Estadística Descriptiva

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**.

Objetivo

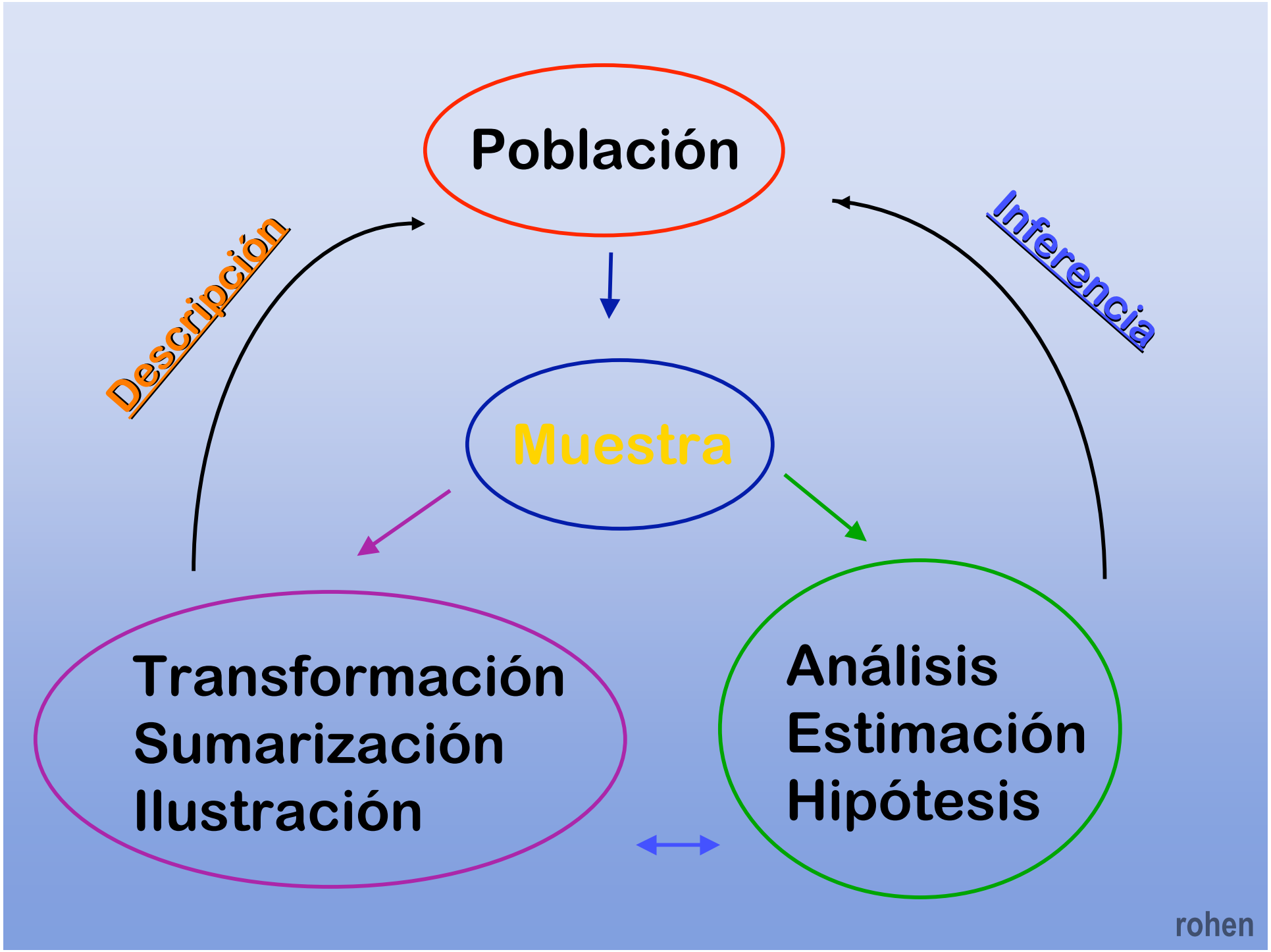
Poder concluir en base a la información contenida en una muestra diferentes aspectos de la realidad (etimación de parámetros)

Identificar problemas; contrastar, a través de los parámetros, diferentes aspectos de la población y tomar decisiones.

