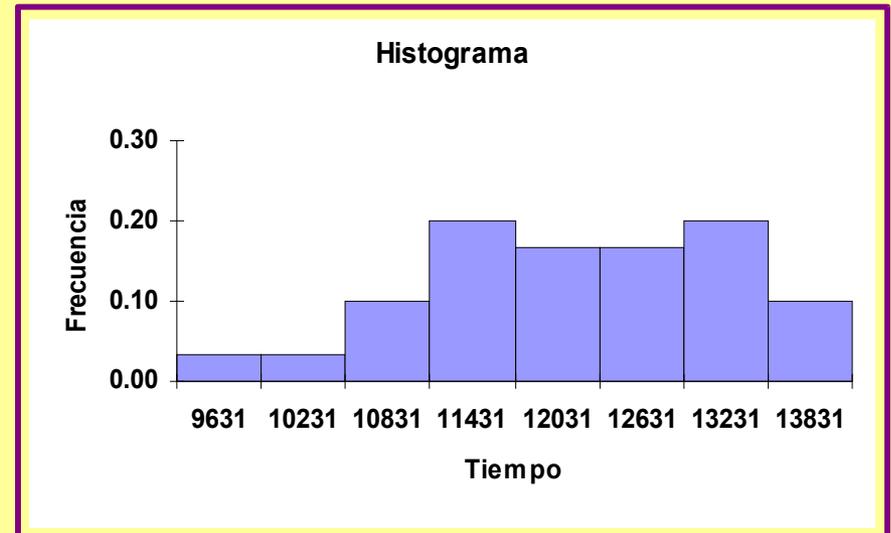
Si las variables son cuantitativas discretas las tablas de frecuencias se realizan con la creación de diferentes clases en base a los datos que toma la variable.

<i>edad</i>	<i>frecuencia</i>	<i>porcentaje</i>
38	3	0.10
39	3	0.10
40	6	0.20
41	1	0.03
42	3	0.10
43	2	0.07
44	4	0.13
45	3	0.10
46	4	0.13
47	0	0.00
48	0	0.00
49	1	0.03
Total	30	1.00



Si las variables son cuantitativas continuas las tablas de frecuencias se realizan con la creación de intervalos numéricos que formarán las diferentes clases.

<i>tiem po</i>	<i>frecuencia</i>	<i>porcentaje</i>
9331-9931	1	0.03
9931-10531	1	0.03
10531-11131	3	0.10
11131-11731	6	0.20
11731-12331	5	0.17
12331-12931	5	0.17
12931-13531	6	0.20
13531-14131	3	0.10
<i>Total</i>	30	1.00



Podemos completar esta tabla de frecuencias con una columna que nos de las Frecuencias Acumuladas ¿qué uso tienen?

tiempo (seg)	frecuencia	porcentaje	Frec. Acum .
9631	1	0.033	0.033
10231	1	0.033	0.067
10831	3	0.100	0.167
11431	6	0.200	0.367
12031	5	0.167	0.533
12631	5	0.167	0.700
13231	6	0.200	0.900
13831	3	0.100	1.000
Total	30	1.00	

a) **0.167**

c) **46.7%**

e) **12,031 seg**

- a) ¿qué frecuencia de individuos tuvieron un tiempo menor a 11,431 seg?
- b) ¿qué porcentaje de individuos tuvieron un tiempo mayor o igual a 12,631 seg?
- c) ¿qué tiempo máximo es el que al menos el 50% de las personas tuvieron?

Otros diagramas de utilidad:

3	38	0	0	0			
6	39	0	0	0			
12	40	0	0	0	0	0	0
13	41	0					
(3)	42	0	0	0			
14	43	0	0				
12	44	0	0	0	0		
8	45	0	0	0			
5	46	0	0	0	0		
1	47						
1	48						
1	49	0					

