Datos

- Descripción de una sola variable.
  - Datos y distribuciones de frecuencias.
  - Medidas de centralización y dispersión.
  - Medidas no centrales.
  - Visualización datos con medidas no centrales
  - Medidas de asimetría y curtosis.
  - Datos atípicos y diagramas caja (Boxplots).
- Descripción conjunta de varias variables.
  - Distribuciones de frecuencia multivariantes
  - Medidas de dependencia lineal
  - Recta de Regresión
  - Matriz de varianzas

Datos: Descripción de una sola variable

Descripción de una sola variable:

- Datos y distribuciones de frecuencias
- Representaciones gráficas
- Medidas de centralización y dispersión
- Medidas no centrales
- Visualización datos con medidas no centrales
- Medidas de asimetría y curtosis
- Datos atípicos y diagramas caja (boxplots)

Datos y distribuciones de frecuencias

- Recordemos que la estadística descriptiva estudia los procedimientos para sintetizar información para un conjunto de datos de una variable **x**:
  - Variables cualitativas, categóricas o atributos.
  - Variables cuantitativas discretas.
  - Variables cuantitativas continuas.
- **Frecuencia absoluta** de un suceso $\mathbf{x_i}$:
  - Es el número de veces que se observa $\mathbf{x_i}$
- **Frecuencia relativa** de un suceso $\mathbf{x_i}$:
  - $f_r(\mathbf{x_i})$=# veces que se observa $\mathbf{x_i}$ /# total de datos
  - Cumple que $\sum f_r(\mathbf{x_i})=1$

Parte de la foto sacada de https://es.wikipedia.org/wiki/Frecuencia_estad%C3%ADstica

A	B		
x_i	n_i	f_i	p_i
1	16	8/25	32%
2	20	2/5	40%
3	9	9/50	18%
4	5	1/10	10%
	N=50	$\sum f_i$=1	$\sum p_i$=100%

Representaciones gráficas

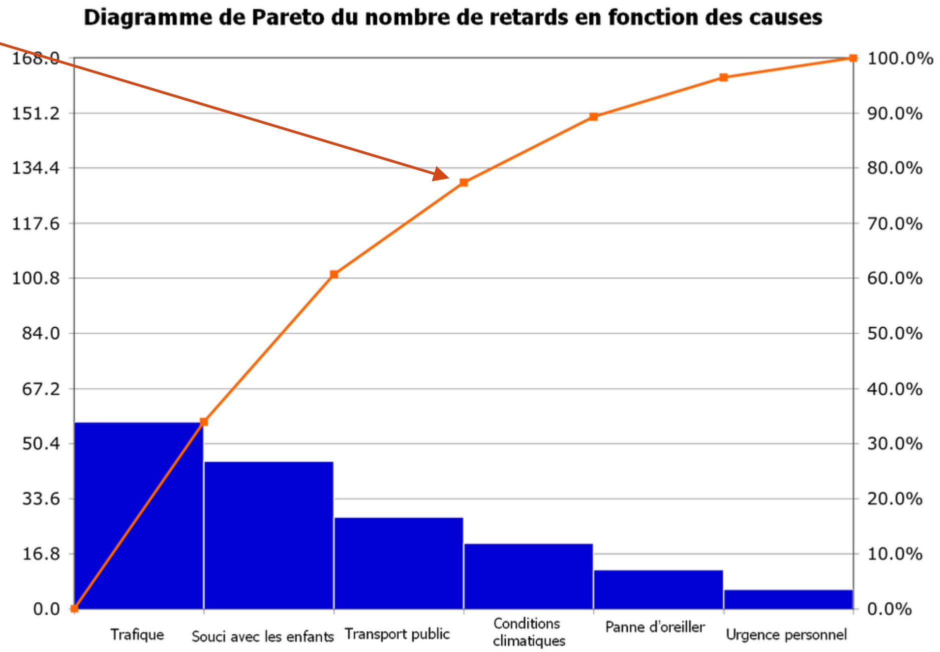
- Hay muchas:
  - Diagrama de Pareto.
  - Diagrama de barras.
  - Histogramas.
  - Gráficos temporales.
  - Otras representaciones.

Representaciones gráficas: Diagrama Pareto

- Para datos cualitativos y representa el principio de Pareto: pocos vitales, muchos triviales.

Para reducir la cantidad de retrasos en un 78%, es suficiente resolver los tres primeros problemas.

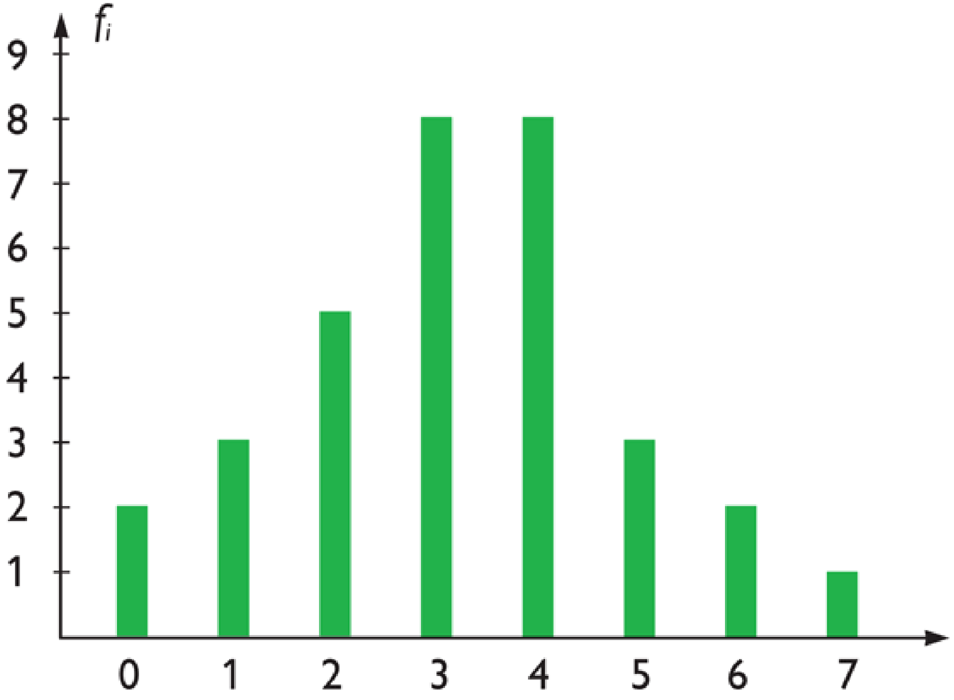
Foto sacada de https://es.wikipedia.org/wiki/Diagrama_de_Pareto#/media/File:Pareto.png



- Organiza los datos de manera que el orden sea descendiente.
- De esta forma se puede asignar orden a las prioridades en el que se toman las decisiones.
- Muestra el principio de Pareto: muchos problemas sin importancia frente a unos pocos muy importantes.

