Tema 1: ESTADÍSTICA DESCRIPTIVA

ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL

1 Conceptos básicos:

Población: Conjunto de individuos que presentan una o varias características en común. Esto es, el conjunto objeto de estudio. Por ejemplo:

- Todos los habitantes de una determinada ciudad.
- Las piezas fabricadas por una misma máquina.
- Todos los enfermos de una misma enfermedad.

¿Cómo podemos estudiar las poblaciones?

- mediante un **censo o estudio exhaustivo** que consiste en observar todos y cada uno de los individuos que integran la población.
- por **muestreo** que consiste en estudiar un subconjunto representativo de la población que se llama **muestra.** Se suele considar una muestra porque no siempre es posible estudiar exhaustivamente toda la población por motivos de tiempo, coste económico u otro tipo de dificultad.

Característica: Propiedad que deseamos observar entre los elementos de la población. Los diferentes estados o valores que presenta una característica se suelen llamar **modalidades de la característica**.

Atendiendo a la **característica** que se estudia, ésta puede clasificarse en:

- **característica cualitativa, categórica o atributo**: representa una cualidad del individuo.
- **característica cuantitativa**: aquellas característica que toman valores numéricos. A las características cuantitativas también se les llaman **variables estadísticas** y se dividen en:
  - **variables estadísticas discretas**: aquellas que toman un número finito o infinito numerable de valores.
  - **variables estadísticas continuas**: aquellas que toman un número no numerable de valores.

Ejemplo: Para los habitantes de una determinada ciudad se pueden estudiar las características: sexo, estado civil, profesión, edad, estatura, nivel de estudios,....

Una vez que hemos clasificado la característica que estudiamos, el siguiente paso es *ordenar y presentar los datos en tablas y gráficos* con el fin de resumir la información que contienen.

2 Distribución de frecuencias:

Supongamos que tenemos una muestra de tamaño n (que puede tomar m ($m \leq n$) modalidades o valores):

$$x_1, x_2, ..., x_n$$

Frecuencia absoluta de la modalidad x_i, n_i : número de veces que aparece x_i en la muestra. Se verifica que:

$$n_1 + n_2 + ... + n_m = n$$

Frecuencia relativa de la modalidad x_i, f_i : proporción de veces que aparece x_i en la muestra, esto es

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

Trivialmente se verifica que:

$$f_1 + f_2 + ... + f_m = 1$$

Se suele presentar en porcentaje sin más que multiplicar por 100, esto es $100 \times f_i\%$.

Los valores que toma la característica, junto con las frecuencias de dichos valores se suelen presentar en una tabla que se llama **distribución de frecuencias** o **tabla de frecuencias**:

valor, x_i	frecuencia absoluta, n_i	frecuencia relativa, f_i
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$		
x_m	n_m	f_m
totales	n	1

Ejemplo: Clasificación de las doradas del Mar Menor según su **sexo** (atributo)

sexo	frecuencia absoluta, n_i	frecuencia relativa, f_i
m	10	$10/27 \simeq 0.37$
h	17	$17/27 \simeq 0.63$
totales	27	1

Ejemplo: Clasificación de las doradas del Mar Menor según la **zona de captura** (atributo que está codificado en valores numéricos: 1,2,3,)

zona de captura	frecuencia absoluta, n_i	frecuencia relativa, f_i
1	9	0.333
2	9	0.333
3	9	0.333
totales	27	1

Ejemplo: Clasificación de las doradas del Mar Menor según la **longitud** (variable cuantitativa que toma "pocos" valores)

longitud	fr. absoluta, n_i	fr. relativa, f_i
1	6	0.22
2	5	0.19
3	6	0.22
4	4	0.15
5	6	0.22
totales	27	1

Ejemplo: Clasificación de las doradas del Mar Menor según el **peso** (variable cuantitativa que toma "muchos" valores). En estos casos, para un mejor y más cómodo manejo de datos se suele **agrupar** los datos en **intervalos de clase** (con igual amplitud):

- **Número de intevalos de clase** $= k$:
  - Es el número entero más próximo por exceso a $\sqrt{n}$. Además k debe verificar que: $5 \leq k \leq 20$.
  - O bien, mediante la *regla de Sturges*: $k = 1 + \log_2 n$
- **Amplitud** $= h \equiv \dfrac{x_{\max.} - x_{\min.}}{k}$
- Los datos deben de clasificarse sin ambigüedad en única clase $\Longrightarrow$

$$[e_0,e_1],\ (e_1,e_2], ..., (e_{k-1},e_k]$$

o bien,

$$[e_0,e_1),\ [e_1,e_2), ..., [e_{k-1},e_k]$$

Conceptos que aparecen con el tratamiento de los datos en clases:

Clase i-ésima: $(e_{i-1},e_i]$
Frecuencia absoluta de la clase i-ésima, n_i : número de observaciones que caen dentro de I_i.

Frecuencia relativa de la clase i-ésima, $f_i \equiv \dfrac{n_i}{n}$.

Frecuencia absoluta acumulada de la clase i-ésima $N_i :\equiv n_1 + n_2 + ... + n_i$

Frecuencia relativa acumulada la clase i-ésima, $F_i :\equiv f_1 + f_2 + ... + f_i = \dfrac{N_i}{n}$.

Marca de la clase i-ésima: $m_i \equiv \dfrac{e_{i-1} + e_i}{2}$, como representante de todas las observaciones de la clase i-ésima.

Y estos valores también se incorporan en la **tabla de frecuencias**:

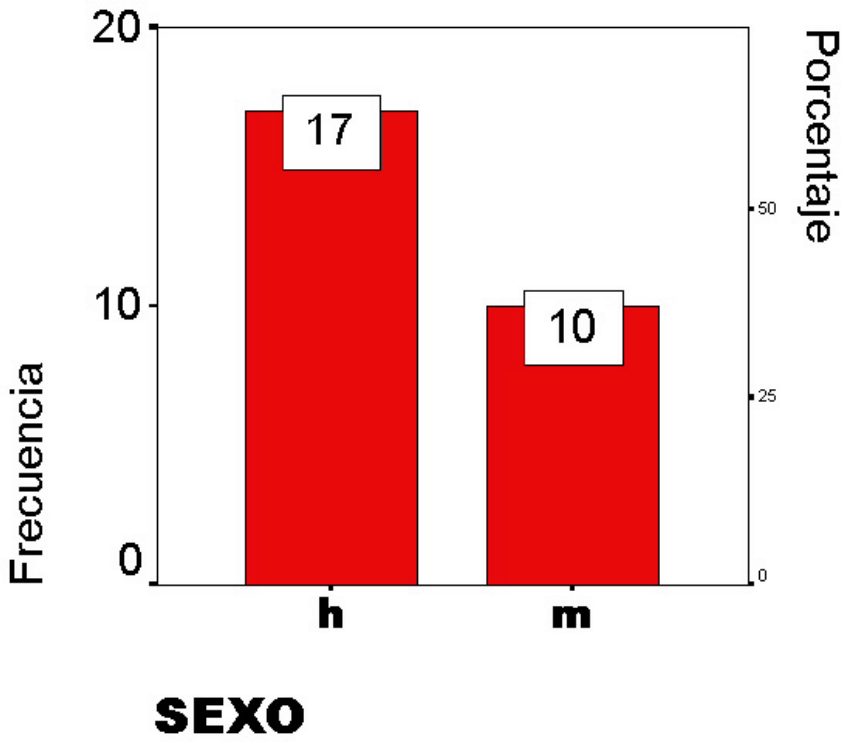
peso	marcas de clase, m_i	frecuencia absoluta, n_i	fr. absoluta acumulada, N_i	frecuencia relativa, f_i	fr. relativa acumulada, F_i
$[0.5, 1.92)$	1.21	6	6	0.22	0.22
$[1.92, 3.34)$	2.63	5	11	0.19	0.41
$[3.34, 4.76)$	4.05	2	13	0.07	0.48
$[4.76, 6.18)$	5.47	4	17	0.15	0.63
$[6.18, 7.6)$	6.89	4	21	0.15	0.78
$[7.6, 9.02]$	8.31	6	27	0.22	1
totales	–	27	–	1	–