-se usa con pocos valores

-los datos están ordenados

-encontramos fácilmente
mínimo y máximo

-encontramos fácilmente los
porcentiles

-da una visión gráfica de la
distribución de los datos

diagrama de tallo y hojas
para la variable edad 38|0
= 38.0

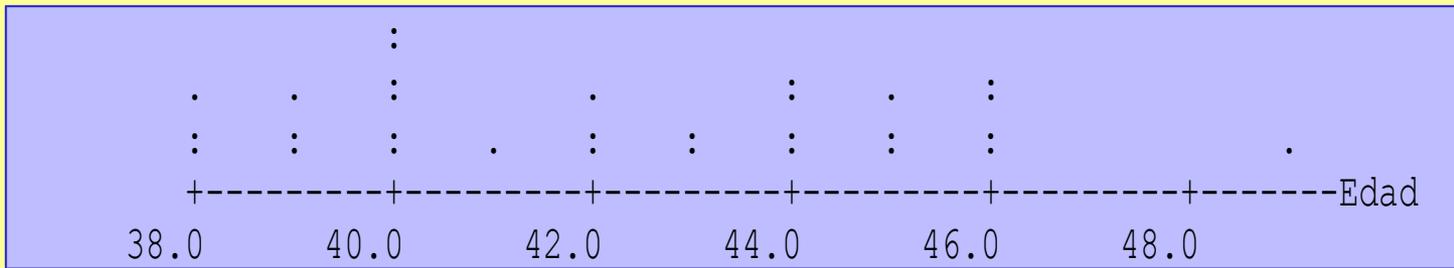
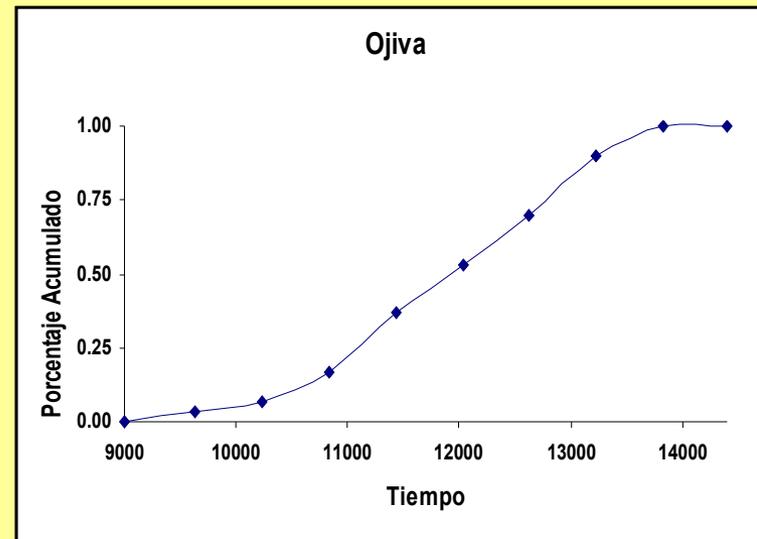
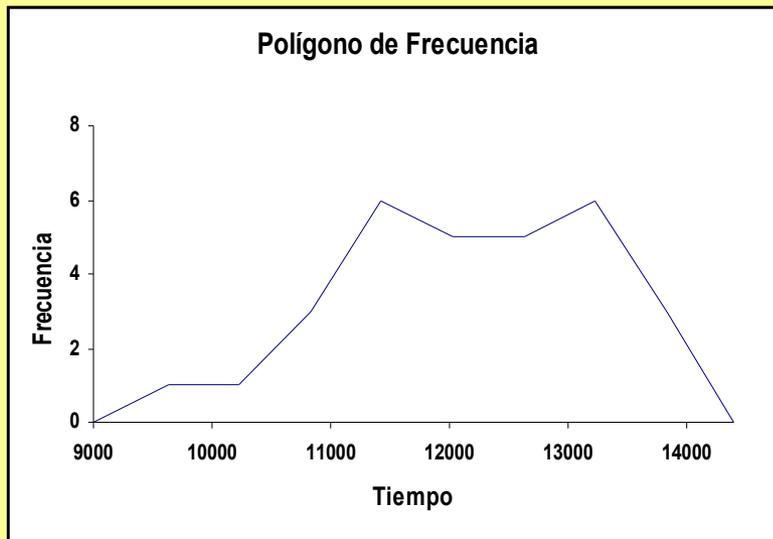


Diagrama de puntos. Útil para cuando tenemos pocos datos discretos

Polígono de Frecuencias

Ojiva



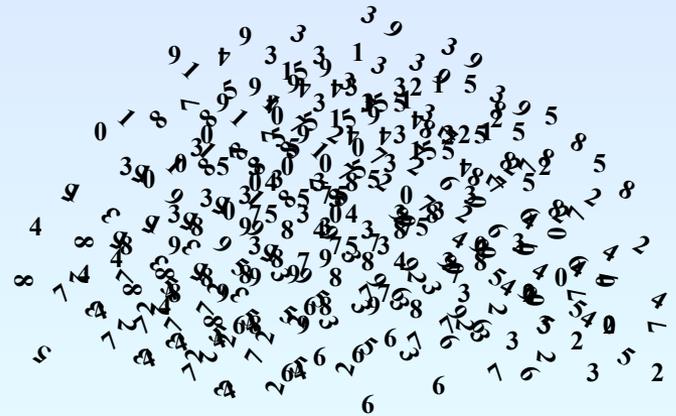
Métodos Numéricos

(válidos solo para datos cuantitativos)

Si pudiéramos escoger entre dos números que nos ayuden a construir una imagen mental burda de la distribución de un bonche de datos ¿Cuáles escogeríamos?

-un número que esté localizado cerca del centro de la distribución

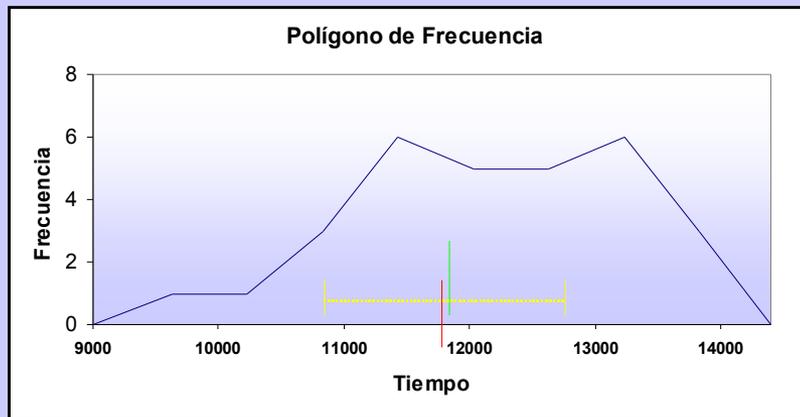
-un número que mida la dispersión de la distribución



Medidas de Tendencia Central

Son números que se localizan cerca del centro o cerca de donde se encuentran los datos con mayor frecuencia:

media, mediana, moda



Medidas de Dispersión

Son números que indican qué tan separados están los datos entre si: **rango, desviación estándar, rango intercuartil**

Medidas de tendencia central

media

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

en el caso en que los datos estén agrupados:

$$\bar{X}_a = \sum_{i=1}^k m_i \left(\frac{f_i}{n} \right)$$

tiem po (seg)	frecuencia	porcentaje	Frec. Acum .
9631	1	0.03	0.03
10231	1	0.03	0.07
10831	3	0.10	0.17
11431	6	0.20	0.37
12031	5	0.17	0.53
12631	5	0.17	0.70
13231	6	0.20	0.90
13831	3	0.10	1.00
<i>Total</i>	30	1	

$$\begin{aligned} \bar{X}_a &= 9631(0.03) + 10231(0.03) + \\ &\quad 10831(0.10) + \dots + 13831(0.10) \\ &= 12,187 \end{aligned}$$

mediana

se localiza el valor central
el valor que toma

$$l(\tilde{X}) = \frac{n+1}{2} \text{ y se observa}$$

3	38	0	0	0		
6	39	0	0	0		
12	40	0	0	0	0	0
13	41	0				
(3)	42	0	0	0		
14	43	0	0			
12	44	0	0	0	0	
8	45	0	0	0		
5	46	0	0	0	0	
1	47					
1	48					
1	49	0				