- Generalmente para variables discretas.

Foto sacada de http://maralboran.org/wikipedia/index.php/Gr%C3%A1ficos_estad%C3%ADsticos



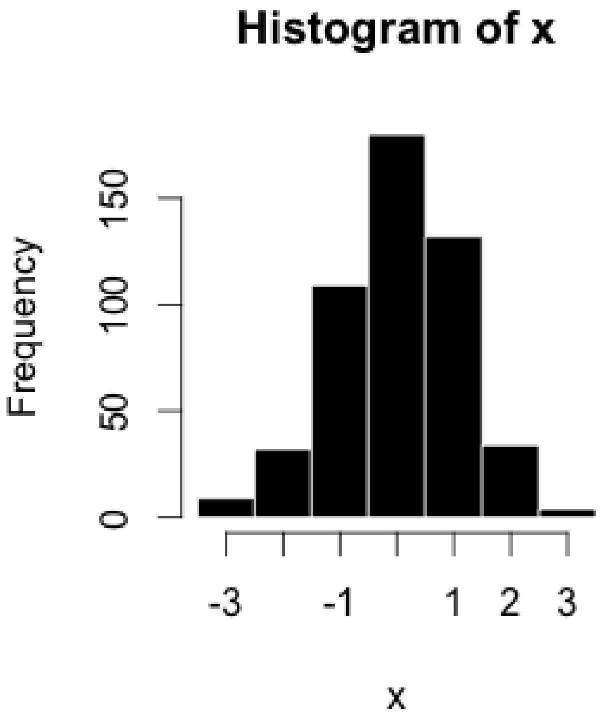
Representaciones gráficas: Histograma

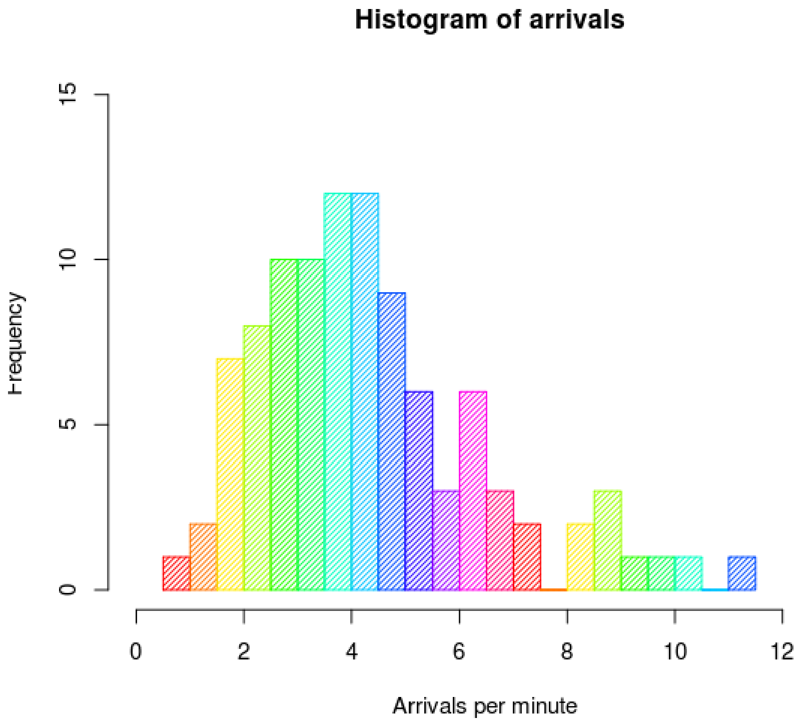
- Conjunto de rectángulos cada uno de los cuales representa un intervalo de agrupación o clase.
- Sus bases son iguales a la amplitud del intervalo, y las alturas se determinan que su área sea proporcional a la frecuencia de cada clase.
- Vale para valores discretos o continuos.
- $k=ceil[(max(x) - min(x))/h]$,
  - donde k es el número de rectángulos,
  - y h es la anchura del rectángulo.
- ¿Que k se selecciona? hay muchas reglas en función del numero de muestras, n:
  - $k=Sqr(n)$
  - Formula de Sturges: $k=ceil(\log_2 n+1)$
  - Regla de Rice: $k=ceil(2n^{1/3})$
  - Etc.

Representaciones gráficas: Histograma

Fotos sacadas de https://en.wikipedia.org/wiki/Histogram

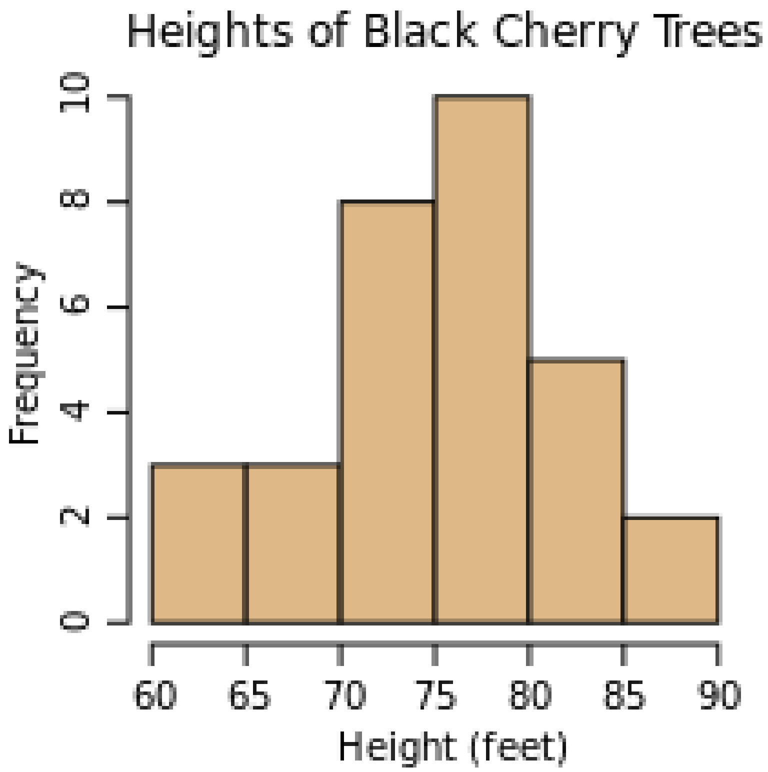
Bin	Count
−3.5	9
−2.5	32
−1.5	109
−0.5	180
0.5	132
1.5	34
2.5	4
3.5	9





Representaciones gráficas: Histograma

- Explicar el significado de un histograma.