3 Representaciones gráficas

Las represntaciones gráficas proporcinan una síntesis visual de la distribución de frecuencias. Las gráficas más utilizadas son las siguientes:

Características cualitativas:

- **Diagrama de Pareto:** En el eje de abscisas se asocia a cada modalidad un rectángulo de base constante y de altura proporcional a la frecuencia correspondiente (las modalidades se suelen disponer en orden decreciente según su frecuencia de aparición).



- **Diagrama de sectores:** A cada modalidad se le asigna un sector circular de amplitud

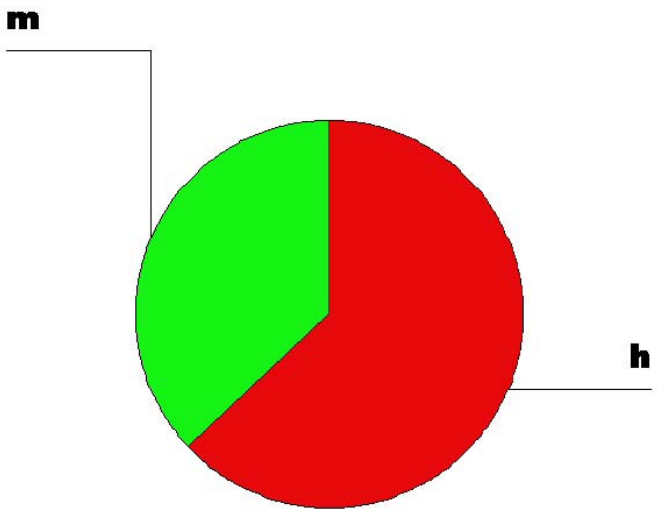
w_i proporcional a su frecuencia relativa$\Rightarrow$

$$\begin{array}{r} w_i = 360^o \times f_i \\ \text{o bien} \qquad \text{, para cada modalidad } i. \\ w_i = 2\pi \times f_i \end{array}$$

El área de los sectores, para la variable **sexo**, sería como sigue:

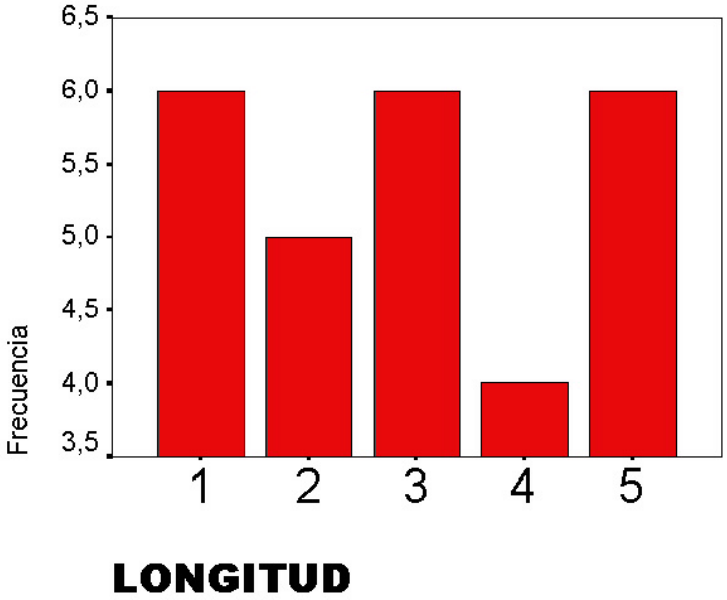
$$\mathbf{h} = 360^o \times 0.63 = 226.67^o$$

$$\mathbf{m} = 360^o \times 0.37 = 133.33^o$$



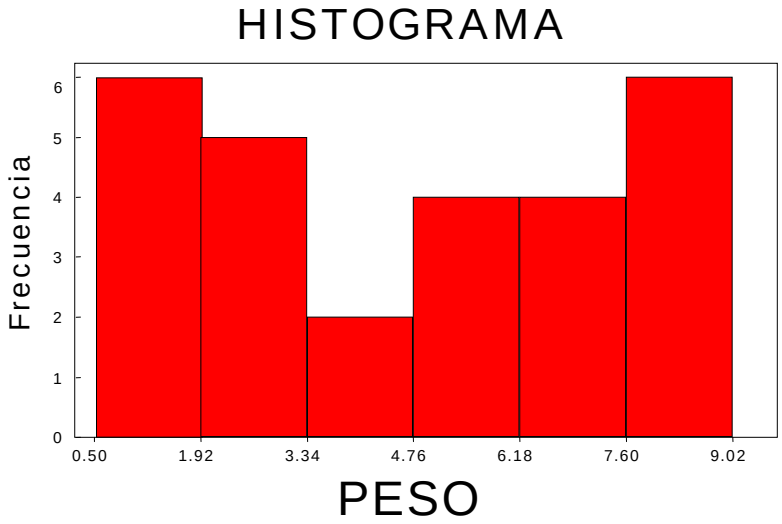
Variables estadísticas discretas o continuas no agrupadas en intervalos de clase:

- **Diagrama de barras:** En el eje de abscisas se representan los valores que tome la variables y sobre cada uno de ellos se dibuja una barra de altura igual o proporcional a la frecuencia absoluta (o relativa) correspondiente.



Variables estadísticas continuas agrupadas en intervalos de clase:

- **Histograma:** En el eje de abscisas se representan los extremos de los intervalos de clase de la variable y sobre cada uno se construye un rectángulo de altura igual a la frecuencia absoluta (cuando todos los intervalos son de la misma amplitud). En otro caso, la altura de cada rectángulo es igual a la frecuencia absoluta de dicha clase dividido por la amplitud de dicha clase.



