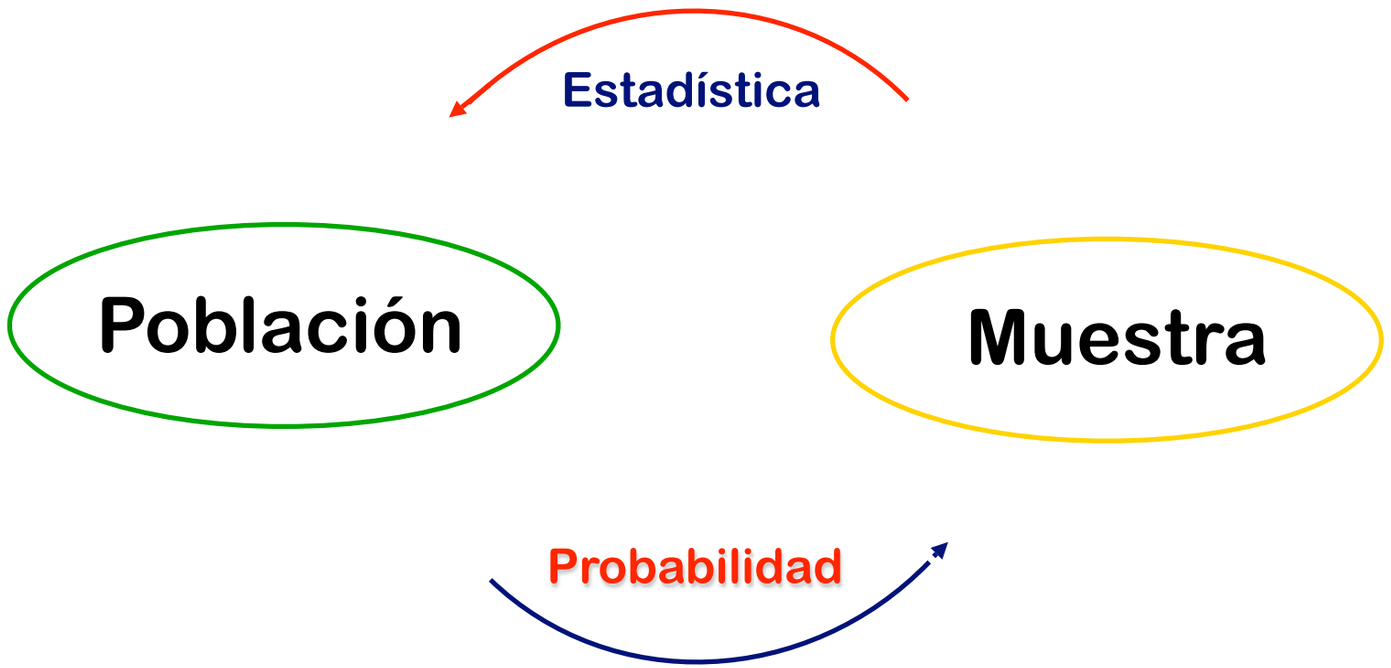

ESTADÍSTICA DESCRIPTIVA
(análisis exploratorio de datos)



rohen

Juntemos la estadística y la probabilidad.

Al realizar un experimento aleatorio muchas veces, esperamos que los resultados obtenidos sean gobernados por sus probabilidades. Así las probabilidades forman un modelo de la realidad y la realidad nos ayuda a establecer dicho modelo

rohen

Objetivo de la Estadística Descriptiva

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

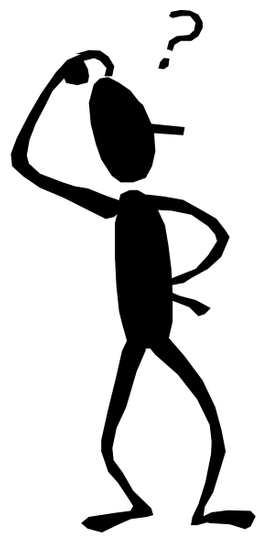
Utilizar esta información para tener resultados, planear o hacer inferencia acerca de la población bajo estudio

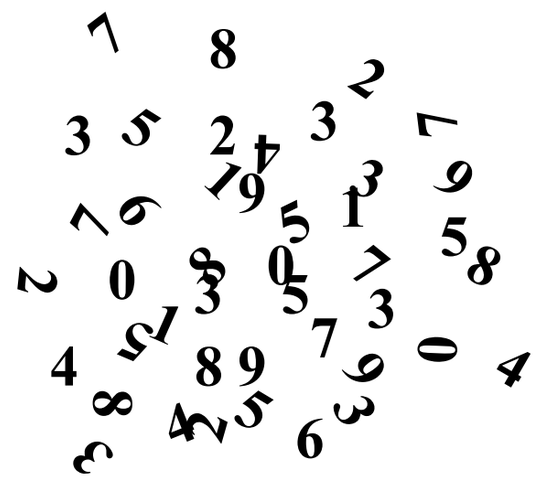
Tomar decisiones es una gran responsabilidad.

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

rohen

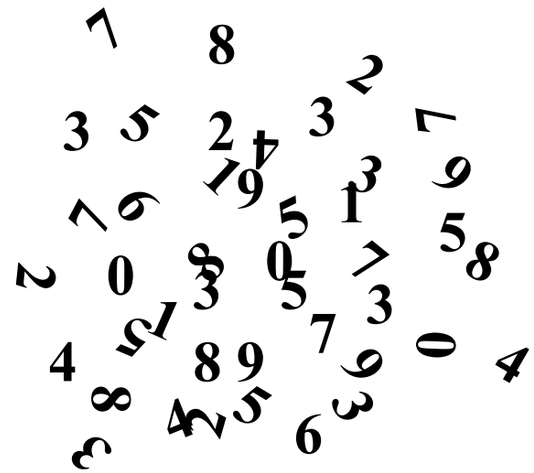
Los **datos** son la materia prima del estadístico. Usa los números para interpretar la realidad. Todos los problemas estadísticos involucran o la recolecta, la descripción y el análisis de los datos, o pensar cómo recolectar, describir y hacer el análisis de los datos.