Fotos sacadas de https://commons.wikimedia.org/wiki/File:Black_cherry_tree_histogram.svg

	Girth	Height	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
11	11.3	79	24.2
12	11.4	76	21.0
13	11.4	76	21.4
14	11.7	69	21.3
15	12.0	75	19.1
16	12.9	74	22.2
17	12.9	85	33.8
18	13.3	86	27.4
19	13.7	71	25.7
20	13.8	64	24.9
21	14.0	78	34.5
22	14.2	80	31.7
23	14.5	74	36.3
24	16.0	72	38.3
25	16.3	77	42.6
26	17.3	81	55.4
27	17.5	82	55.7
28	17.9	80	58.3
29	18.0	80	51.5
30	18.0	80	51.0
31	20.6	87	77.0

Medidas de centralización y dispersión

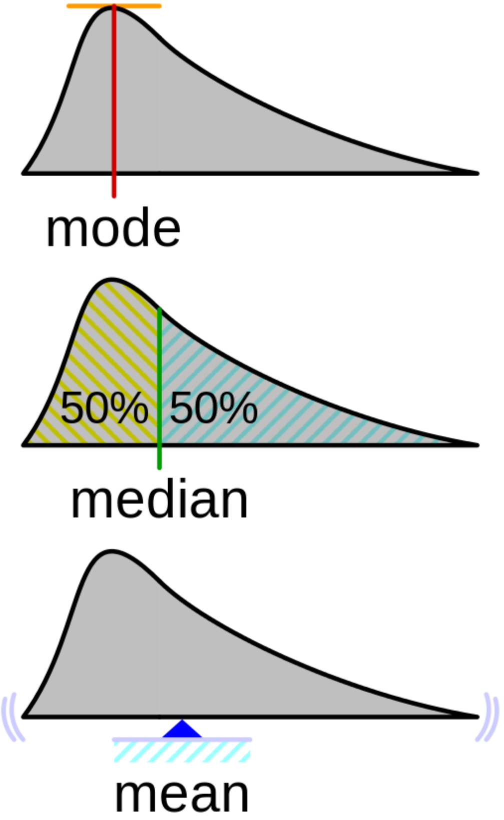
- Medidas de centralización:
  - **Media:** $\langle x\rangle=(x_1+x_2+\ldots+x_n)/n=\sum x_i/n$, o en frecuencias $\langle x\rangle=\sum x_i\, fr(x_i)$. **Es muy sensible a observaciones atípicas.**
  - **Mediana:** valor tal que ordenados en magnitud los datos, el 50% es menor que ella y el 50% mayor. Es el valor central si el número de datos es impar, o la media de los dos centrales si es par. Se suele denominar por **Med. No es tan sensible a observaciones atípicas.** Si la media y mediana difieren mucho es posible que exista heterogeneidad de poblaciones.
  - **Moda:** valor que se repite más.
- Medidas de dispersión:
  - **Desviación típica:** $s=[\sum(x_i-\langle x\rangle)^2/n]^{1/2}$, o $s=[\sum(x_i^2/ n) - \langle x\rangle^2]^{1/2}$

  o en frecuencias $s=[\sum(x_i- \langle x\rangle)^2 f_r(x_i)]^{1/2}$
  - **Coeficiente de variación:** $CV=s/|\langle x\rangle| > 0$ (mide la variabilidad, el inverso es el coeficiente de señal-ruido).
    - Si CV es grande (>1.5 normalmente), indica posible error en los datos, o datos heterogéneos.

Medidas de centralización y dispersión

- Medidas de dispersión:
  - Otras medidas de dispersión:
    - Asociadas a mediana:
      - **MEDA**=mediana|x_i-Med|,
      - Como la mediana, MEDA no es tan alterada por datos extremos.
      - Este tipo de medidas que no se ven aceptadas por los datos extremos, se denotan como medidas **robustas** o resistentes.
      - Med y MEDA cumplen que al menos el 50% de los datos esta en el intervalo (Med-MEDA, Med+MEDA).

Foto sacada de https://es.wikipedia.org/wiki/Mediana_(estad%C3%ADstica)



Medidas de centralización y dispersión

- La interpretación que se suele dar a la media y la desviación (información conjunta de media y dispersión):
  - En el intervalo ($<x>-ks$, $<x>+ks$) existe como mínimo el $100(1-(1/k^2))\%$ de las observaciones:
    - Por ejemplo si la media es 500 y la desviación típica 20:
      - Para dos desviaciones típicas tenemos que en el intervalo (460, 540) estarán como mínimo el $100(1-(1/2^2))\%=75\%$ de las observaciones.
      - Para tres desviaciones típicas tenemos que en el intervalo (440, 560) estarán como mínimo el $100(1-(1/3^2))\%=89\%$ de las observaciones.

Medidas de centralización y dispersión

- Esto es equivalente a que (demostración en el libro):
  - $f_r(|x_i - \langle x \rangle| \geq ks) \geq 1-(1/k^2)$ que nos permite concluir que en cualquier distribución de datos se encuentra, al menos:
    - Entre la media y dos desviaciones típicas el 75% de las observaciones.
    - Entre la media y tres desviaciones típicas el 89% de las observaciones.
    - Y así sucesivamente.
- Esta desigualdad es la famosa ***desigualdad de Tchebychev***.

Medidas de posicionamiento

- Percentil p: es el menor valor superior al p% de los datos ordenados. Si el numero de datos es impar Med es el percentil 50. Es una medida no central usada en estadística que indica el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones.
- Cuartiles: aquellos valores que dividen la distribución en cuatro partes iguales.
  - Primer cuartil Q_1: es el percentil 25.
  - Segundo cuartil Q_2: es la mediana.
  - Tercer cuartil Q_3: es el percentil 75.

Medidas de posicionamiento

- Los percentiles y cuartiles se utilizan para construir medidas de dispersión basadas en datos ordenados:
  - Rango intercuartílico: es la diferencia ente los pecentiles 75 y 25.
- Hay muchos tipos ***cuantiles***:
  - Percentiles, Quartiles, Deciles, Etc.
- El concepto de mediana se generaliza mediante los cuantiles.
- Un cuantil de ***orden k*** (donde k es un valor entre 0 y 1), será el valor de la variable que deja por debajo de si una ***proporción*** de k observaciones del total n de todas observaciones, que se han ordenado previamente en magnitud.
- Es decir habrá $n_x k$ observaciones con valores menores o igual al cuantil k.
- Así la media resulta ser el cuantil de orden ½.