4 Características que debemos identificar ante un conjunto de datos

NOTA: A partir de ahora, salvo que se diga lo contrario, nos centraremos en **datos cuantitativos no agrupados**; esto es,

$$\text{muestra de tamaño } \mathbf{n}\text{: } x_1, x_2, ..., x_n$$

A continuación vamos a definir medidas numéricas que describen los aspectos más relevantes de la distribución de frecuencias. Estas características se clasifican según la información que tratan de resumir en:

- **Medidas de posición o localización:** describen cómo se comportan globalmente los datos y localizan la distribución de frecuencias.

- **Medidas de dispersión:** miden la variabilidad de los datos entre sí o respecto de una medida de centralización.

- **Medidas de forma:** informan sobre la asimetría de la distribución (**medidas de asimetría**) y sobre la concentración de las observaciones en torno a la zona central (**medidas de apuntamiento o kurtosis**).

4.1 Medidas de Posición

4.1.1 Medidas de Posición Central

Indican dónde se sitúa la zona central de la distribución de frecuencias.

- **Media aritmética:** .

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

Si los datos están agrupados en intervalos de clase, se toman como observaciones las marcas de clase, esto es:

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{k} m_i \times n_i}$$

donde m_i es la marca de clase del intervalo i y n_i es la frecuencia absoluta del intervalo i, con $i = 1, ..., k$.

Es la medida de centralización más utilizada ya que es el centro de gravedad del conjunto de datos, sin embargo, es muy sensible a valores extremos lo que la hace poco representativa. Además es sensible a cambios de escala y traslaciones. Veamos qué significa estos comentarios con ejemplos:

Además es sensible a cambios de escala y traslaciones

- **Mediana, M_e:** Es aquel valor que divide en dos partes iguales la distribución de frecuencias. Esto es,

$$\underbrace{x_{(1)}, x_{(2)}, ...,}_{50\% \text{ de los datos}} \boxed{M_e}, \underbrace{....., x_{(n)}}_{50\% \text{ de los datos}}$$

Forma de calcularla para datos no agrupados:

  - Ordenamos los datos de menor a mayor:

$$x_{(1)}, x_{(2)}, ..., x_{(n)}$$

  - Entonces,

$$\boxed{M_e = \begin{cases} x_{(\frac{n+1}{2})} & \text{(dato central)} & \text{si } n \text{ es IMPAR} \\ \frac{x_{(\frac{n}{2})} + x_{((\frac{n}{2})+1)}}{2} & \text{(promedio de los dos datos centrales)} & \text{si } n \text{ es PAR} \end{cases}}$$

Si los datos están agrupados en intervalos de clase, hablamos de **intervalo mediano** como el primer intervalo de clase cuya frecuencia absoluta acumulada es $\geq \frac{n}{2}$. Si necesitamos un valor numérico para la mediana, podemos tomar la marca de clase del intervalo mediano.

- **Moda, M_o**: Valor de la variable que presenta mayor frecuencia. Existe para datos cualitativos.

Si los datos están agrupados en intervalos de clase, hablamos de **intervalo modal** como aquel que presenta mayor frecuencia. Si necesitamos un valor numérico para la moda, podemos tomar la marca de clase del intervalo modal.

Notar que puede existir más de una moda, así como no existir.

Puede no situarse en el centro de la distribución de frecuencias.

4.1.2 Medidas de posición no central

Proporcionan información sobre la estructura interna de los datos.

- **Cuantiles**: Se define el cuantil de valor α $(0 < \alpha < 1)$ de una distibución de frecuencias como el valor C_α que deja a su izquierda el $100\alpha\%$ de las observaciones y a la derecha el $100(1-\alpha)\%$ restantes de las observaciones.

Casos particulares de cuantiles para valores concretos de α:

- **Percentiles**: $P_1, P_2, ..., P_{99}$, para $\alpha = 0.01,\ 0.02, ..., 0.99$, respectivamente.
- **Deciles**: $D_1, D_2, ..., D_9$, para $\alpha = 0.1,\ 0.2, ..., 0.9$, respectivamente.
- **Cuartiles**: Q_1, Q_2, Q_3, para $\alpha = 0.25,\ 0.50,\ 0.75$, respectivamente.

  Forma de calcular los cuartiles para datos no agrupados:

  - Ordenamos los datos de menor a mayor:

    $$x_{(1)}, x_{(2)}, ..., x_{(n)}$$

    y determinamos la M_e.
  - Entonces Q_1 es la mediana del conjunto de datos que hay a la izquierda de la M_e (excluida la M_e) y Q_3 es la mediana del conjunto de datos que hay a la derecha de la M_e (excluida la M_e).

Veámos cómo se calcula la M_e y los cuartiles con unos ejemplos:
a) Consideremos una muestra de tamaño 11 que toma los valores siguientes:

5 5.3 6.1 7 7.2 7.5 7.8 8.1 8.6 8.9 9

b) Consideremos una muestra de tamaño 10 que toma los valores siguientes:

5 5.3 6.1 7 7.2 7.5 7.8 8.1 8.6 8.9

4.2 Medidas de Dispersión

4.2.1 Medidas de Dispersión Absoluta

- **Rango o recorrido**: Amplitud del intevalo donde se encuentran distribuidas todas las observaciones.

$$\boxed{R = x_{\max.} - x_{\min}}$$

Es muy sensible a valores extremos.

- **Rango Intercuartílico**: Amplitud del intevalo donde se encuentran distribuidas el 50% de las observaciones.

$$\boxed{RIQ = Q_3 - Q_1}$$

- **Varianza**: Medida de dispersión asociada la media aritmética y se define por:

$$\boxed{s_X^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}(\overline{x^2} - \overline{x}^2)}$$

Si los datos están agrupados en intervalos de clase, se toman como observaciones las marcas de clase, esto es:

$$\boxed{s_X^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(m_i - \overline{x})^2 \times n_i = \frac{n}{n-1}(\overline{x^2} - \overline{x}^2)}$$

donde m_i es la marca de clase del intervalo i y n_i es la frecuencia absoluta del intervalo i, con $i = 1, ..., k$.

A la raíz cuadarada positiva de la varianza se le denomina **desviación típica** o **estándar** y se denota por s o s_X.

La varianza viene dada en unidades al cuadrado, mientras que la desviación típica viene en las mismas unidades físicas de los datos.

Como la media, es muy sensible a valores extremos. Además es sensible a cambios de escala pero no a traslaciones. Veámoslo con ejemplos:

4.2.2 Medidas de Dispersión Relativa

Medida adimensional, esto es, no tiene unidades cuyo objetivo es comparar la dispersión de distribuciones que se miden en distintas unidades. Normalmente este concepto tiene sentido pleno para magnitudes, es decir, para variables no negativas.

- **Coeficiente de variación de Pearson**:

$$\boxed{CV = \frac{s}{\overline{x}}}$$

  Interpretación: Mide la representatividad de la media como medida que resume toda la información de la variable cuando comparamos dos o más distribuciones de frecuencias. Cuanto menor sea el valor de este coeficiente mayor representatividad de la media, ya que significa que los datos están más agrupados entorno a su valor medio.