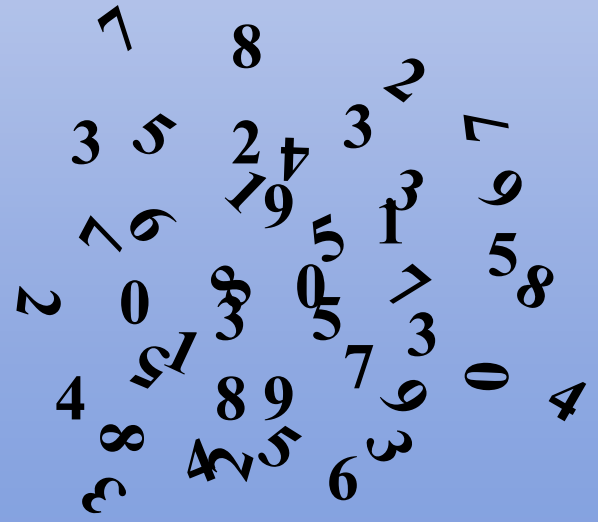
El uso de la probabilidad es una herramienta fundamental.

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**

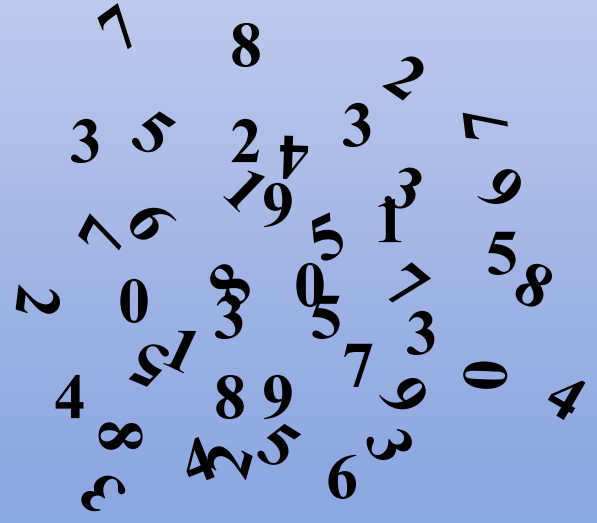

Población
Descripción
Inferencia
Muestra
Transformación
Sumarización
Ilustración
Análisis
Estimación
Hipótesis
rohen

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.



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

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**

¿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

tabla de distribución de frecuencias

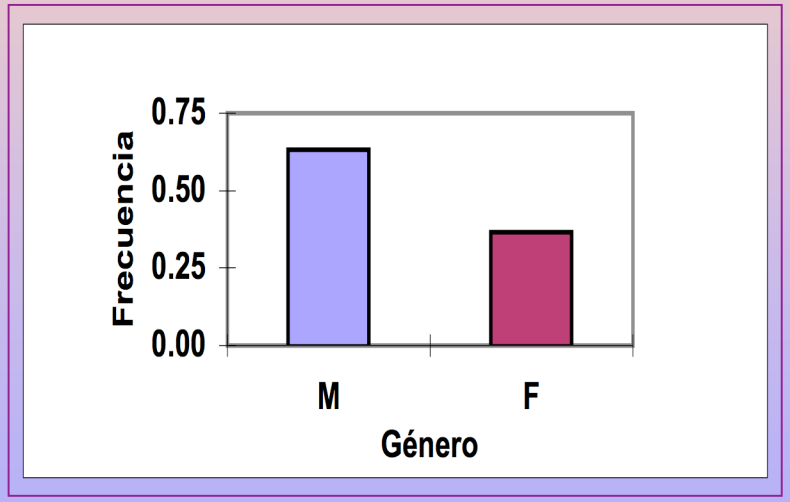
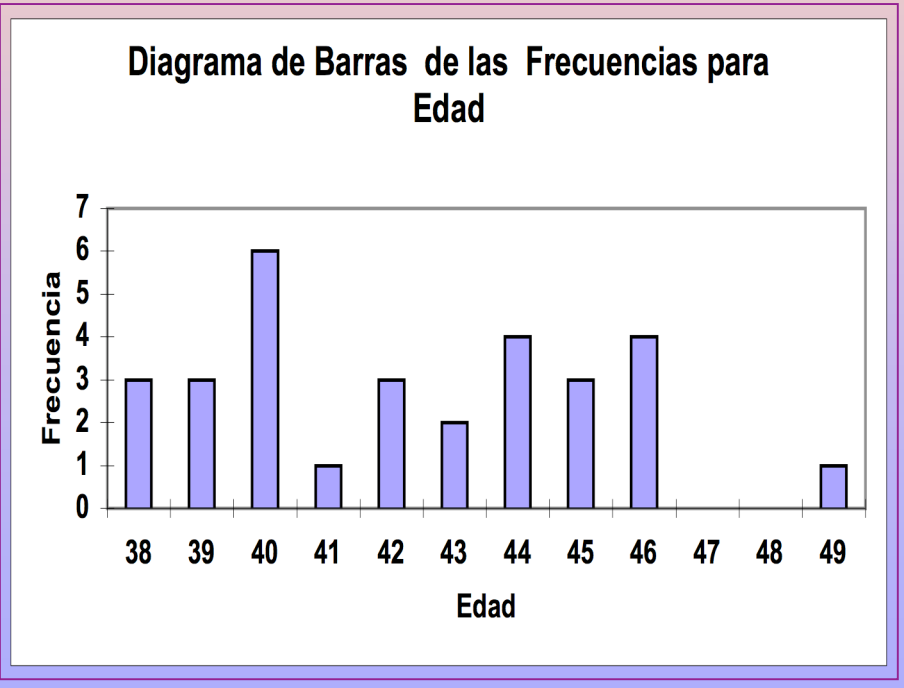