Medidas de posicionamiento

- En general los cuantiles son puntos tomados a intervalos regulares de la función de distribución de una variable aleatoria.
- Así cuantil de orden p de una distribución es el valor de la variable x_p que indica que por debajo de esta se encuentra un porcentaje p dado de observaciones. En este caso estamos en distribuciones, y $0 < p < 1$.
  - Ej: El cuantil de orden 0,27 es el valor de la x que deja un 27% de probailidad a su izquierda, y por tanto el cuantil de orden 0,50 se corresponde con la mediana de la distribución.

Medidas de posicionamiento

- Resumiendo los tipos de cuantiles más habituales son:
  - Percentiles: dividen a la distribución en cien partes.
  - Cuartiles: dividen a la distribución en cuatro partes (corresponden a los cuantiles 0,25; 0,50 y 0,75).
  - Quintiles: dividen a la distribución en cinco partes (corresponden a los cuantiles 0,20; 0,40; 0,60 y 0,80).
  - Deciles: dividen a la distribución en diez partes.

Medidas de posicionamiento

- Existen muchos métodos para calcular los percentiles:

Weisstein, Eric W. "Quartile." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/Quartile.html

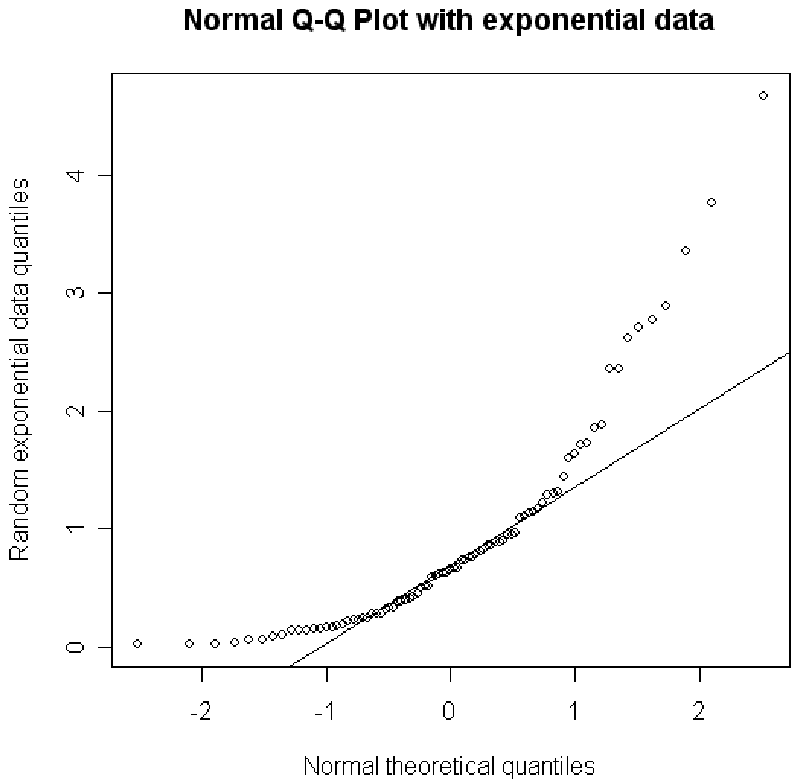
method	1st quartile	1st quartile	3rd quartile	3rd quartile
	n odd	n even	n odd	n even
Minitab	$\frac{n+1}{4}$	$\frac{n+1}{4}$	$\frac{3n+3}{4}$	$\frac{3n+3}{4}$
Tukey (Hoaglin et al. 1983)	$\frac{n+3}{4}$	$\frac{n+2}{4}$	$\frac{3n+1}{4}$	$\frac{3n+2}{4}$
Moore and McCabe (2002)	$\frac{n+1}{4}$	$\frac{n+2}{4}$	$\frac{3n+3}{4}$	$\frac{3n+2}{4}$
Mendenhall and Sincich (1995)	$\left[\frac{n+1}{4}\right]$	$\left[\frac{n+1}{4}\right]$	$\left[\frac{3n+3}{4}\right]$	$\left[\frac{3n+3}{4}\right]$
Freund and Perles (1987)	$\frac{n+3}{4}$	$\frac{n+3}{4}$	$\frac{3n+1}{4}$	$\frac{3n+1}{4}$

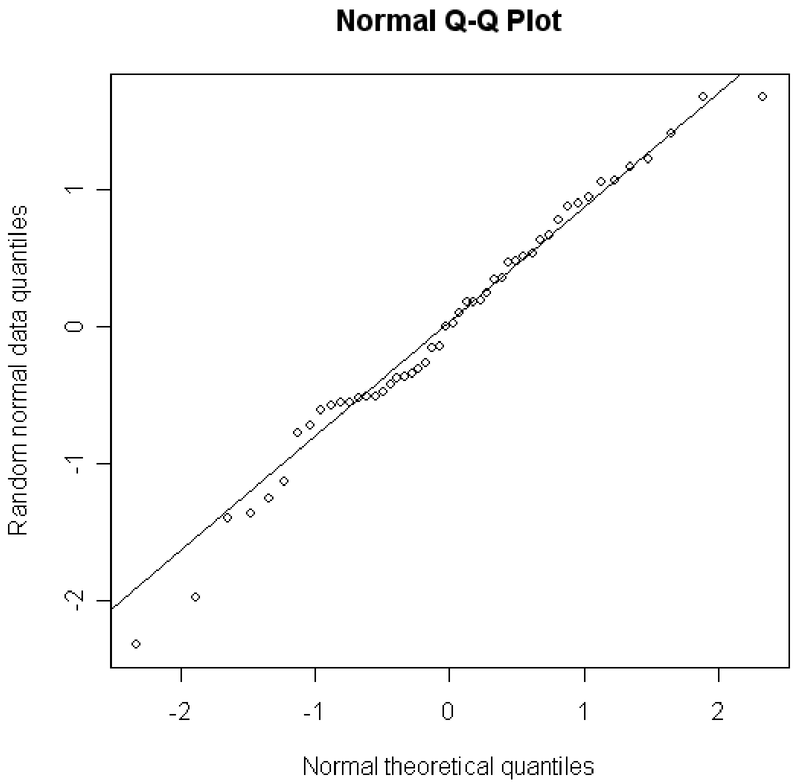
Visualización datos con medidas de posicionamiento: Gráficos Q-Q

- Los gráficos Q-Q (Q-Q plots) se caracterizan por visualizar de una manera muy rápida y sencilla como se diferencian los datos de dos distribuciones de observaciones.
- Se basan en representar enfrentados en un gráfico x-y los cuantiles de ambas distribuciones. El "Q" viene de cuantil en inglés.
- Si todos los cuantiles son iguales aparecerá la recta x=y en el gráfico, y significará los dos conjuntos de datos se distribuyen de manera idéntica.
- Generalmente una de la distribuciones es conocida (por ejemplo una normal), para contrastar si los datos observados se ajustan a la distribución conocida.