rohen

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



rohen

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

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

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

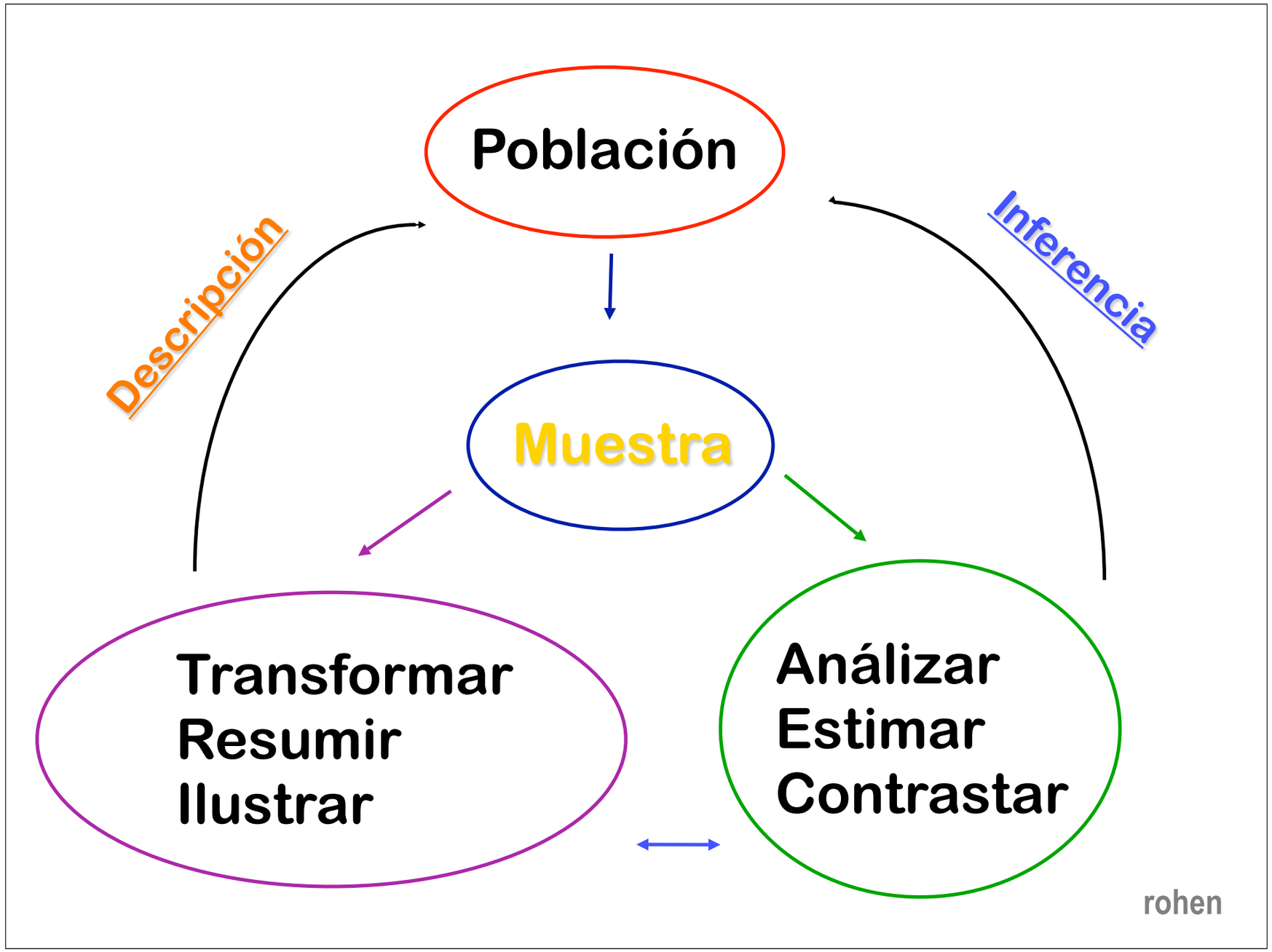
¿Por qué muestreamos?

- Poblaciones muy grandes

- Respuesta rápida

- Destrucción de la muestra

- Costo


Población
Descripción
Inferencia
Muestra
Transformar
Resumir
Ilustrar
Análizar
Estimar
Contrastar
rohen

Cualquiera que sea nuestro objetivo:

- describir a la población,
- analizar o
- pronosticar el comportamiento de la población,