diagrama de barras

Gráfico de Pastel

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

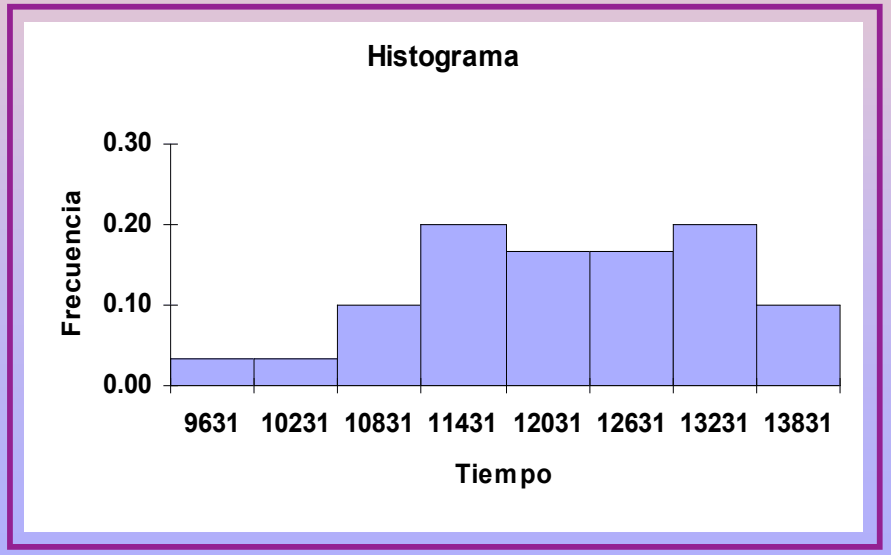
edad	*frecuencia*	*porcentaje*
38	3	0.10
39	3	0.10
40	6	0.20
41	1	0.03
42	3	0.10
43	2	0.07
44	4	0.13
45	3	0.10
46	4	0.13
47	0	0.00
48	0	0.00
49	1	0.03
Total	30	1.00