Visualización datos con medidas de posicionamiento: Gráficos Q-Q

Imágenes extraídas de https://es.wikipedia.org/wiki/Gr%C3%A1fico_Q-Q

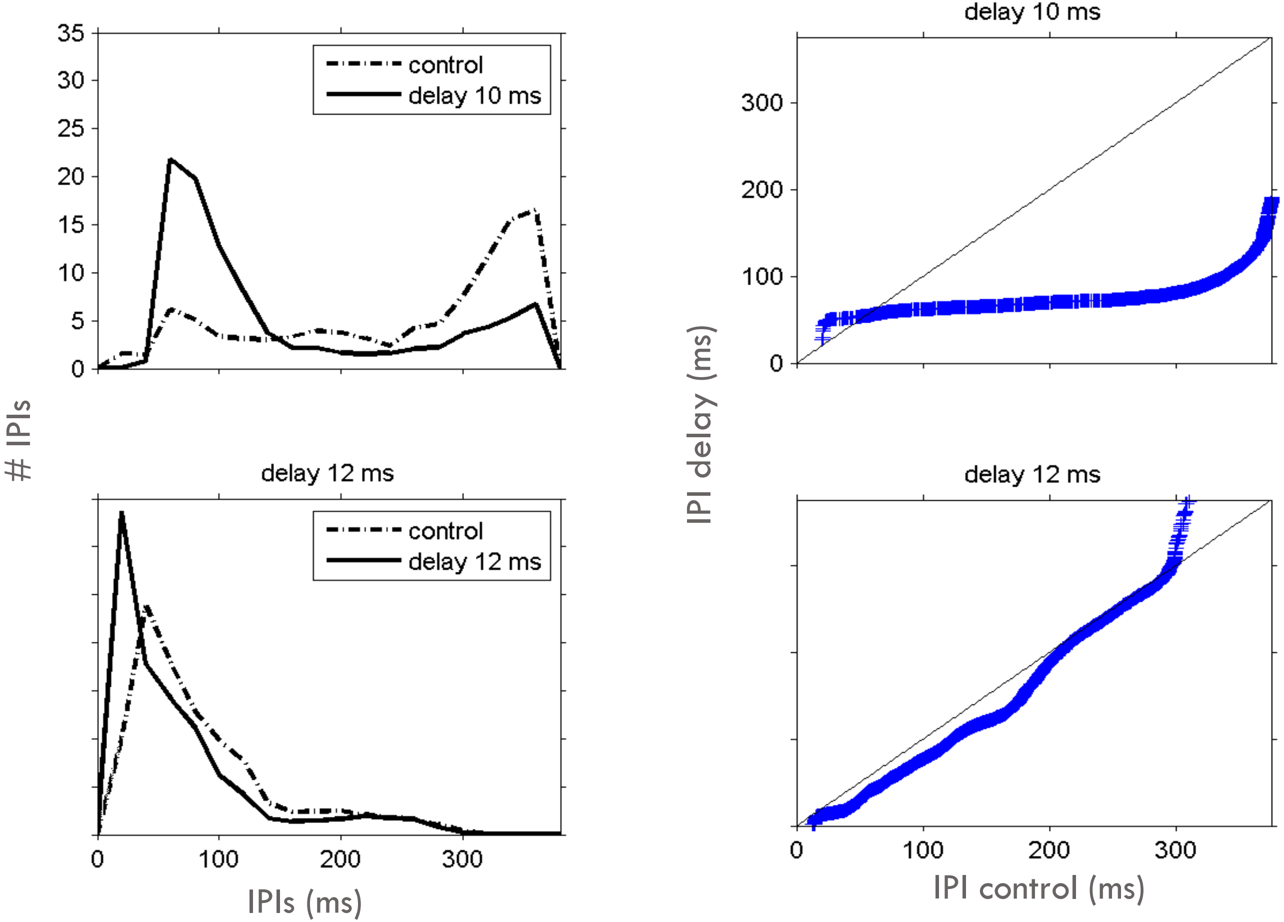




Visualización datos con medidas de posicionamiento: Gráficos Q-Q

- Para el calculo del cuantil ***k*** y para ***n*** datos se utilizan diversas formulas:
  - La más habitual es $k/(n+1)$
  - Para graficas simétricas $(k-a)/(n+1-2a)$, con $a \in (0,0.5)$, con $a=0$ retomamos la primera, con $a=0.5$ se suele utilizar para distribuciones normales.
  - $(k-(1/3))/(n+(1/3))$.
  - $(k-0.3175)/(n+0.365)$.
  - $(k-0.326)/(n+0.348)$.
  - Etc.

Visualización datos con medidas de posicionamiento: Gráficos Q-Q



Medidas de asimetría y curtosis

- Informan de aspectos importantes de la forma de la distribución
- Coeficiente de asimetría (medida adimensional):
  - $CA=\sum(x_i-<x>)^3/ns^3$
  - En un conjunto de datos simétricos respecto a la media el sumatorio es nulo.
  - Si CA es negativo la distribución se alarga para valores inferiores a la media (la cola de la distribución se extiende a la izquierda). La masa de la distribución se concentra en el lado derecho.
  - Si CA es positivo la distribución se alarga para valores superiores a la media (la cola de la distribución se extiende a la derecha). La masa de la distribución se concentra en el lado izquierdo.