La muestra debe ser

Representativa

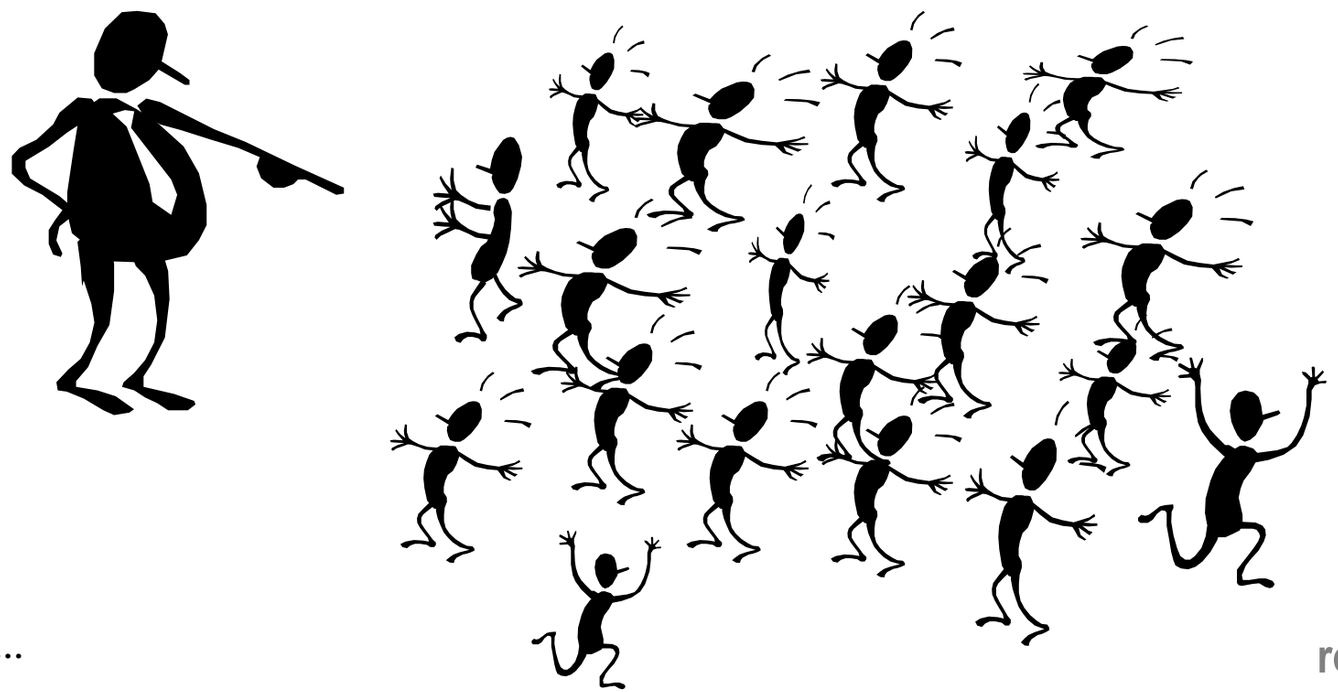
para que sea

Confiable

Muestreo Aleatorio

- Cada unidad tiene la misma oportunidad de ser elegida

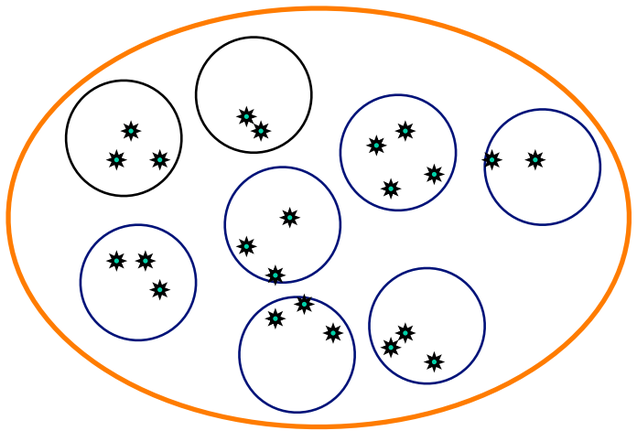
- La selección de una unidad no tiene influencia sobre la elección de otra unidad



Muestreo Estratificado

- Divide a la población en grupos homogéneos

- Se extrae una muestra aleatoria simple de cada grupo o estrato, proporcional al tamaño de éste



Muestreo por Conglomerado

- Divide a la población en grupos

-Se extrae una muestra aleatoria simple de los grupos

-Se muestrean <u>todos</u> los elementos del grupo seleccionado

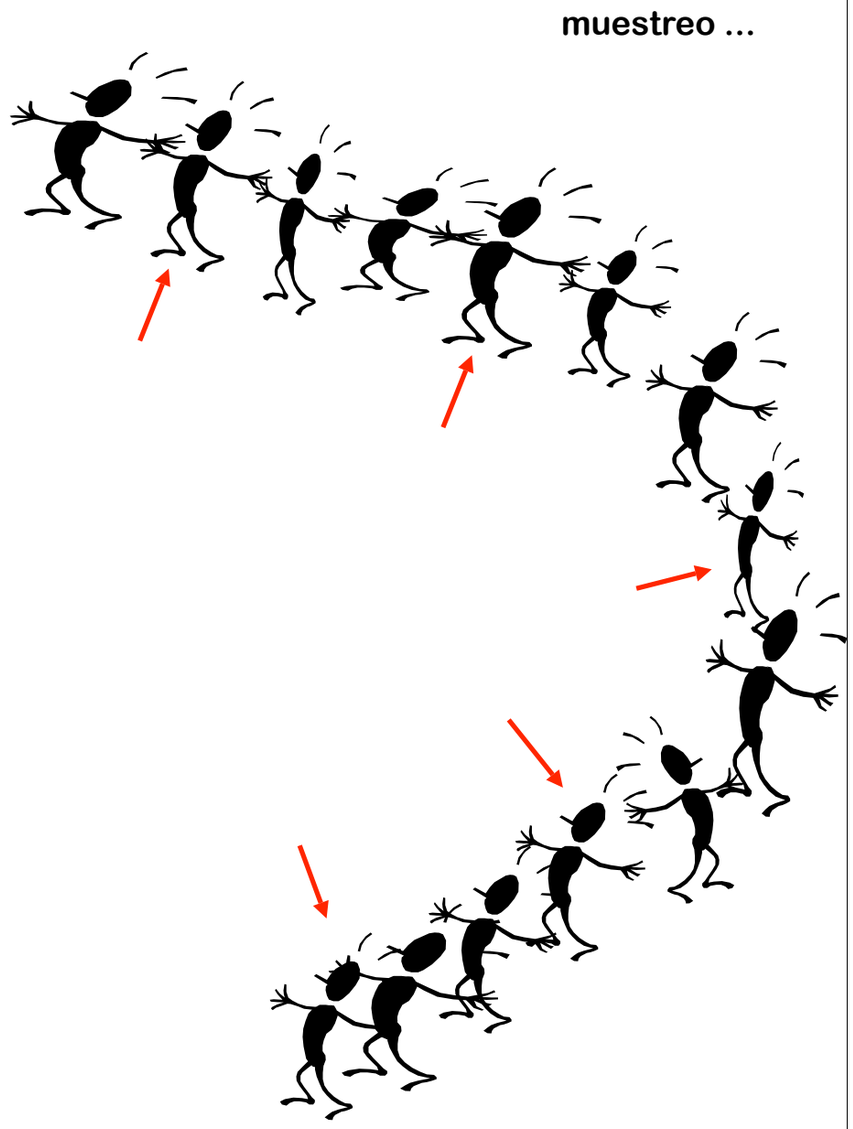
Muestreo Sistemático

- 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



Algunos conceptos importantes antes de empezar a describir...

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 $\overline{X}, S, n, \tilde{X}$

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**

rohen

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))

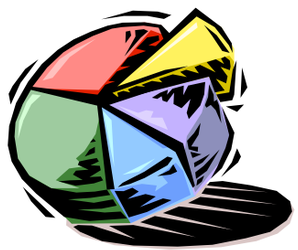
Ordinal: Las categorías se pueden arreglar en orden, pero las distancias entre las clases no necesariamente son iguales (intensidad del dolor, escolaridad, nivel socioeconómico)

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

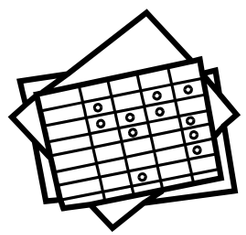
Intervalo: Son medidas en las que las distancias entre los valores es significativa pero no existe un cero absoluto (el cero no es ausencia de atributo) . No tiene sentido hacer cociente o producto (temperatura, usos horarios)