$$l(\tilde{X}) = \frac{30+1}{2} = 15.5$$

el valor que toma la
variable es 42.0 años

para datos agrupados se estima como

$$\tilde{X}_a = A + \frac{0.5 - P_{m-1}}{p_m} (B - A)$$

donde **A** y **B** son los límites inferior y superior de la clase que contiene a la mediana, **P_{m-1}** es la frecuencia acumulada hasta la clase anterior a la que contiene la mediana y **p_m** es la frecuencia relativa de la clase que contiene a la mediana

tiempo	frec.	porc.	Frec Acum
9331-9931	1	0.033	0.033
9931-10531	1	0.033	0.067
10531-11131	3	0.100	0.167
11131-11731	6	0.200	0.367
11731-12331	5	0.167	0.533
12331-12931	5	0.167	0.700
12931-13531	6	0.200	0.900
13531-14131	3	0.100	1.000
Total	30	1	

$$\tilde{X}_a = 11731 + \frac{0.5 - 0.367}{0.167} (600)$$

moda

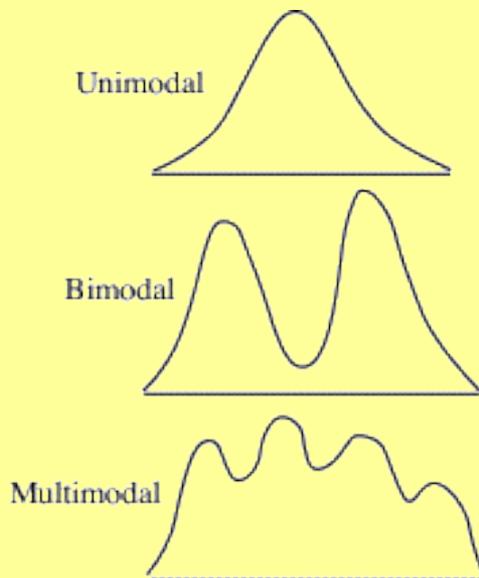
es el valor con la frecuencia mas alta.

La distribución puede ser unimodar, multimodal

cuando los datos están agrupados podemos hablar de una clase modal o estimar la moda:

$$Mo = A + \left(\frac{d_1}{d_1 + d_2} \right) (B - A)$$

donde **A** y **B** son los límites inferior y superior de la clase modal, $d_1 = f_{mo} - f_{mo-1}$ y $d_2 = f_{mo} - f_{mo+1}$



Medidas de Dispersión

rango se define como la diferencia entre el valor máximo y el mínimo:

$$Rango = max - min$$

Es una medida **sensible** a valores extremos y no es muy informativa ya que es **insensible** a datos intermedios

amplitud intercuartílica es la distancia entre el percentil 75 y el percentil 25:

$$AI = P_{75} - P_{25}$$

Nos da una idea de la distancia entre los valores que determinan el 50% de los datos centrales

Varianza es una variación promedio alrededor de la media, definida como

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

un problema de la varianza es que tiene las unidades al cuadrado y su interpretación no es fácil, por lo que usamos su raíz: **desviación estándar**

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

es sensible a valores extremos.

Si los datos están agrupados en k intervalos, la varianza se estima como:

$$S_a^2 = \frac{\sum_{i=1}^k f_i (m_i - \bar{X}_a)^2}{n-1}$$

para nuestro ejemplo ¿qué valor toma S_a^2 ?

Hay algunas formas de poner juntos a la desviación estándar y a la media muestrales

Coeficiente de Variación: es una medida de variación relativa y expresa la desviación estándar como un porcentaje de la media aritmética. Se obtiene como

$$CV = \frac{S}{\bar{X}} \times 100$$

por su falta de dimensiones es muy útil para comparar variación entre diferentes poblaciones, que a simple vista serían difíciles de comparar.

Creación de Intervalos:

con s y \bar{X} se pueden formar intervalos de la forma $\bar{X} \mp kS$ y obtener el número de observaciones que caen dentro de ese intervalo.

Si nuestra distribución muestral tiene una forma mas o menos simétrica y acampanada podemos usar la regla empírica:

alrededor del 69% de las observaciones cae dentro de una desviación estándar de la media

alrededor del 95% de las observaciones cae dentro de dos desviaciones estándar de la media

alrededor del 97.7% de las observaciones cae dentro de tres desviaciones estándar de la media

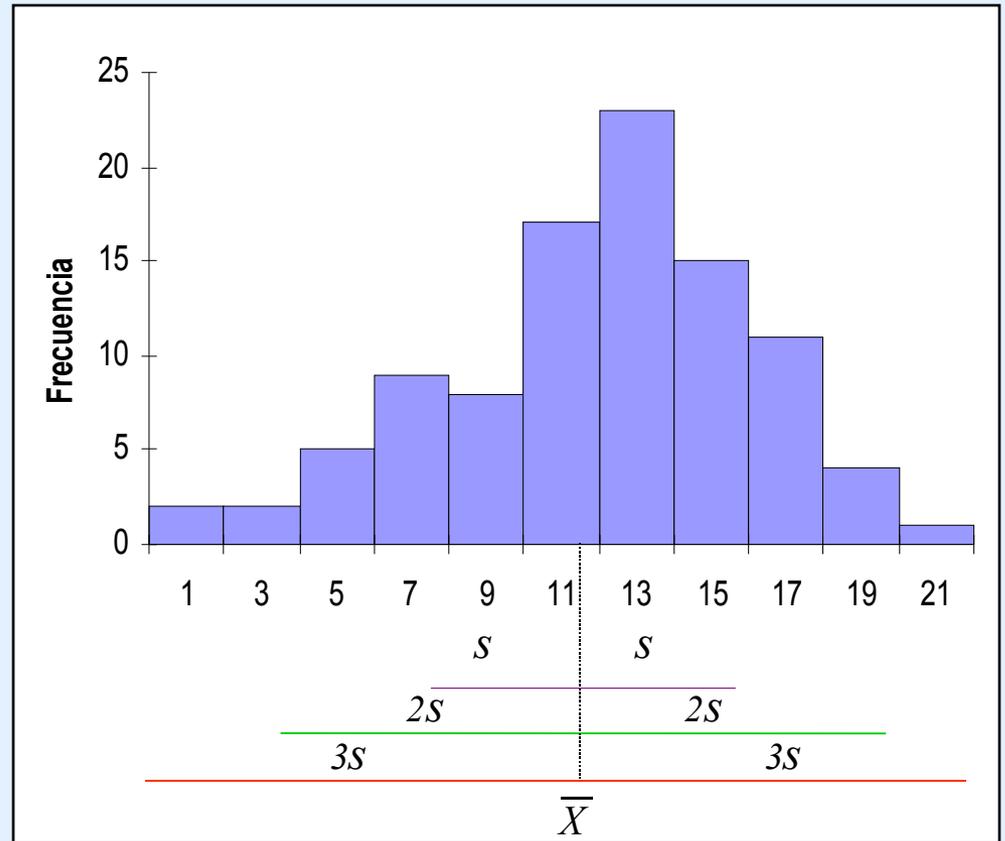
Monóxido de Carbono en el humo de los cigarrros