Es sensible a traslaciones pero no a cambios de escala.

4.3 Medidas de Forma

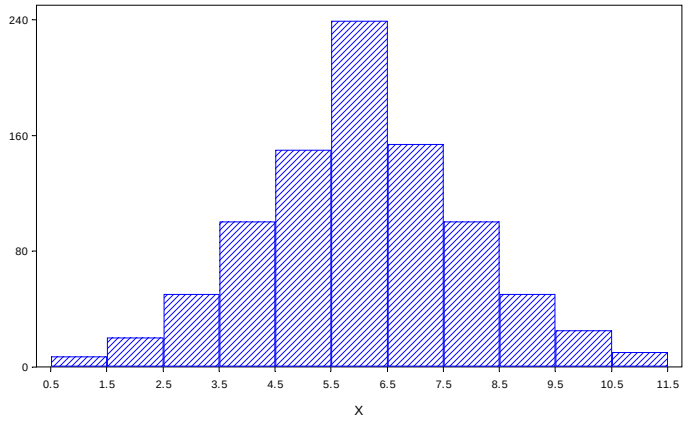
Nos proporcionan información sobre el perfil de la distribución de frecuencias.

Los atributos relacionados con la forma los vamos a establecer de manera aproximada observando la correspondiente representación gráfica de los datos y éstos son:

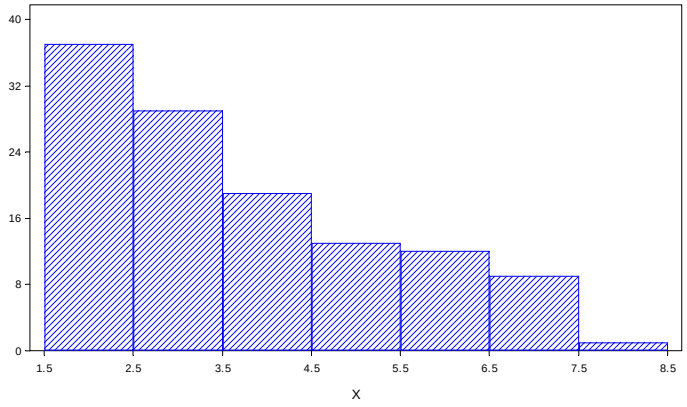
- **Asimetría (Skew):**
  Coeficiente de asimetría de Fisher:

$$\boxed{CA_f = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}}$$

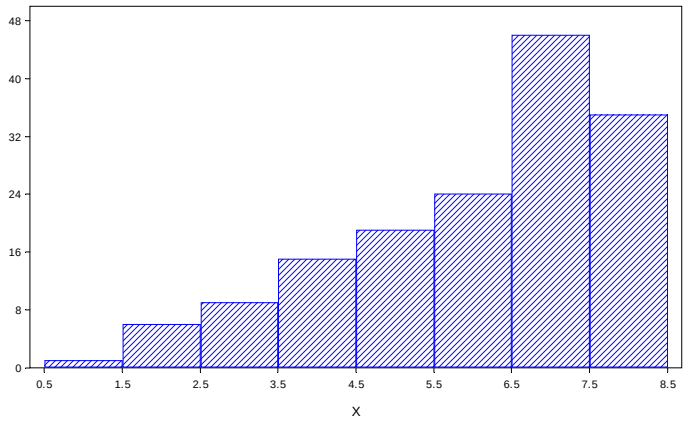
  Una **distribución de frecuencias** es **simétrica** si su correspondiente representación gráfica (diagrama de barra o histograma) es simétrica respecto de un eje vertical. ($CA_f \simeq 0$)



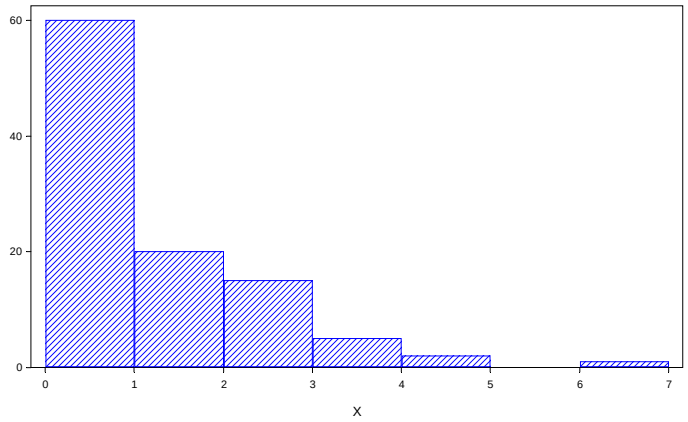
Distribución asimétrica a la derecha si las observaciones están desplazadas hacia la derecha. $(CA_f > 0)$



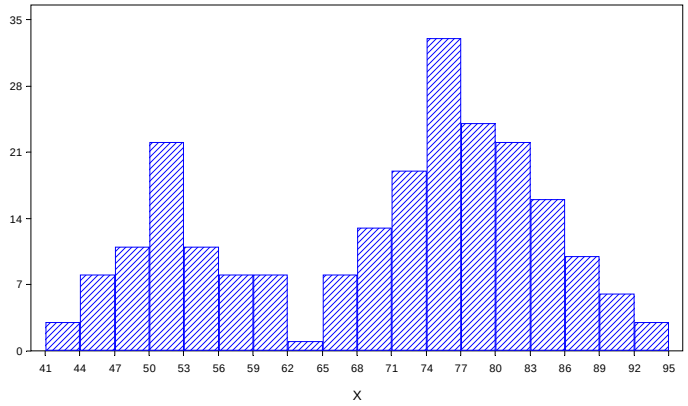
Distribución asimétrica a la izquierda si las observaciones están desplazadas hacia la izquierda. $(CA_f < 0)$



- **Existencia o no de datos atípicos o anómalos** que son datos que a simple vista se encuentran muy alejados del resto de los datos (en el apartado siguiente veremos una forma para identificarlos). Estos valores pueden ser debidos a un error de medida o de transcripción de los datos o corresponde a un verdadero valor de la variable.
  Distribución asimétrica con existencia de valores extremos



- **Unimodal, bimodal o multimodal**. Cuando existe más de una moda, en ocasiones es posible identificar aproximadamente la **existencia de subgrupos de datos**. En estos casos, las medidas resumen globalmente pueden llegar a ser engañosas, por lo que siempre que sea posible, conviene explorar las características en dichos subgrupos de datos.
  Por ejempo los ingresos familiares, se espera globalmente dos grupos de datos según si una o dos personas de la unidad familiar trabajan.



5 Diagrama de caja y bigotes

Es un resumen gráfico que permite visualizar, para un conjunto de datos, la tendencia central, la dispersión y la presencia de valores extremos. Otra característica de este tipo de gráfico es que nos da información sobre la asimetría del conjunto de datos.
La mayor utilidad de los diagramas de caja y bigotes es para comparar dos o más conjuntos de datos.