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

género	*frecuencia*	*porcentaje*
0	19	0.63
1	11	0.37
Total	30	1.00

tabla de distribución de frecuencias

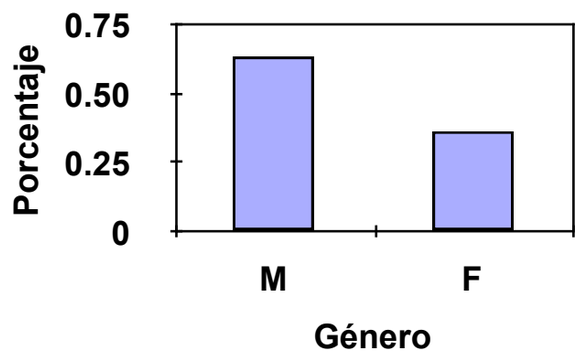


diagrama de barras

rohen

Gráfico de pastel

o

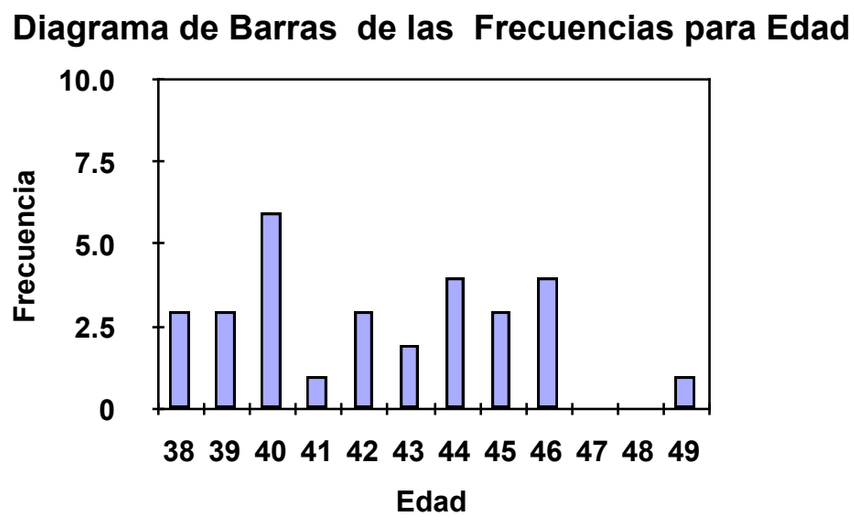
diagrama

circular

rohen

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

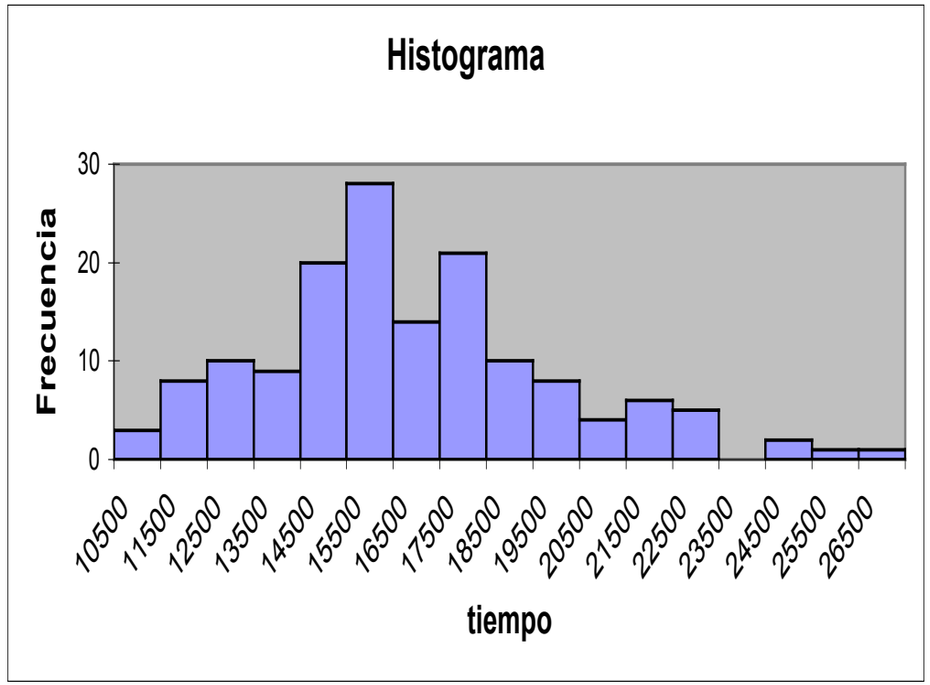
edad	*frecuencia*	*frecuencia relativa*
38	3	0.10
39	3	0.10
40	6	0.20
41	1	0.03
42	3	0.10
43	2	0.07
44	4	0.13
45	3	0.10
46	4	0.13
47	0	0.00
48	0	0.00
49	1	0.03
Total	30	1.00



rohen

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

tiempo	*Frecuencia*
10500	3
11500	8
12500	10
13500	9
14500	20
15500	28
16500	14
17500	21
18500	10
19500	8
20500	4
21500	6
22500	5
23500	0
24500	2
25500	1
26500	1



rohen

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

tiempo	*frecuencia*	*Frec. Relat.*	*Frec. Rel. Acum.*
9331- 9931	20	0.033	0.033
9931-10531	20	0.033	0.067
10531-11131	60	0.100	0.167
11131-11731	120	0.200	0.367
11731-12331	100	0.167	0.533
12331-12931	100	0.167	0.700
12931-13531	120	0.200	0.900
13531-14131	60	0.100	1.000

a) 0.167

b) 46.7%

c) 12,331 min.

a) ¿qué frecuencia de personas tuvieron un tiempo promedio menor a 11,131 segundos?

b) ¿qué porcentaje de personas tuvieron un tiempo promedio mayor o igual a 12,331 segundos?

c) ¿cuál es el máximo de minutos promedio que al menos el 50% de los atletas tuvieron?

rohen

Otros diagramas de utilidad:

18	0 0
19	0 0 0 0
20	0 0 0 0 0 0 0
21	0 0 0 0 0
22	0 0 0
23	0 0
24	0 0 0
25	0
26	0

diagrama de tallo y hojas
para la variable edad
18|0 = 18.0

-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

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?