Intervalos alrededor de la media

$n = 372$

$\bar{X} = 11.66$

$s = 4.089$



$\bar{X} \mp s$ (7.57 , 15.75) 264 obs. 70.96%

$\bar{X} \mp 2s$ (3.48 , 19.84) 353 obs. 94.89%

$\bar{X} \mp 3s$ (-0.61 , 23.93) 372 obs. 100.00%

medidas de dispersión...

Medidas de Posición Relativa

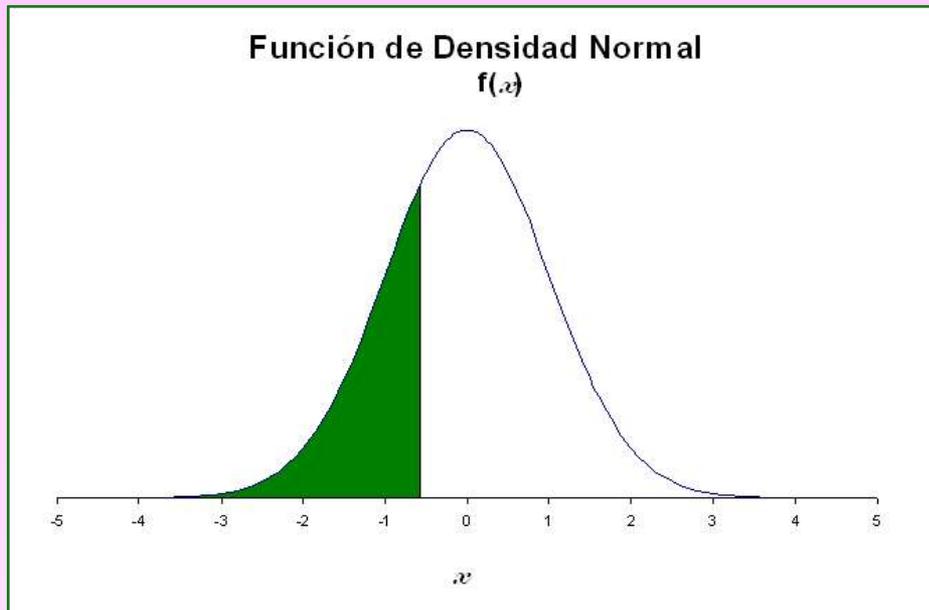
Son medidas descriptivas que localizan la posición de una medición en relación a otras mediciones.

Una medida que expresa esta posición en términos de un porcentaje es llamado **porcentil**

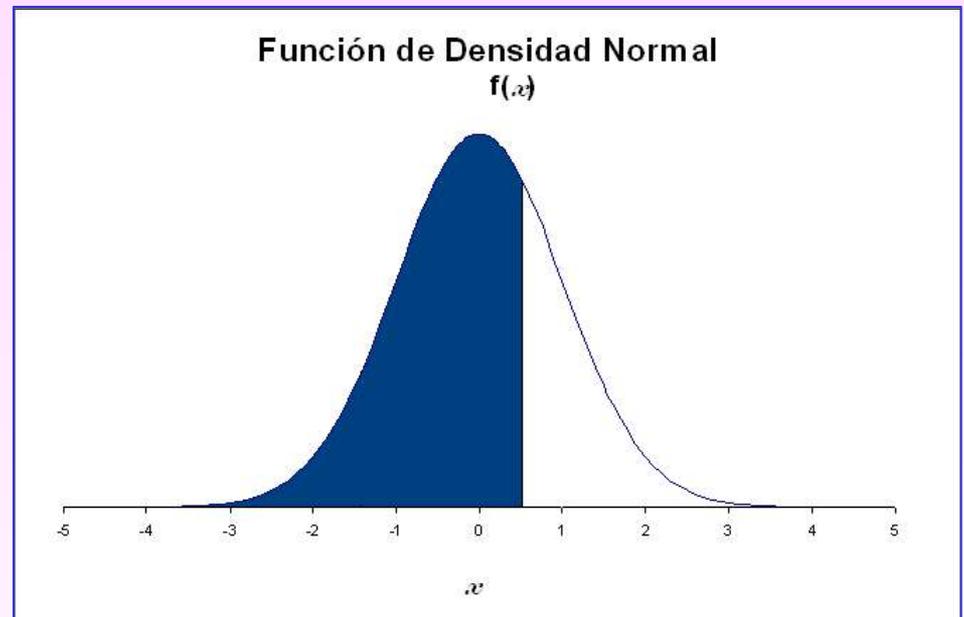
El **porcentil** de orden α (P_α) es el valor de la variable por debajo del cual se encuentra una frecuencia acumulada α .

medidas de posición relativa...