Un diagrama de caja y bigotes se construye de la siguiente manera:

1. Ordenamos los datos de la muestra de menor a mayor y obtenemos los tres cuartiles.
2. Dibujamos un rectángulo cuyos extremos son Q_1 y Q_3 y dividimos el rectángulo por un segmento central a la altura de la M_e
3. Calculamos los límites superior e inferior admisibles que nos servirán para identificar los datos atípicos. Éstos son:

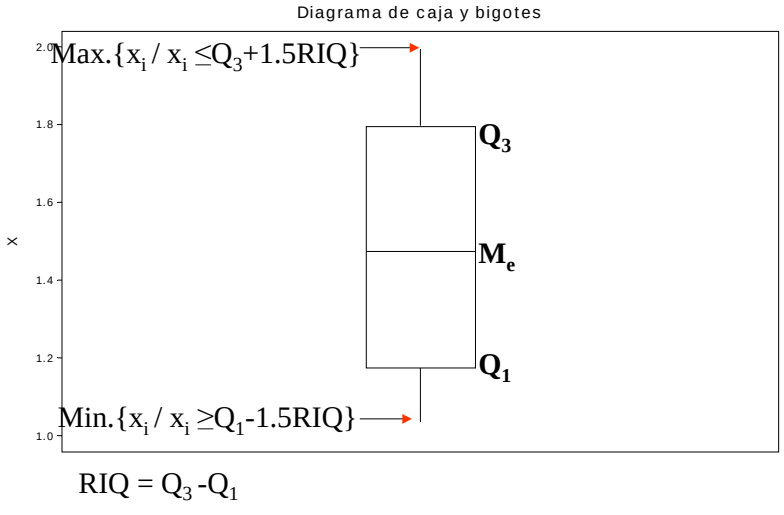
$$L_{SUPERIOR} = Q_3 + 1.5 \times RIQ$$

$$L_{INFERIOR} = Q_1 - 1.5 \times RIQ$$

Clasificamos como **datos atípicos** aquellas observaciones situadas fuera del intervalo

$$[Q_1 - 1.5 \times RIQ, Q_3 + 1.5 \times RIQ]$$

Los segmentos $1.5 \times RIQ$ (llamados **bigotes**) se acortan hasta: el dato del conjunto inmediatamente superior a $Q_1 - 1.5 \times RIQ$ para el bigote inferior, esto es, hasta el $mín.x_i$ tal que $x_i \geq Q_1 - 1.5 \times RIQ$; y el dato inmediatamente anterior a $Q_3 + 1.5 \times RIQ$ para el bigote superior, esto es, hasta el $máx.x_i$ tal que $x_i \leq Q_3 + 1.5 \times RIQ$.



Cuando existen datos atípicos en el conjunto, representamos los datos atípicos como puntos

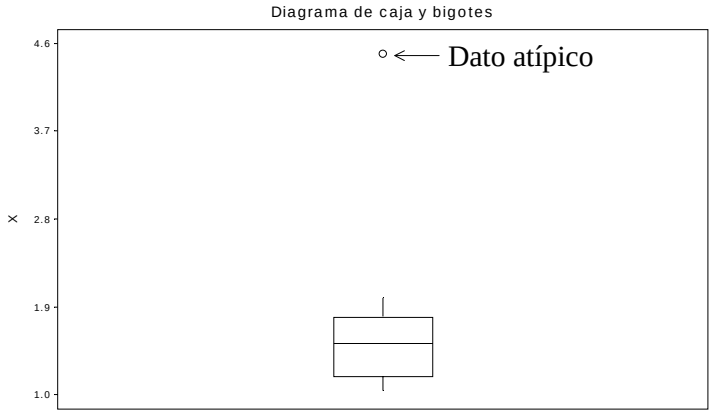
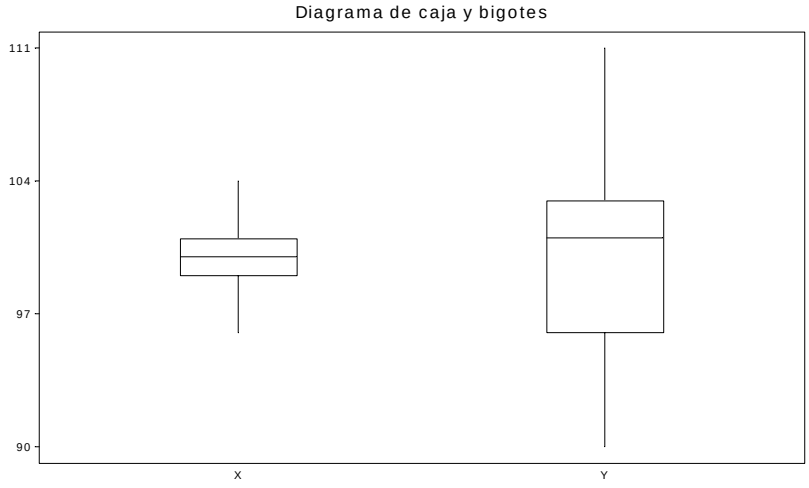
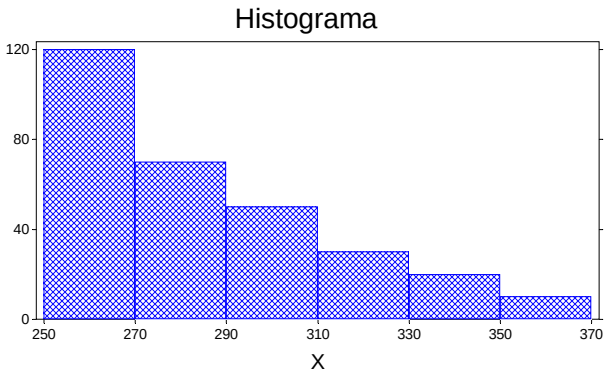
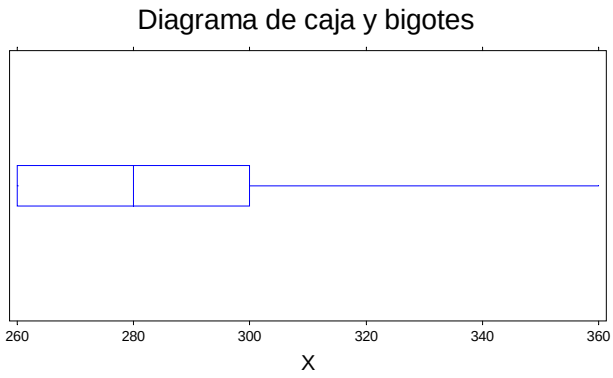


Diagrama de caja y bigote múltiple, cuando lo utilizamos para comparar dos o más conjuntos de datos:

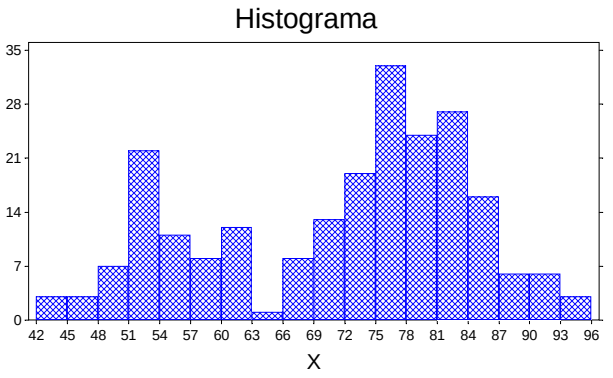


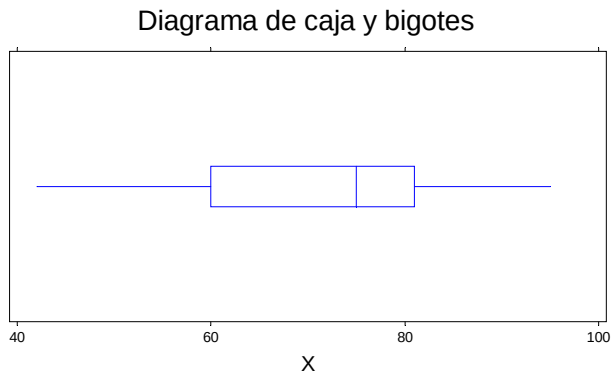
Notar que la información que nos proporciona el histograma en poblaciones unimodales también nos la proporcina el diagrama de caja y bigotes:





Sin embargo, esto no es cierto para poblaciones bimodales:





ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL

Dada una población, es muy frecuente estudiar dos o más características entre los individuos que la integran. Estas dos características se denotan por X e Y. La variable que representa conjuntamente estas dos características se denota por (X, Y) y se denomina por **variable estadística bidimensional**

$$\boxed{\begin{array}{l} X: \ x_1, x_2, ..., x_i, ..., x_k \\ \\ Y: \ y_1, y_2, ..., y_j, ..., y_p \end{array}}$$

Cada individuo de la población presentará un valor x_i de X y un valor y_j de Y.

Siempre podemos empezar estudiando cada característica por separado (como lo hemos hecho en el tema anterior), pero *el objetivo es el estudio conjunto de las características para determinar si existe o no dependencia entre ellas y, en caso afirmativo, determinar de qué grado.*

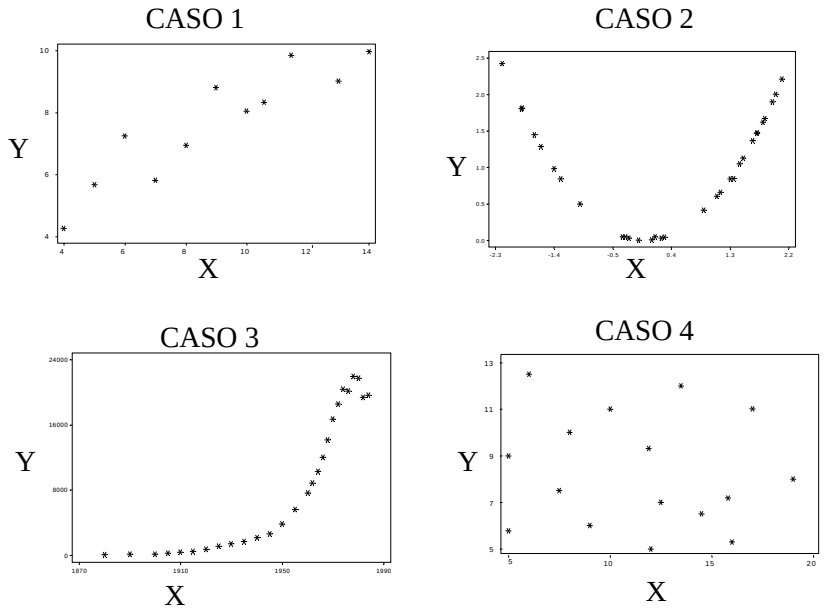
1 Regresión y Correlación

Supongamos que tenemos una muestra de tamaño n de una **variable estadística bidimensional** cuantitativa no agrupada (X, Y) formada por los pares

$$\boxed{\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}}$$

y queremos estudiar su **dependencia estadística**; esto es, por una parte, encontrar una función que exprese lo mejor posible el tipo de relación de una variable con la otra (**Teoría de la regresión**), y, por otra, estimar el grado de dependencia mutua entre las variables (**Teoría de la Correlación**).

El primer paso para estudiar el comportamiento de dos variables es el llamado **diagrama de dispersión** o **nube de puntos** que es una representación gráfica de los valores (x_i, y_i) en el plano.



Aspectos que debemos de identificar en la nube de puntos:

- Si valores altos de una de las variables están asociados con valores altos de la otra de las variables (dependencia positiva), o por el contrario, con valores bajos de la otra de las variables (dependencia negativa).

- Si la evolución de una de las variables en función de la otra sigue un patrón reconocible: una recta, una parábola, una exponencial, etc...

1.1 Planteamiento del problema de regresión

Tenemos dos variables

X: **Variable independiente** o **explicativa**

Y: **Variable dependiente** o **respuesta**

Los datos vienen dados en forma de pares:

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

Objetivo: Pretendemos escribir la variable Y en términos de la variable X mediante una expresión de la forma

$$Y = f(X)$$

donde f sea la "mejor" función que relaciona X con Y.

Por conocimientos previos sobre el fenómeno que estudiamos o con la nube de puntos tenemos una idea de la relación que liga a las variables X e Y.

1.1.1 Criterio de los mínimos cuadrados

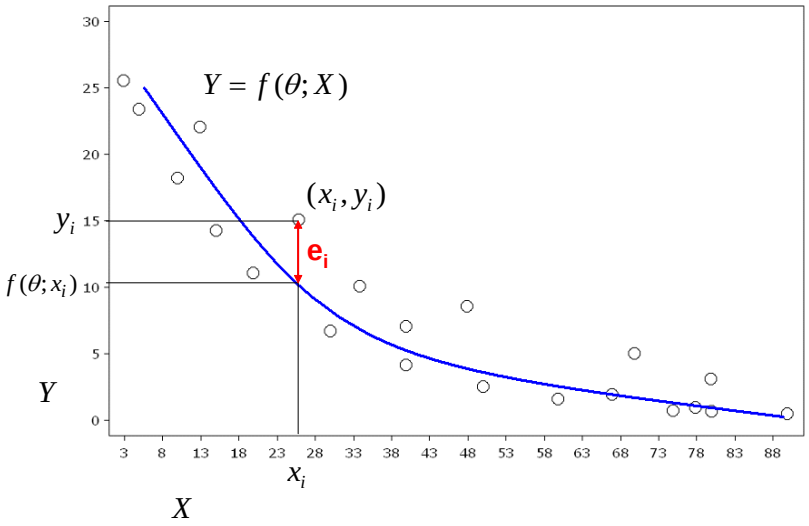
Queremos escribir $Y = f(X)$, donde $f(X)$ dependerá de unos parámetros.
Por ejemplo: una recta $Y = b + aX$, una parábola $Y = c + bX + aX^2$, una exponencial $Y = \beta e^{\alpha X}$,....