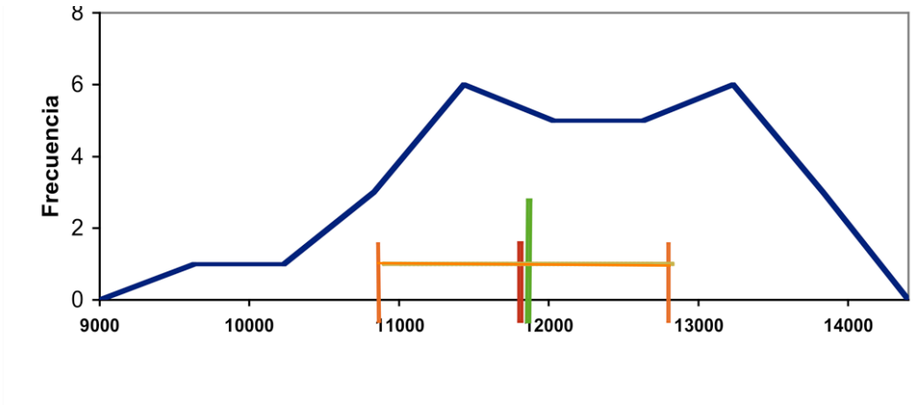
-un número que esté localizado cerca del centro de la distribución

-un número que mida la dispersión de la distribución

rohen

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

rohen

Medidas de tendencia central

media

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

{ Es sensible a valores extremos

Mediana

Se localiza el valor central como $l(\tilde{X}) = \frac{n+1}{2}$

Se observa el valor que toma la posición central

3	38	0 2 5
6	39	1 7 8
12	40	0 0 1 3 7 9
13	41	5
(3)	42	4 6 8
14	43	3 6 7
12	44	1 3 5 6
8	45	0 7 9
5	46	2 2 3 8
1	47	
1	48	4

$$l(\tilde{X}) = \frac{32}{2} = 16$$

El valor que toma la mediana es 42.8 cm.

Diagrama de tallo y hoja para la variable tamaño
38|0 = 38.0 cm.

Si n es par, $l(\tilde{X})$ es fraccionaria y se observan los valores que toman las dos posiciones centrales

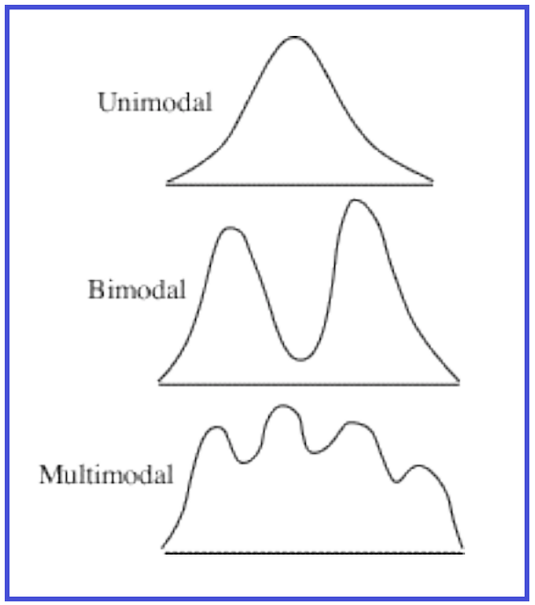
3	38	0 2 5
6	39	1 7 8
12	40	0 0 1 3 7 9
13	41	5
(3)	42	4 6 8
14	43	3 6
12	44	1 3 5 6
8	45	0 7 9
5	46	2 2 3 8
1	47	
1	48	4

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

se promedian los valores centrales. El valor que toma la mediana es 42.7 pesos

moda

es el valor con la frecuencia mas alta.
La distribución puede ser unimodal o multimodal



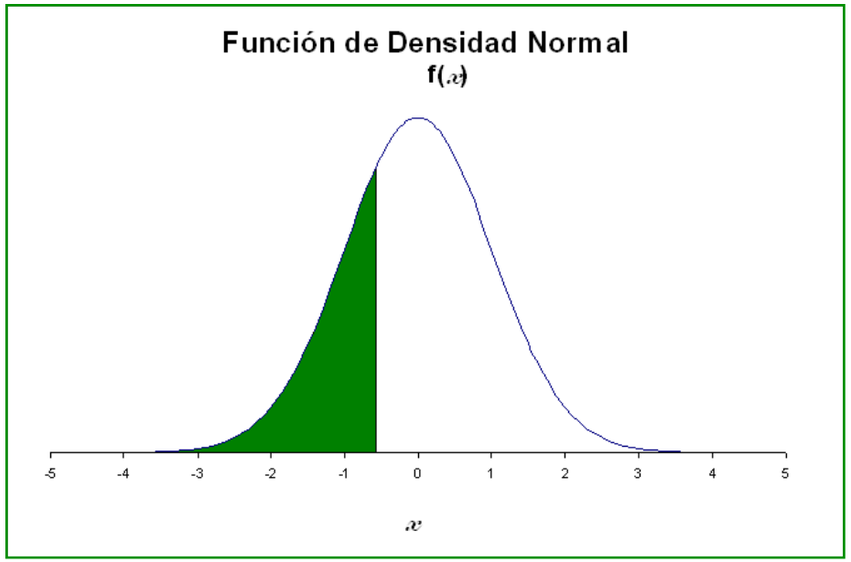
cuando los datos están agrupados podemos hablar de una clase modal que no necesariamente coinsidirá con la moda, en caso de que ésta exista.

Medidas de Posición Relativa

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

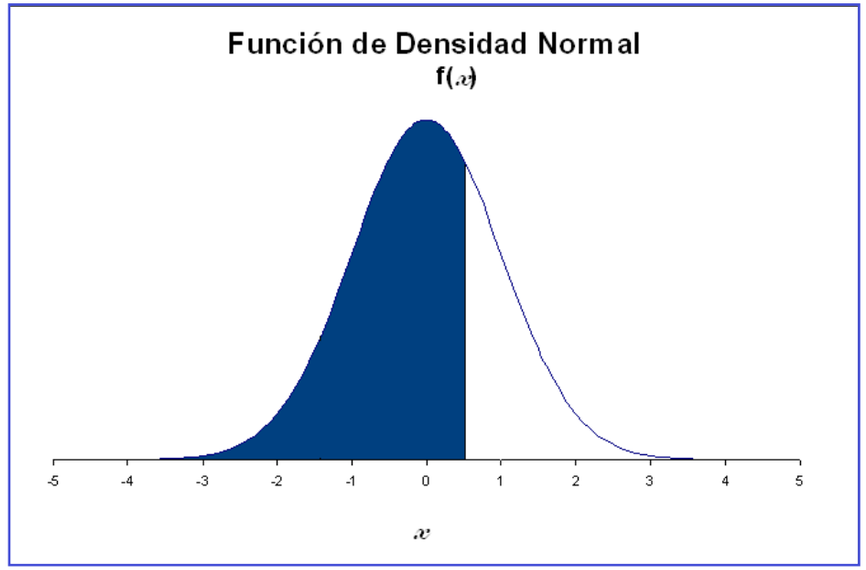
Una medida que expresa esta posición en términos de un porcentaje es llamado **porcentil**

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



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



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

3	38	0 2 5
6	39	1 7 8
12	40	0 0 1 3 7 9
13	41	5
(3)	42	4 6 8
14	43	3 6
12	44	1 3 5 6
8	45	0 7 9
5	46	2 2 3 8
1	47	
1	48	4

-los datos se ordenan de menor a mayor

-se encuentra la localización de los porcentiles:

$$l(P_\alpha) = \left(\frac{\alpha}{100}\right)(n+1)$$

-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 porcentiles 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

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

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

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

$\Rightarrow$ $Q_3 = P_{75} = 12.75$

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

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

Los cuantiles (deciles, quintiles, cuartiles) son muy útiles para comparar poblaciones de diferente tamaño

rohen

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 porcentil 75 y el porcentil 25:

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

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

rohen

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

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

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

desviación estándar

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

es sensible a valores extremos.