El porcetil 25 o primer cuartil $Q_1 = -0.675$ deja a su izquierda el 25% de las observaciones



El porcetil 70, es decir, $P_{70} = 0.525$ deja a su izquierda el 70% de las observaciones



medidas de posición relativa...

El diagrama de tallo y hojas, nos ayuda a localizarlos rápidamente

3	38	0	0	0			
6	39	0	0	0			
12	40	0	0	0	0	0	0
13	41	0					
(3)	42	0	0	0			
14	43	0	0				
12	44	0	0	0	0		
8	45	0	0	0			
5	46	0	0	0	0		
1	47						
1	48						
1	49	0					

-los datos se ordenan de menor a mayor

-se encuentra la localización de los percentiles:

$$l(P_{\alpha}) = \left(\frac{\alpha}{100} \right) (n)$$

-se lee el valor de dicha observación

-si la localización es fraccionaria se toma el promedio del valor en la localización anterior y posterior

los percentiles no necesariamente son números observados

medidas de posición relativa...

1	9		6				
3	10		2	3			
6	10		6	7	9		
11	11		0	1	3	3	3
(5)	11		7	7	8	8	8
14	12		0	0	1	2	3
9	12		6	6	9	9	
5	13		0	2	3	3	4

$$l(Q_1) = (25/100)31 = 7.75$$

$$\Rightarrow Q_1 = P_{25} = 11.05$$

$$l(Q_3) = (75/100)31 = 23.25$$

$$\Rightarrow Q_3 = P_{75} = 12.65$$

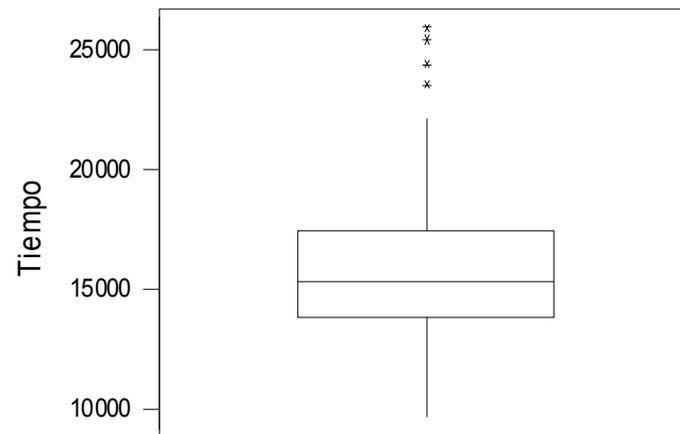
$$l(Q_2) = (50/100)31 = 15.5$$

$$\Rightarrow Q_2 = P_{50} = 11.8$$

Los cuartiles Q_1 y Q_3 son muy útiles para construir lo que llamamos **diagrama de caja y brazos**.

Diagrama de Caja y Brazos

Nos permite ver la distribución de los datos, el máximo, el mínimo, la localización de los Cuartiles, y la dispersión por cuartiles. Nos permitirá ver si existe un sesgo así como puntos extremos.



Análisis Exploratorio de Datos

Para hacer estadística diferente a la descriptiva, podemos usar todas las técnicas hasta ahora aprendidas y hacer algún análisis comparativo o asociativo.

El problema de comparación consiste en **contrastar** las distribuciones de frecuencia de una variable de interés para dos o más **subpoblaciones** basándose en los datos de la muestra.

En el problema de comparación surgen algunas preguntas:

¿Hay alguna diferencia en las distribuciones poblacionales?

¿Cuál es la naturaleza de esas diferencias?

¿Qué tan grandes son esas diferencias?

El análisis exploratorio nos ayudará a darnos una idea de las respuestas a estas preguntas

comparación...

La comparación de las distribuciones de frecuencia entre subpoblaciones cuando la variable de interés es cualitativa se hace con una tabla de contingencias o tabulación cruzada

	Hábitos de tabaquismo			
Género	Nunca ha fumado	Dejó de fumar	Fuma actualmente	Total
Masculino	154	25	185	364
Femenino	127	11	38	176
Total	281	36	223	540

las frecuencias pueden ser relativas o absolutas y nos dan una idea de qué tan frecuente se presentan simultáneamente ambos atributos en una población

comparación...

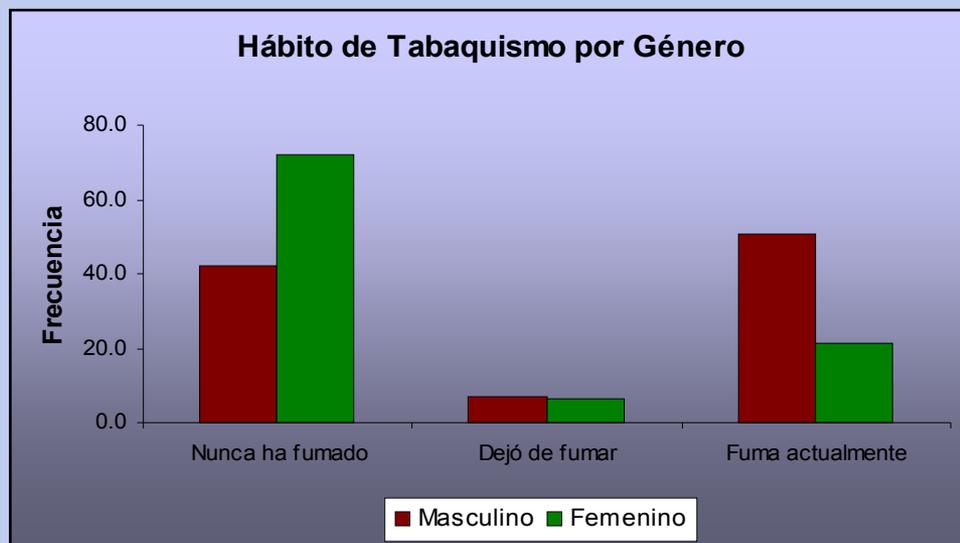
El objetivo de la comparación es ver si una característica determinada varía relativo a alguna subclase, por lo que se calculan las frecuencias relativas condicionales f_{ij}/f_i ó p_{ij}/p_i (de ésta manera compensamos por diferencias de tamaños) ...