Luego formalmente tendremos $Y = f(\theta; X)$, con θ el vector de parámetros. La forma de la curva está fijada pero tenemos que determinar los parámetros que determinen la "mejor" curva, dentro de la familia paramétrica elegida, que relacione X con Y.

Definimos el **residuo** como

$$e_i = y_i - f(\theta; x_i) \text{ para cada } i = 1, 2, ..., n$$

que es el error que cometemos cuando aproximamos y_i por el correspondiente valor de x_i calculado sobre la curva. Gráficamente sería:



Criterio de ajuste:

$$\text{Elegimos } \theta \text{ tal que minimice } \underbrace{\sum_{i=1}^{n} [y_i - f(\theta; x_i)]^2}_{\mathbf{SC}(\boldsymbol{\theta})}$$

donde $SC(\theta)$ es la **suma de cuadrados de los residuos**.

No siempre es posible encontrar la forma explícita de este mínimo y tendremos que recurrir a métodos numéricos para obtenerlo. Nosotros nos vamos a centrar en una clase de familias paramétricas para las cuales es posible determinar explícitamente la forma de $f(\theta; X)$.

Recta de regresión mínimo cuadrática

Supongamos que la nube de puntos obtenida tiende a condensarse aproximadamente a lo largo de una recta$\Longrightarrow$ Esto nos indica que existe una dependencia lineal entre las características.

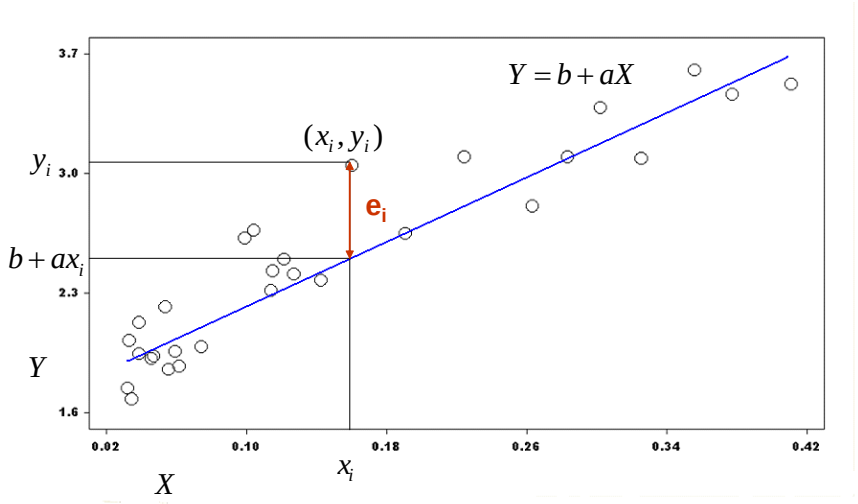
Para determinar la "mejor" recta que aproxima a la nube de puntos tenemos que encontrar dos parámetros: **a** (**pendiente**) y **b** (**ordenada en el origen**) de manera que ambas características se relacionen por

$$Y = b + aX$$

En este caso, los **residuos** tendrán la forma

$$e_i = y_i - (b + ax_i) \text{ para cada } i = 1, 2, ..., n$$

que es el error que cometemos cuando aproximamos y_i por el correspondiente valor de x_i calculado sobre la recta de ecuación $Y = b + aX$. Gráficamente sería:



Criterio de ajuste:

Elegimos **a** y **b** tal que minimice $\underbrace{\sum_{i=1}^{n} [y_i - (b + ax_i)]^2}_{\mathbf{SC(a,b)}}$

donde $SC(a, b)$ es la **suma de cuadrados de los residuos**.
Los posibles mínimos son las soluciones del sistema

$$\left.\begin{array}{l} \frac{\partial}{\partial a} SC(a,b) = 0 \\ \frac{\partial}{\partial b} SC(a,b) = 0 \end{array}\right\} \Rightarrow \left.\begin{array}{l} 2\sum_{i=1}^{n} (y_i - b - ax_i)(-x_i) = 0 \\ 2\sum_{i=1}^{n} (y_i - b - ax_i)(-1) = 0 \end{array}\right\}$$

que después de manipularlo algebraicamente obtenemos que:

$$\begin{aligned} \widehat{a} &= \frac{\overline{xy} - \overline{x} \times \overline{y}}{\overline{x^2} - \overline{x}^2} \\ \widehat{b} &= \overline{y} - \widehat{a} \times \overline{x} \end{aligned}$$

Definimos la **covarianza** entre X e Y por:

$$\boxed{s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{n}{n-1}(\overline{x \cdot y} - \overline{x} \cdot \overline{y})}$$

que es una medida descriptiva conjunta que nos proporciona información sobre la relación lineal que existe entre ambas características.

Con lo cual, la **pendiente** de la recta la podemos escribir como:

$$\boxed{\widehat{a} = \frac{\frac{n}{n-1}\left(\overline{xy} - \overline{x} \times \overline{y}\right)}{\frac{n}{n-1}\left(\overline{x^2} - \overline{x}^2\right)} = \frac{s_{xy}}{s_x^2}}$$

y la **ordenada en el origen** por:

$$\boxed{\widehat{b} = \overline{y} - \frac{s_{xy}}{s_x^2} \times \overline{x}}$$

De donde, la **recta de regresión mínimo cuadrática de** Y **sobre** X es:

$$\boxed{Y = (\overline{y} - \frac{s_{xy}}{s_x^2} \times \overline{x}) + \frac{s_{xy}}{s_x^2} \times X}$$

Bondad del ajuste. Coeficiente de correlación.

La suma de cuadrados de los residos con los valores ajustados de la pendiente y la ordenada se puede expresar como:

$$SC(\widehat{a},\widehat{b}) = (n-1)S_y^2\left(1 - \frac{s_{xy}^2}{s_x^2 \times s_y^2}\right)$$

Si $SC(\widehat{a},\widehat{b}) = 0 \Longrightarrow e_i^2 = 0,\ \forall i = 1,2,...,n \Longrightarrow$ El ajuste es perfecto pues la recta pasa por todos los puntos.
Cuanto más pequeño sea $SC(\widehat{a},\widehat{b})$, menores serán los residuos$\Longrightarrow$ Mejor será el ajuste de la recta a la nube de puntos.
Inconveniente de la $SC(\widehat{a},\widehat{b})$ como medida de la bondad del ajuste, es que no podemos determinar a partir de qué valor de ésta hemos conseguido un buen ajuste.

Introducimos el **coeficiente de correlación entre** X **e** Y definido por:

$$\boxed{r = \frac{s_{xy}}{s_x \times s_y}}$$

Y la cantidad

$$R^2 = \frac{s_{xy}^2}{s_x^2 \times s_y^2}$$

se llama **coeficiente de determinación** y se interpreta como la fracción de variación de Y explicada por la variable X.