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

Creación de Intervalos:

con S y $\overline{X}$ se pueden formar intervalos de la forma $\overline{X} \pm kS$ 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:

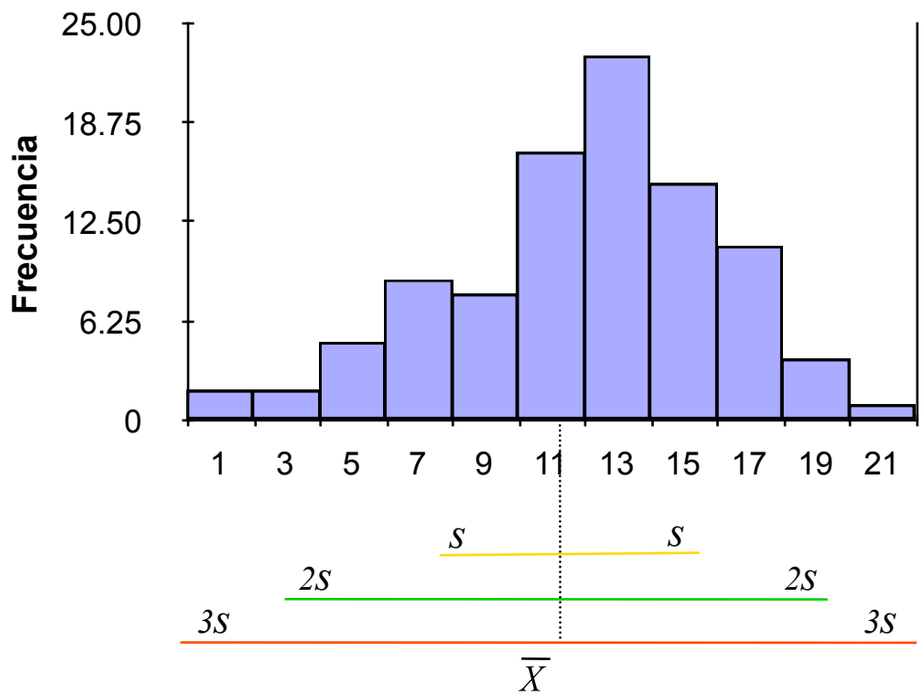
alrededor del 69% de las observaciones cae dentro de una desviación estándar de la media

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

alrededor del 99.7% de las observaciones cae dentro de tres desviaciones estándar de la media

Monóxido de Carbono en el humo de los cigarros

Intervalos alrededor de la media



$n = 372$

$\overline{X} = 11.66$

$S = 4.089$

$\overline{X} \pm S$	(7.57 , 15.75)	264 obs.	70.96%
$\overline{X} \pm 2S$	(3.48 , 19.84)	353 obs.	94.89%
$\overline{X} \pm 3S$	(0.0 , 23.93)	372 obs.	100.00%

rohen

medidas de dispersión...

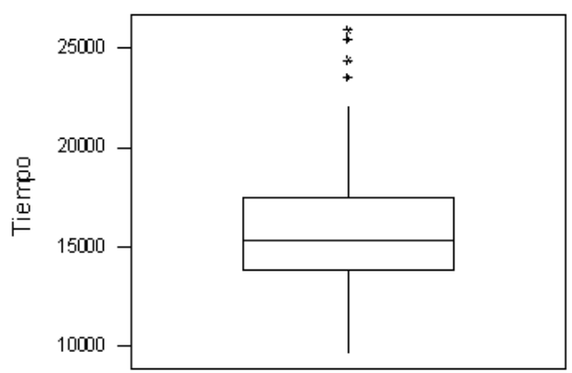
Coeficiente de Variación: es una medida de variación relativa y expresa la desviación estándar como un porcentaje de la media aritmética. Se obtiene como

$$CV = \frac{S}{\overline{X}} \text{x} 100$$

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

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 mas 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

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 f_{ij} pueden ser relativas o absolutas y nos dan una idea de qué tan frecuentemente se presentan simultáneamente ambos atributos en una población

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) ...

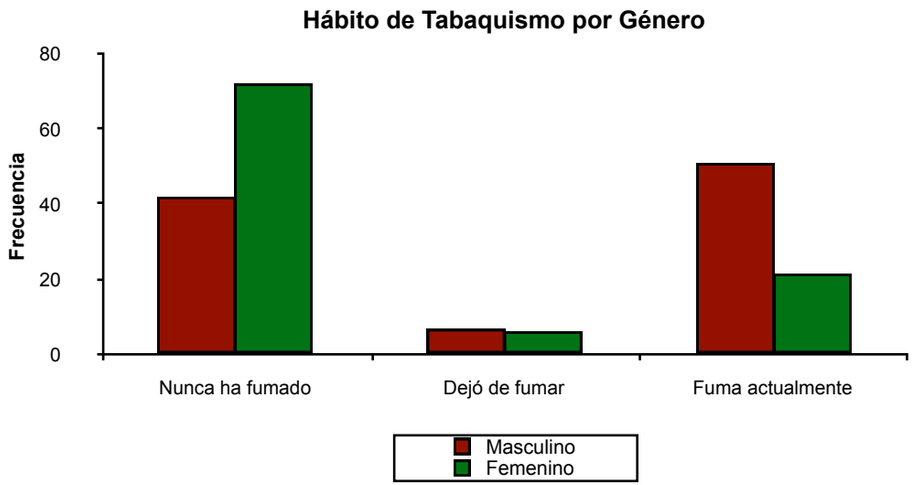
	Hábitos de tabaquismo (%)			
Género	**Nunca ha fumado**	**Dejó de fumar**	**Fuma actualmente**	**Total**
Masculino	28.5	4.6	34.3	**67.4**
Femenino	23.5	2.1	7.0	**32.6**
Total	**52.0**	**6.7**	**41.3**	**100**

comparación...