	Hábitos de tabaquismo (%)			
Género	Nunca ha fumado	Dejó de fumar	Fuma actualmente	Total
Masculino	28.5	4.6	34.3	67.4
Femenino	23.5	2.1	7.0	32.6
Total	52.0	6.7	41.3	100

comparación...

... y calculamos las frecuencias relativas condicionadas a género

	Hábitos de tabaquismo condicionadas a género (%)			
Género	Nunca ha fumado	Dejó de fumar	Fuma actualmente	Total
Masculino	42.3	6.8	50.9	100
Femenino	72.1	6.5	21.5	100
Total	52.0	6.7	41.3	100

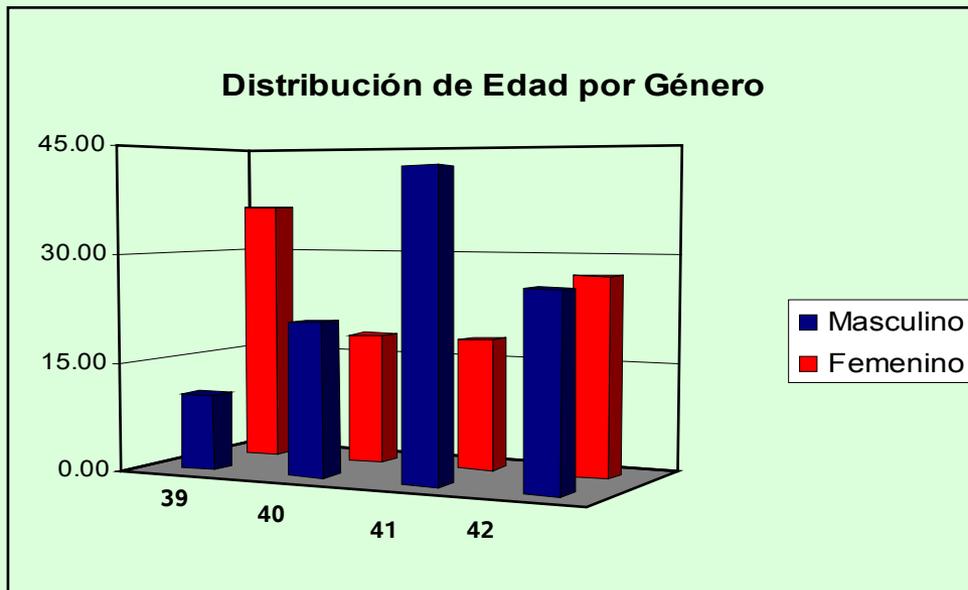


¿el hábito de tabaquismo difiere si se es hombre o mujer?

comparación...

Si la variable a analizar es discreta se puede tratar como si fuera cualitativa.

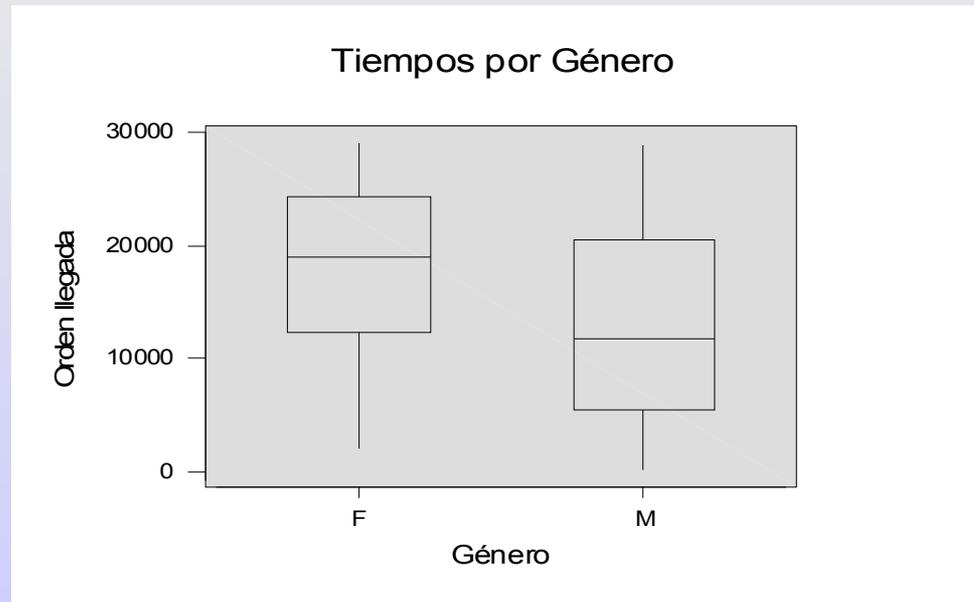
	Edad en años condicionada a género (%)				
Género	39	40	41	42	Total
Masculino	10.53	21.05	42.11	26.32	100
Femenino	36.36	18.18	18.18	27.27	100
Total	20.00	20.00	33.33	26.67	100



¿hay alguna diferencia entre géneros con respecto a la edad?

comparación...