Medidas de asimetría y curtosis

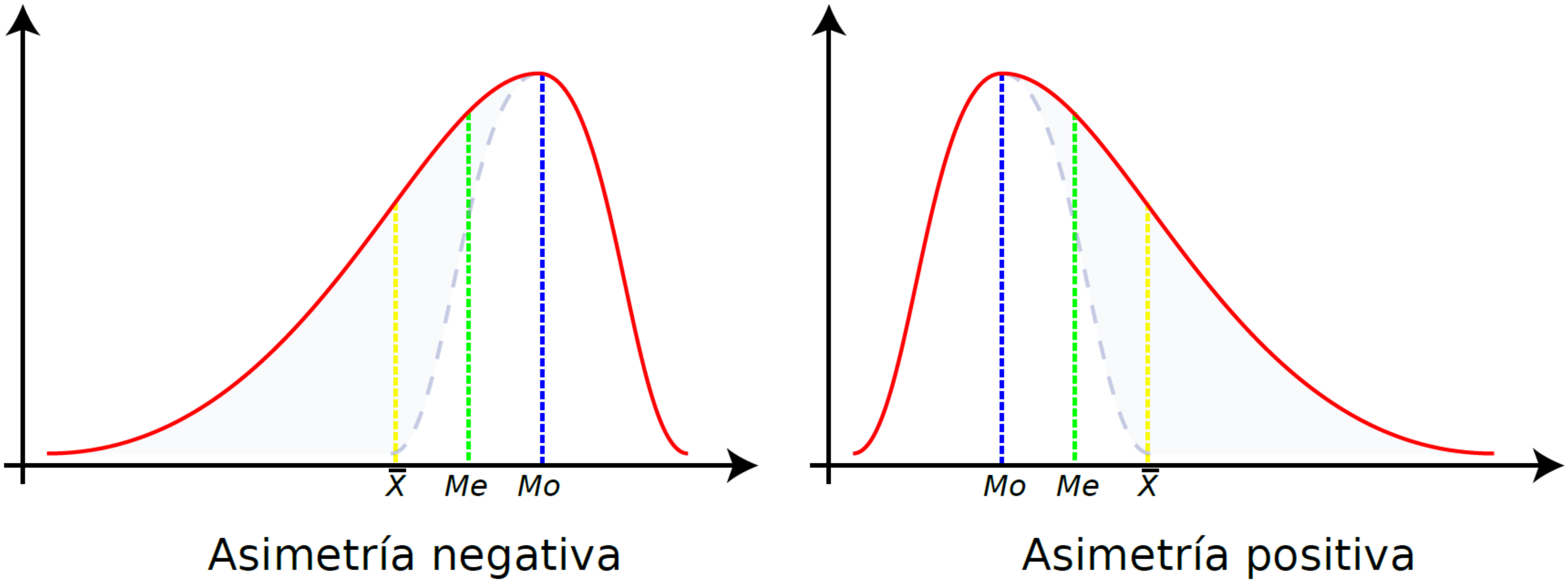


Imagen extraída de https://es.wikipedia.org/wiki/Asimetr%C3%ADa_estad%C3%ADstica

Medidas de asimetría y curtosis

- **Coeficiente de curtosis (apuntamiento y aplastamiento):**
  - $CA_p=\sum(x_i-<x>)^4/ns^4$, se suele definir como $CA_p=\sum(x_i-<x>)^4/ns^4-3$ ya que la curtosis de la normal es 3, y así la curtosis se expresa en función de la normal.
  - Mide como la frecuencia relativa de unos datos observados se reparte entre el centro y los extremos. A la curtosis también se le llama *apuntamiento*.
  - De esta forma se puede medir la heterogeneidad y presencia de valores atípicos de la distribución (ver ejemplos del libro):
    - Cuando el apuntamiento es muy bajo se dice que la distribución es heterogénea.
    - Sin embargo, cuanto el apuntamiento es muy alto indica la presencia de valores atípicos en la distribución.
    - Valores intermedios de este coeficiente indica la presencia de homogeneidad en la muestra.
  - También se puede ver este coeficiente en comparación con la distribución normal:
    - Si CA_p <0 la distribución es más aplanada que la normal (platicúrtica).
    - Si CA_p =0 la distribución es igual de apuntada que la normal (mesocúrtica).
    - Si CA_p >0 la distribución es más apuntada que la normal (leptocúrtica).

Medidas de asimetría y curtosis

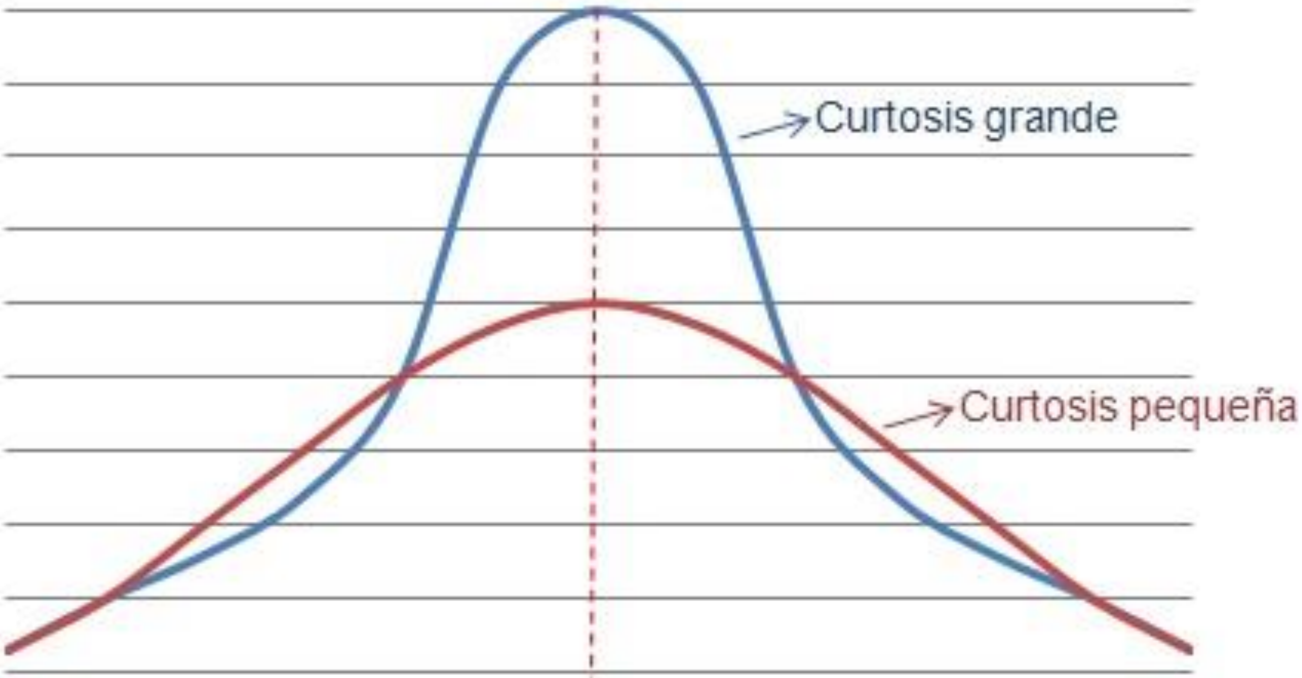


Imagen extraída de http://www.universoformulas.com/estadistica/descriptiva/curtosis/