… y calculamos las frecuencias relativas condicionadas a género

Hábito de Tabaquismo condicionado 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**



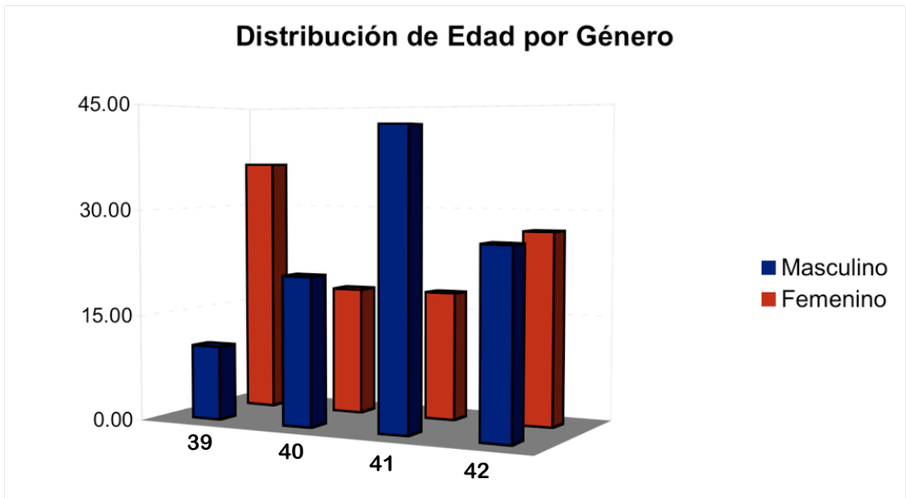
¿el hábito de tabaquismo difiere si se es hombre o mujer?

rohen

comparación...

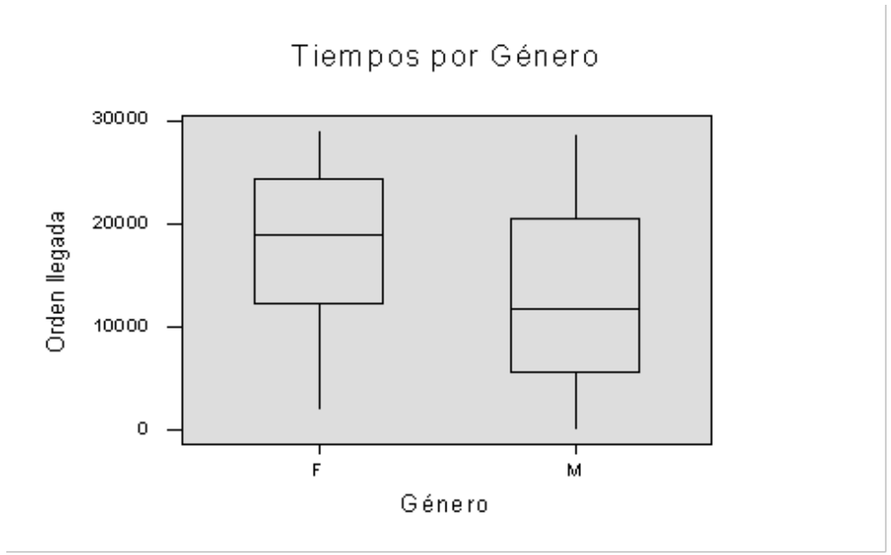
Si la variable a analizar es discreta se puede tratar como si fuera cualitativa.

	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?

En el caso de que la variable a analizar sea discreta o contínua podemos estar interesados en comparar tanto la localización como la dispersión entre las distribuciones de las subpoblaciones. Una manera de hacerlo es por medio de un diagrama esquemático



¿Qué género tiene mejores tiempos?

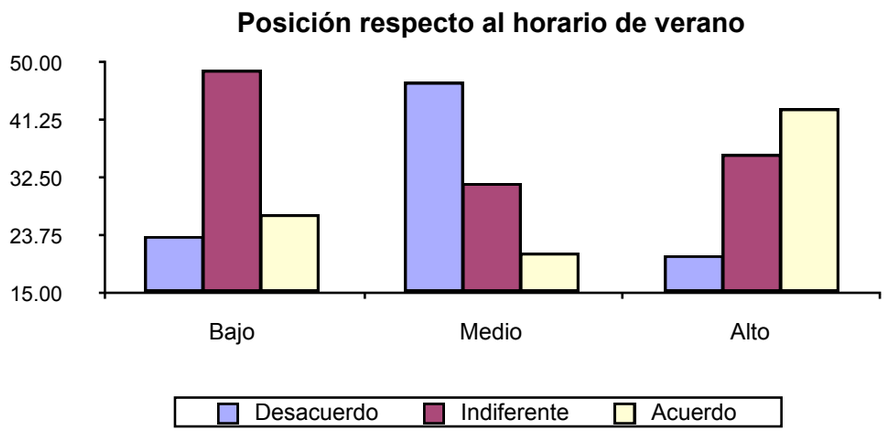
¿Cuál tiene mayor dispersión?

Muchas veces es importante saber si una variable influye sobre el comportaminto 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 ...

rohen

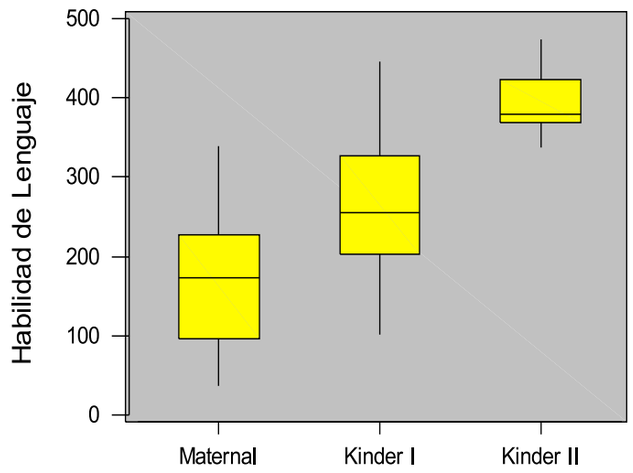
Una Variable Ordinal y otra Cuantitativa

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.