En el caso de que la variable a analizar sea cont nua podemos estar interesados en comparar tanto la localizaci n como la dispersi n entre las distribuciones de frecuencia de las subpoblaciones. Una manera de hacerlo es por medio de un diagrama esquem tico



**  Qui n tiene tiempos m s altos?
  qui n tiene mayor dispersi n?**

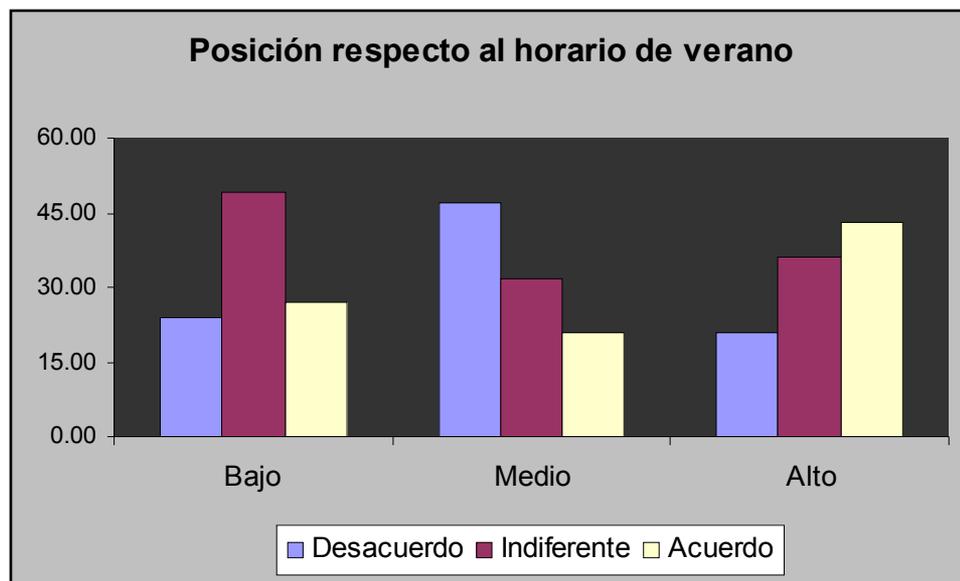
Muchas veces es importante saber si una variable influye sobre el comportamiento de otra variable. Con ello estudiamos el problema de asociación.

Ambas Variables Ordinales

El uso de la tabla de contingencia y su correspondiente diagrama de barras es de gran utilidad para asociar variables cualitativas en escala ordinal.

Ésta tabla se presenta con las frecuencias relativas condicionadas a las clases de una de las variables

	Posición respecto al horario de verano				
		Desacuerdo	Indiferente	Acuerdo	Total
Nivel Socioeconómico	Bajo	23.90	49.02	27.07	100.00
	Medio	47.02	31.93	21.05	100.00
	Alto	20.69	36.21	43.10	100.00



¿A mayor nivel socioeconómico, mayor aceptación?

asociación ...

Una Variable Ordinal y otra Cuantitativa

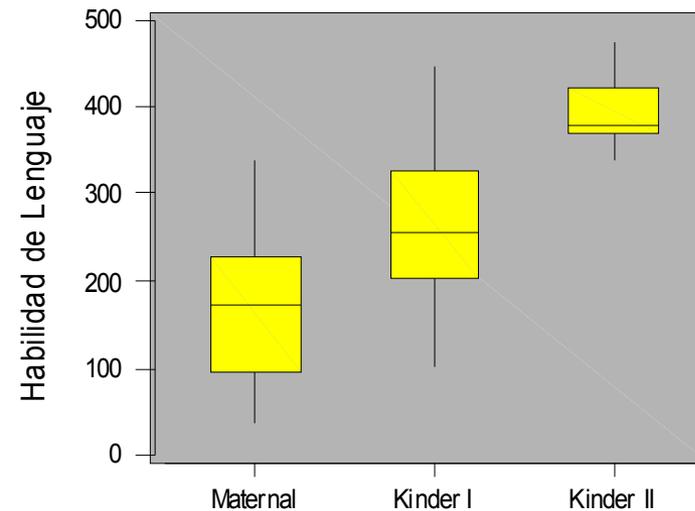
Una manera de evidenciar la posible asociación entre las variables es a través del diagrama esquemático.

Éste diagrama nos da una idea de cómo dependen la variable cuantitativa, no solo en localización sino también en dispersión con respecto al aumento o disminución en escala de la variable cualitativa ordinal.

asociación ...

Grado Escolar		
Maternal	Kinder I	Kinder II
68	255	425
35	202	370
145	317	380
173	327	476
190	247	410
225	100	358
340	448	338
123	412	373
228	228	377
	192	467
	297	388

¿Qué nos dice este diagrama esquemático?



Ambas Variables Cuantitativas

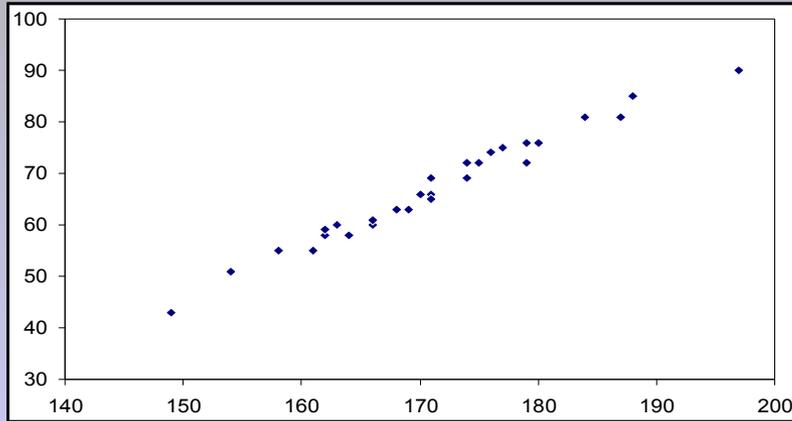
Para este caso el diagrama de dispersión es muy usado para asociar variables cuantitativas.

Consiste en graficar parejas de valores (x_i, y_i) correapondientes a un solo individuo, sobre un plano cartesiano.