rohen

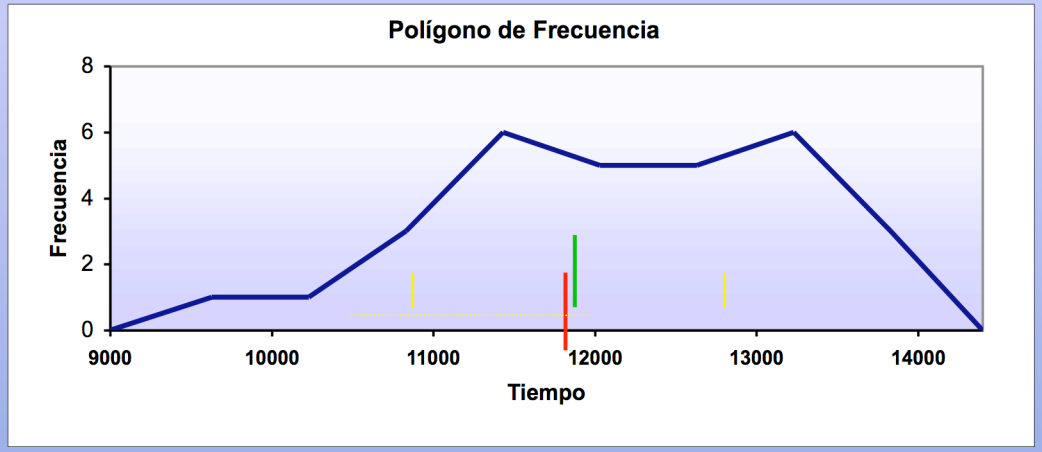
Para variables cuantitativas continuas las tablas de frecuencias se realizan con la creación de intervalos numéricos que formarán las diferentes clases.

tiempo	*frecuencia*	*porcentaje*
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
Total	30	1.00



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**

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

Mediana se localiza el valor central de los datos y se observa el valor que toma

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

Moda es el valor con la frecuencia más alta

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

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.

Creación de Intervalos:

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

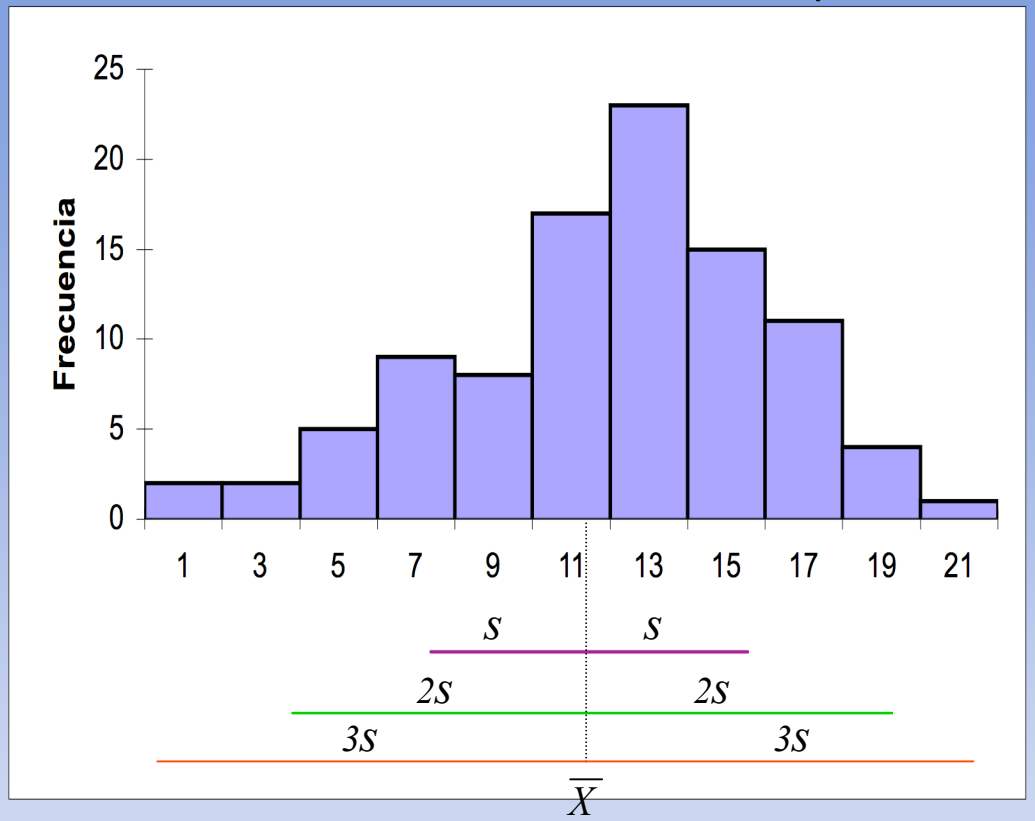
alrededor del 69% de las observaciones cae dentro de una desviación estándar de la media

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

alrededor del 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

$\overline{X} \mp S$	(7.57 , 15.75)	264 obs.	70.96%
$\overline{X} \mp 2S$	(3.48 , 19.84)	353 obs.	94.89%
$\overline{X} \mp 3S$	(-0.61 , 23.93)	372 obs.	100.00%

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?