Datos atípicos y diagramas caja (boxplots)

- Datos atípicos: es muy frecuente que los datos presenten cierta observaciones que se ha han medido erróneamente o que se han transcrito mal.
  - En un histograma pueden ser los datos en los extremos aislados, pero no solo pueden estar en los extremos.
  - Para detectarlos se utilizan valores de centralización y dispersión que estén poco afectados por estos (Mediana y MEDA).
  - Se suele utilizar esta regla para considerar valores sospechosos:
    - $x > \text{Med}(x) \pm (4.5 \times \text{MEDA}(x))$
  - Este criterio se utiliza mucho pero no tiene en cuenta la asimetría de la distribución.
  - Sin embargo las medidas cuantiles tiene en cuenta la asimetría y por tanto se pueden utilizar para esto.
  - Así considerando los cuartiles y el rango intercuartílico, se suele utilizar esta regla para considerar valores atípicos:
    - $x < Q_1 - 1.5(Q_3-Q_1)$
    - $x > Q_3 + 1.5(Q_3-Q_1)$

Datos atípicos y diagramas caja (boxplots)

- Diagramas caja (boxplots): representación semigráfica de una distribución que muestra sus características principales, señalando los posible datos atípicos (ejercicio 2.5 del libro). El cálculo:
  - Se ordenan los datos y se sacan los cuartiles.
  - Se pintan los cuartiles en los extremos en un rectángulo con la mediana como separación en el interior.
  - Se calculan los límites admisibles de valores atípicos:
    - $LI < Q_1 - 1.5(Q_3-Q_1)$
    - $LS > Q_3 + 1.5(Q_3-Q_1)$
  - Se pintan dos líneas (bigotes) que salen del rectángulo indicando estos límites.
  - Algunas veces los "bigotes" son las líneas que van desde cada extremo del rectángulo central hasta el valor más alejado no atípico, es decir, que está dentro del intervalo (LI, LS).

Datos atípicos y diagramas caja (boxplots)

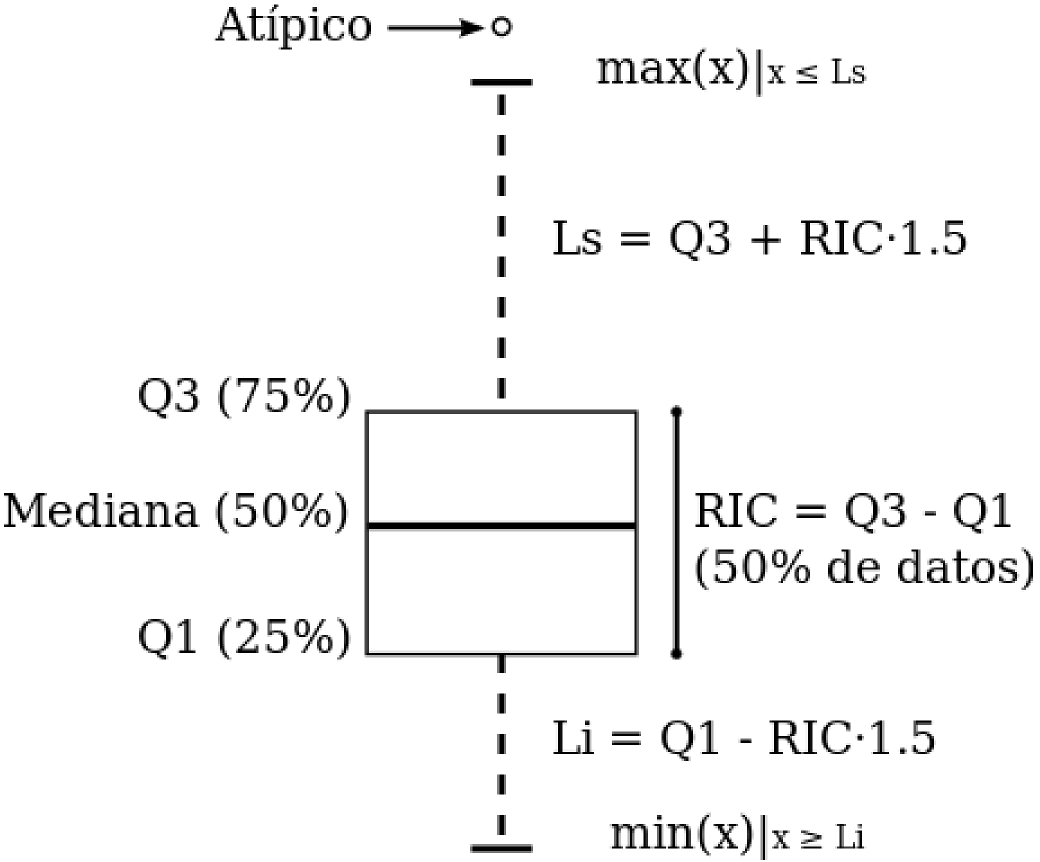
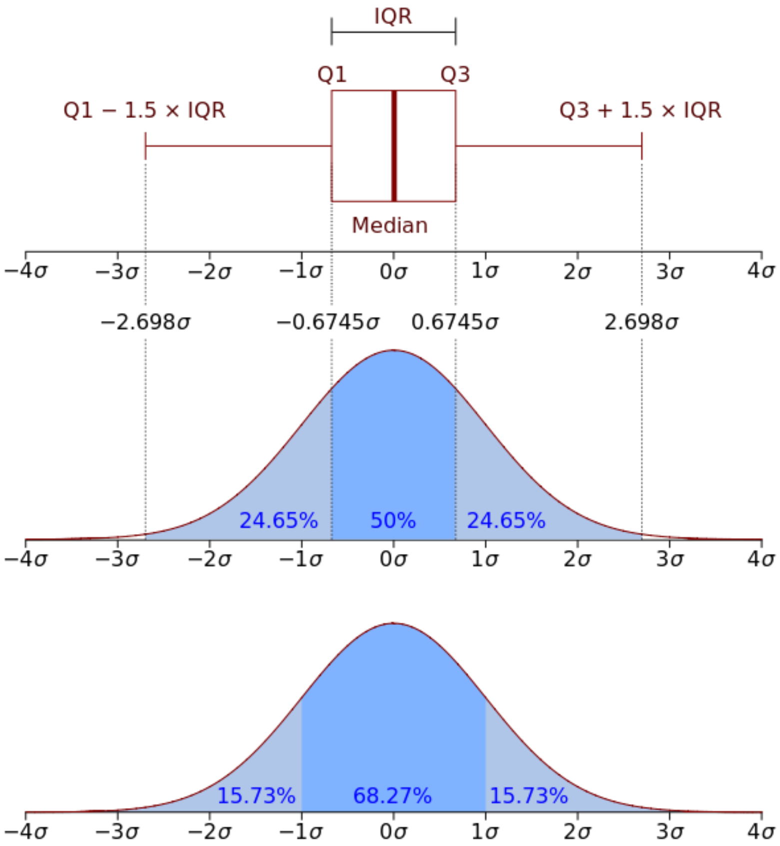