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

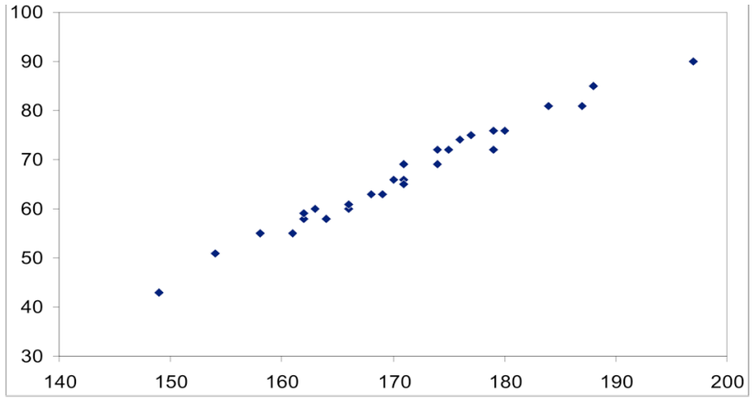
Para este caso el diagrama de dispersión es muy usado para asociar variables cuantitativas.

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

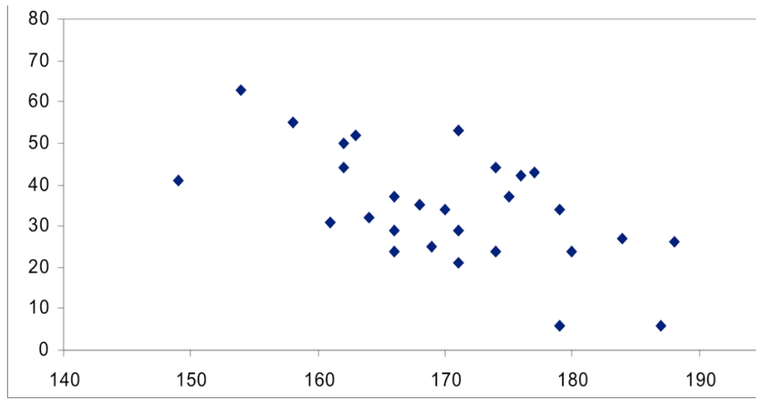
Una medida de asociación que complementa este diagrama es el coeficiente de correlación (medida de relación lineal entre las variables) obtenido como

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})/(n-1)}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2/(n-1)}} = \frac{S_{xy}}{S_x S_y}$$

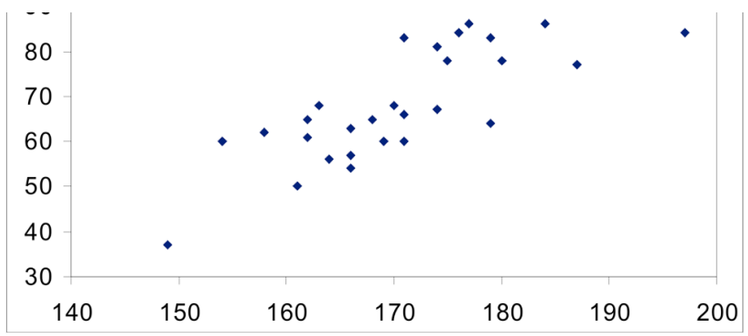
asociación ...



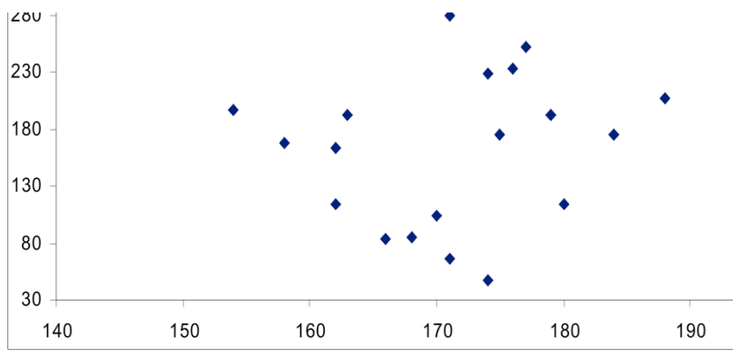
$r = 0.99$



$r = -0.7$



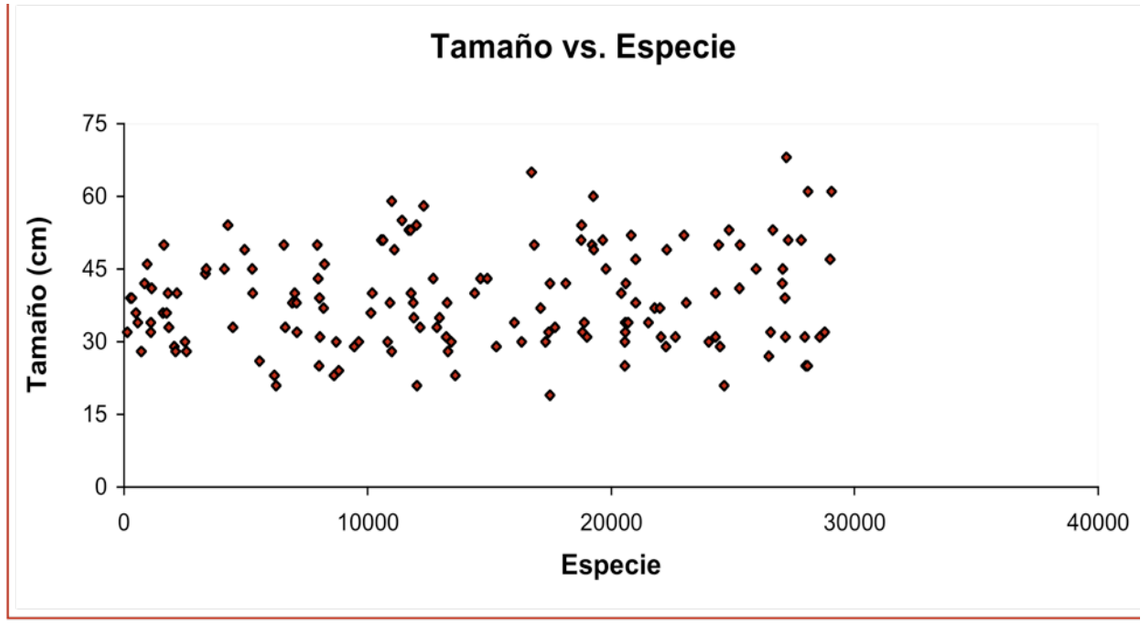
$r = 0.8$



$r = 0.1$

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

rohen



asociación ...

$r = 0.130$

¿Existe alguna relación lineal entre el tamaño y especie?

¿Confirma el valor de *r* esta relación?