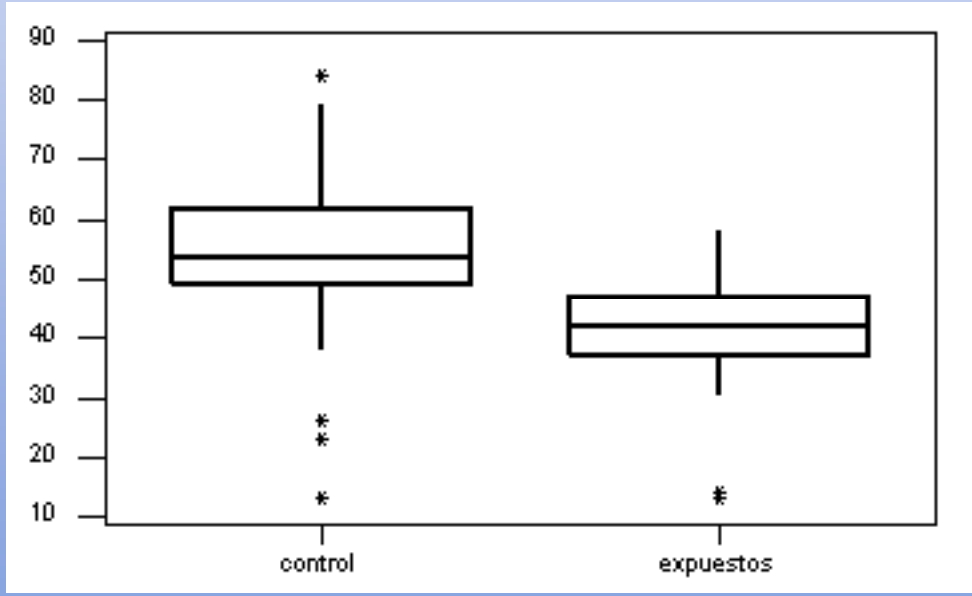
¿Qué tan grandes son esas diferencias?

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

Comparación entre dos poblaciones

Control: muestra no expuesta a plomo ambiental

Expuestos: muestra expuesta a plomo ambiental



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

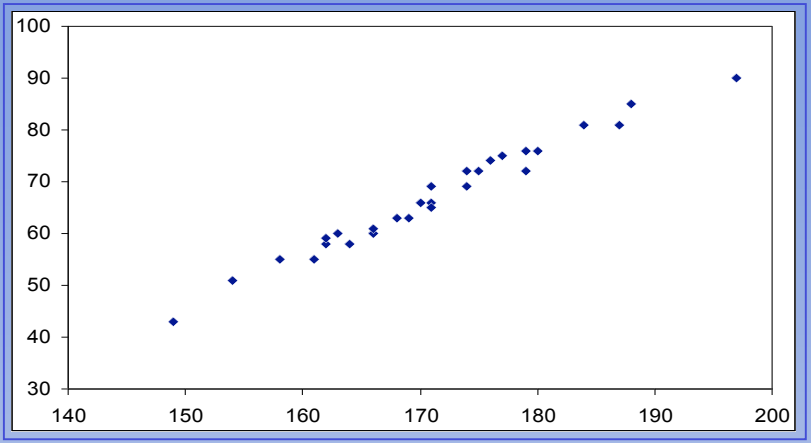
Para este caso el diagrama de dispersión es muy útil. Consiste en graficar parejas de valores (x_i, y_i) correspondientes a un solo individuo, sobre un plano cartesiano.