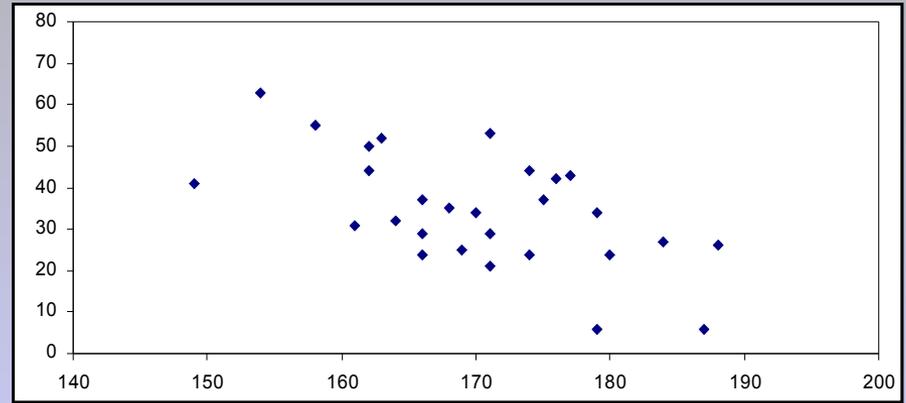
Una medida de asociación que complementa este diagrama es el coeficiente de correlación (medida de **relación lineal** entre las variables) obtenido como

$$r(x, y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) / (n-1)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)}} = \frac{S_{xy}}{S_x S_y}$$

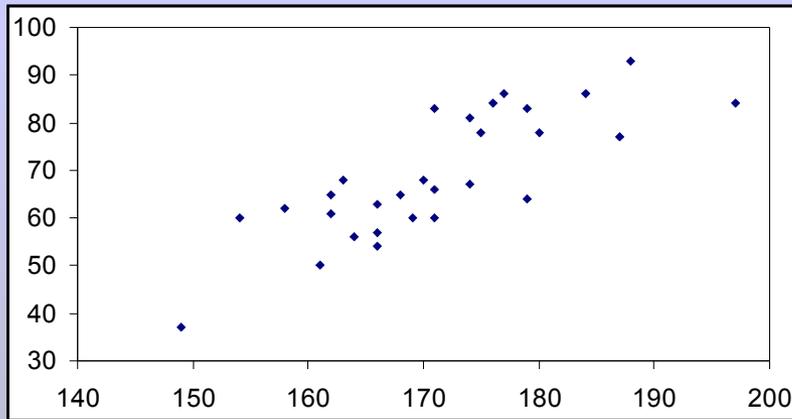
asociación ...



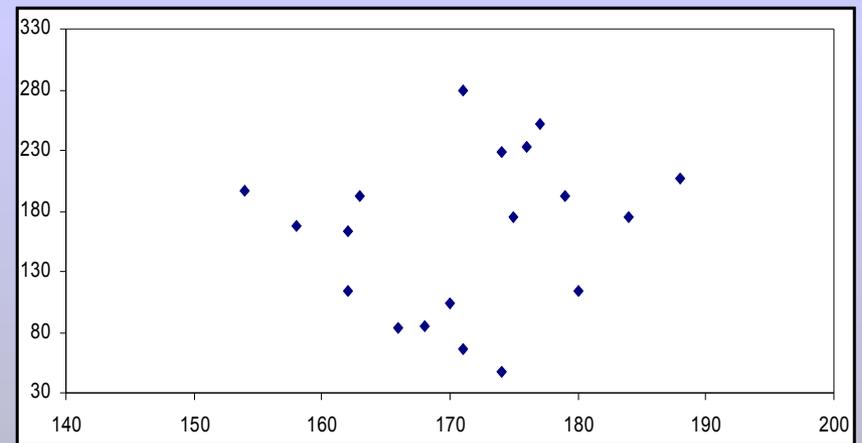
$$r = 0.99$$



$$r = -0.7$$



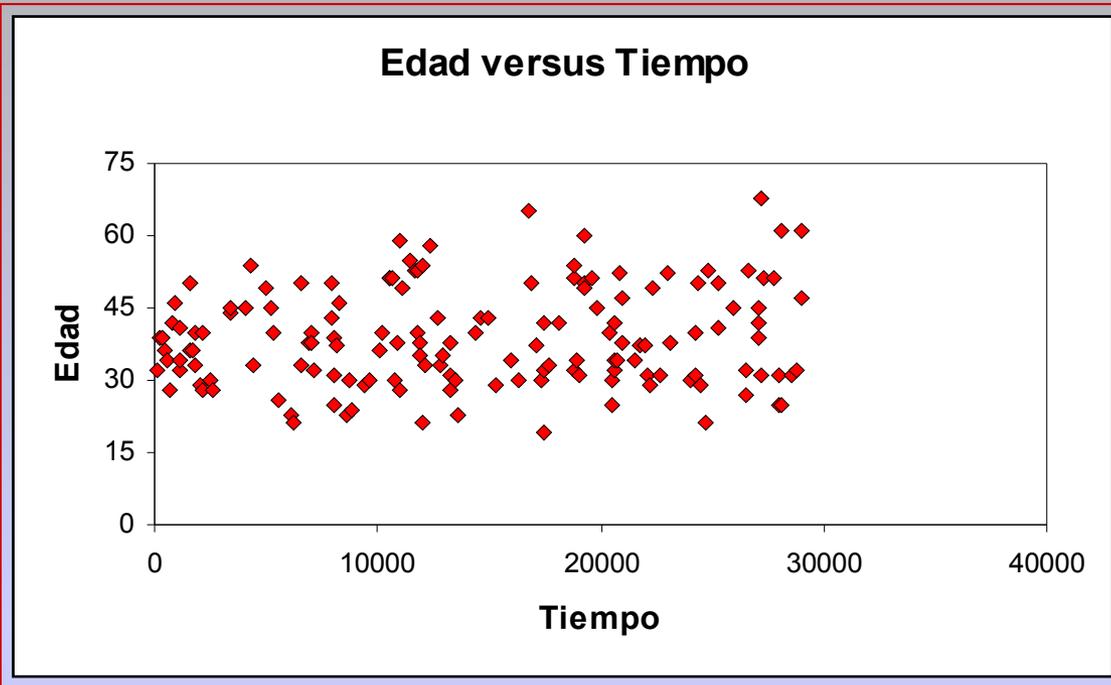
$$r = 0.8$$



$$r = 0.1$$

¿Se puede decir que si r es cero, las variables son independientes?

asociación ...



$$r = 0.130$$

¿Existe alguna relación lineal entre el tiempo que tomó correr el maratón y la edad de los participantes? ¿Confirma el valor de r esta relación?