A partir de esta cantidad, la suma de cuadrados se puede escribir como:

$$\boxed{SC(\widehat{a},\widehat{b}) = (n-1)s_y^2\,(1-R^2)}$$

Propiedades de r y R^2:

- **Coeficiente de determinación:** $0 \leq R^2 \leq 1$. Además:
  - Si $R^2 = 1 \Longrightarrow SC(\widehat{a},\widehat{b}) = 0 \Longrightarrow$ El ajuste es perfecto, la recta de regresión pasa por todos los puntos.
  - Si $R^2 = 0 \Longrightarrow \nexists$ dependencia lineal entre X e Y.
  - Cuanto más se aproxime R^2 a 1 mejor es el ajuste lineal ya que significa que la varianza residual es menor. Podemos tomar como referencia que si

$$r_{xy}^2 > 0.8 \Rightarrow \qquad \text{el ajuste es bueno}$$

$$r_{xy}^2 > 0.9 \Rightarrow \quad \text{el ajuste es muy bueno}$$

- $R^2 = r^2$, para la **regresión lineal simple**.

- **Coeficiente de correlación:**
  - Medida adimensional tal que $-1 \leq r \leq 1$.
  - Mide el grado de dependencia lineal entre X e Y.
  - Interpretación del signo:

$$\text{Si } r_{xy} > 0 \implies \text{Dependencia positiva entre } X \text{ e } Y.$$

$$\text{Si } r_{xy} < 0 \implies \text{Dependencia negativa entre } X \text{ e } Y.$$

Caso particular: Recta forzada a pasar por el origen

En ocasiones, consideracions físicas sobre el fenómeno que estamos estudiando nos llevan a pensar que si $X = 0 \Rightarrow Y = 0$. Por ejemplo: Concentración de un producto (Y) en función de tiempo (X) que se va creando en una reacción química. Cuando empezamos la reacción ($X = 0$), todavía no puede haber producto, por lo tanto, $Y = 0$.

Entonces, la recta que queremos ajustar es

$$Y = aX$$

En este caso los **residuos** son

$$e_i = y_i - ax_i \text{ para cada } i = 1, 2, ..., n$$

Criterio de ajuste:

$$\boxed{\text{Elegimos } \mathbf{a} \text{ tal que minimice } \underbrace{\sum_{i=1}^{n} [y_i - ax_i]^2}_{\mathbf{SC(a)}}}$$

donde $SC(a)$ es la **suma de cuadrados de los residuos**.

Los posibles mínimos son las soluciones de la ecuación

$$\left. SC'(a) = 0\} \Rightarrow 2\sum_{i=1}^{n} (y_i - ax_i)(-x_i) = 0 \right\}$$

cuya solución es:

$$\boxed{\widehat{a} = \frac{\overline{xy}}{\overline{x^2}}}$$

Además, es fácil comprobar que se trata de un mínimo.

2 Modelos linealizables

Modelo Exponencial: $\boxed{Y = \beta e^{\alpha X}} \quad \Longrightarrow$

Transformación: Tomamos logaritmos

$$\ln(Y) = \ln(\beta) + \alpha X$$

Cambios de variable:

$$\boxed{\ln(Y) = Y'}$$

Se obtiene la recta de regresión de Y' sobre X:

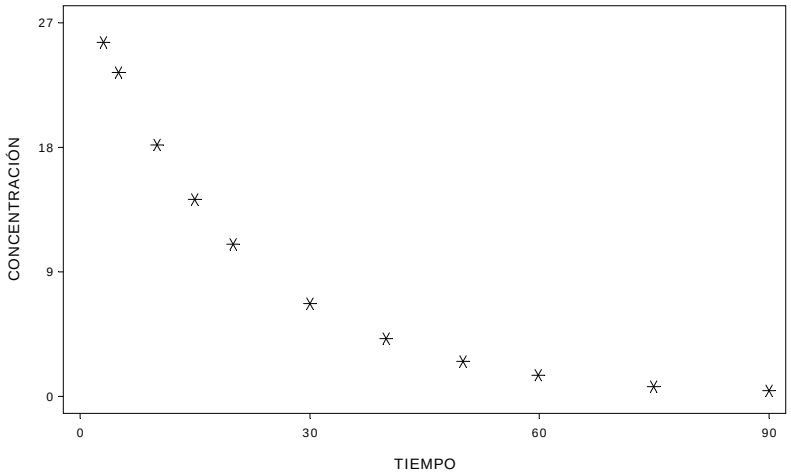
$$\boxed{Y' = b + aX}$$

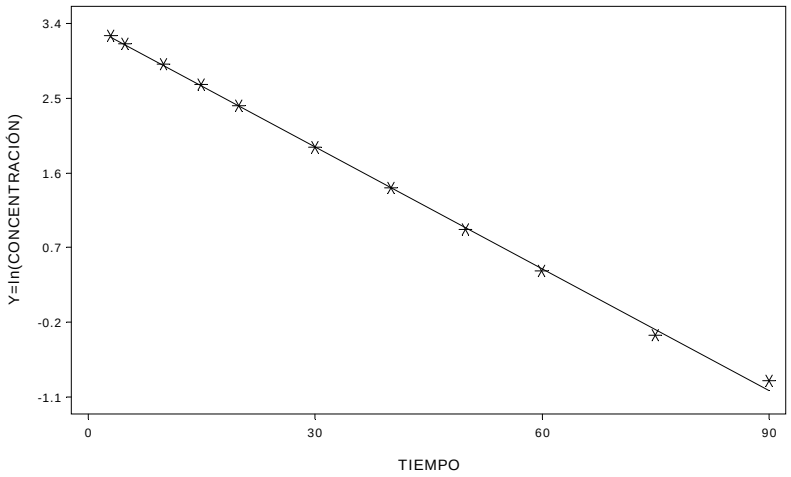
Se deshace el cambio considerado:

$$\text{Como } b = \ln(\beta) \Longrightarrow \boxed{\widehat{\beta} = e^{\widehat{b}}} \qquad \text{y} \qquad \boxed{\widehat{\alpha} = \widehat{a}}$$

Ejemplo: `Concentración de éster en función del tiempo`

$$CONCENTRACION = \beta e^{\alpha \cdot TIEMPO}$$





<u>Modelo Potencial:</u> $\boxed{Y = \beta X^{\alpha}} \quad \Longrightarrow$

Transformación: Tomamos logaritmos

$$\ln(Y) = \ln(\beta) + \alpha \ln(X)$$

Cambios de variable:

$$\boxed{\begin{array}{c} \ln(Y) = Y' \\ \\ \ln(X) = X' \end{array}}$$

Se obtiene la recta de regresión de Y' sobre X':

$$\boxed{Y' = b + aX'}$$

Se deshace el cambio considerado:

$$\text{Como } b = \ln(\beta) \Longrightarrow \boxed{\widehat{\beta} = e^{\widehat{b}}} \qquad \text{y} \qquad \boxed{\widehat{\alpha} = \widehat{a}}$$

Ejemplo: `Volumen ocupado por un gas a distintas presiones`

$$VOLUMEN = \beta \cdot PRESION^{\alpha}$$