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

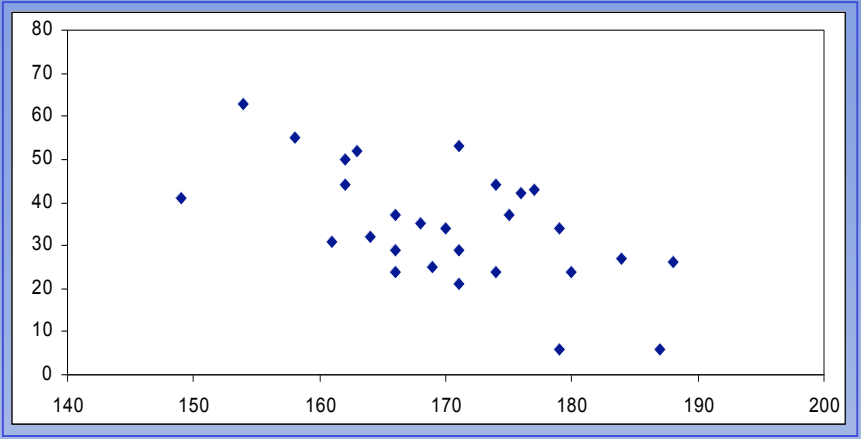
$$r(x,y) = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)/(n-1)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2/(n-1)}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2/(n-1)}} = \frac{S_{xy}}{S_x S_y}$$

rohen

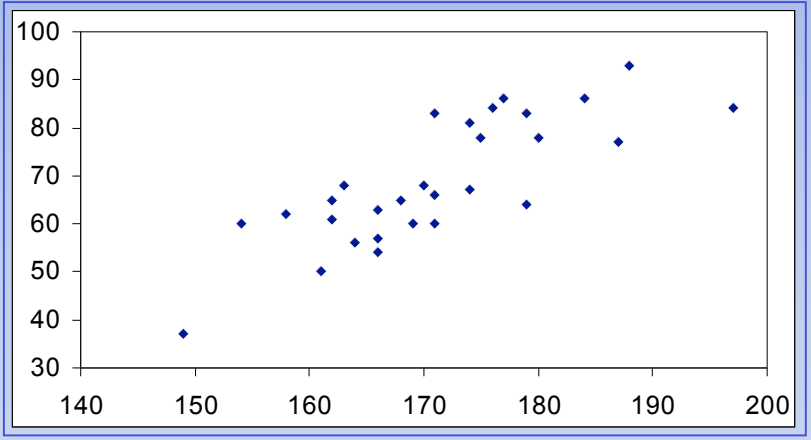
asociación ...



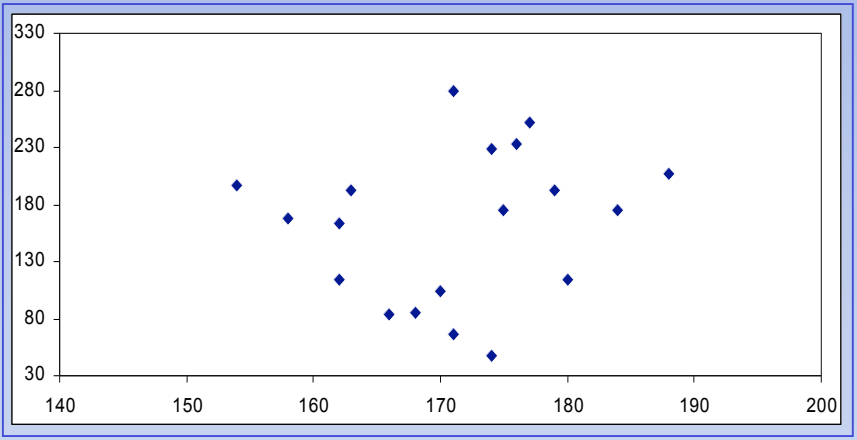
$r = 0.99$



$r = -0.7$



$r = 0.8$



$r = 0.1$

rohen

Referencias:

http://www.hrc.es/bioest/M_docente.html

Triola, M. **Elementary Statistics (9th Ed.)** Addison-Wesley Longman, 2000

Zar, Jerrold H.- Biostatistical Analysis.- 4rd ed.- Prentice Hall, Inc

Rosner, B.- Fundamentals of Biostatistics. 6th Ed. Brooks/Cole Publishing Co., 2006