Imagen extraída de https://es.wikipedia.org/wiki/Diagrama_de_caja

Datos atípicos y diagramas caja (boxplots)

Imagen extraída de https://en.wikipedia.org/wiki/Box_plot



Datos atípicos y diagramas caja (boxplots)

i	1	2	3	4	5
	6	7	8	9	10
	11	12	13	14	15
	16	17	18	19	20
x_i	9	6	7	7	3
	9	10	1	8	7
	9	9	8	10	5
	10	10	9	10	8
x_i (sort)	1	3	5	6	7
	7	7	8	8	8
	9	V9	9	9	9
	10	10	10	10	10

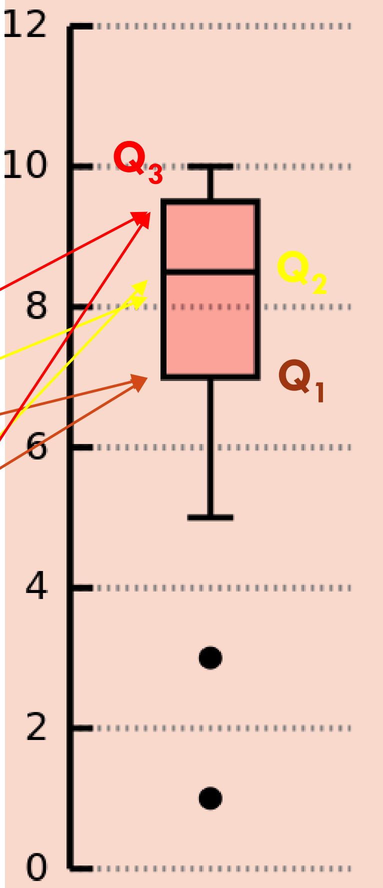


Imagen extraída de https://de.wikipedia.org/wiki/Boxplot

Datos atípicos y diagramas caja (boxplots)

- Q3=(9+10)/2=9,5
- Q2=(8+9)/2=8,5
- Q1=(7+7)/2=7
- RIC=Q3-Q1=2,5
  - LI $< Q_1 - 1{,}5(Q_3\text{-}Q_1) = 7\text{-}1{,}5*2{,}5=7\text{-}3{,}75=3{,}25$.
  - LS $> Q_3 + 1.5(Q_3\text{-}Q_1) = 9.5+1{,}5*2{,}5=9{,}5+3{,}75=13{,}25$.
- El bigote inferior llega bajando hasta el dato antes de 3,25, es decir el dato 5.
- El bigote superior llega hasta el dato antes de 13,25, es decir 10.
- Los que están por debajo de 3,25 son atípicos (los valores 1 y 3).
- Los que están por encima de 13,25 son atípicos (ningún valor).

Datos atípicos Multivariantes

- La metodología univariante para detectar datos atípicos no tiene porque funcionar en problemas multivariantes.
- En problemas multivariantes se puede utilizar la distancia de Mahalanobis.
- Esta medida se puede ver como la distancia al centro de gravedad de las variables aleatorias (media multivariante), ponderada por la matriz de covarianaza.
- Para datos multivariantes distribuidos normalmente, los valores de la distancia de Mahalanobis se puede demostrar que tienen aproximadamente una distribución chi-cuadrado con p grados de libertad.
- Así, cuando un dato multivariante tenga una distancia de Mahalanobis muy grande será atípico.
- Para una lectura rápida de estas técnicas y otras: aquí 1, aquí 2 y aquí 3.

Ejemplos para realizar en casa

- Realizar para casa los ejemplos del libro:
  - 2-1, 2-2, 2-3, 2-4 y 2-5.

Ejemplos para realizar en casa (ejemplo 2.1, tabla 2.2)

X	frecuencia	frecuencia relativa
0	40	0,44
1	26	0,29
2	14	0,16
3	6	0,07
4	3	0,03
5	0	0,00
6	1	0,01
TOTAL	90	1

Distribución de frecuencias de la variable: número de llamadas recibidas en una centralita en períodos de un minuto

Otro criterio para la mediana (caso discreto):

mediana es el valor x_m tal que $fr(x<x_m) < 0,5$ pero $fr(x \leq x_m) \geq 0,5$ (diferente la transparencia 18).

Ejemplos para realizar en casa (ejemplo 2, tabla 2.3)

Intervalo	Centro del intervalo	Frecuencia relativa
20-24	22	0,30
25-29	27	0,40
30-34	32	0,20
35-39	37	0,07
40-44	42	0,03

Distribución de la variable: tiempo en minutos al realizar una operación

Datos: Descripción conjunta de variables

Descripción conjunta de varias variables.

- Distribuciones de frecuencia multivariantes
- Medidas de dependencia lineal
- Recta de Regresión
- Matriz de varianzas

Distribuciones de frecuencia multivariantes

- Uno de los objetivos del análisis estadístico es encontrar las relaciones que existen entre un grupo de variables.
- Ahora supondremos que nuestro conjunto de datos contiene los valores de las variables (x, y), que se han medido conjuntamente en una población.
- El análisis que veremos aquí se generaliza para cualquier número de variables.

Distribuciones de frecuencia multivariantes: D. Conjunta

- La distribución conjunta de frecuencias de dos variables (x, y) es una tabla que representa los valores observados de ambas variables, y las frecuencias relativas de aparición de cada pareja de valores.
- La suma de las frecuencias relativas
  - $\sum_i \sum_j \; f_r(\mathbf{x_i}, \mathbf{y_j})=1$,
    - siendo $f_r(\mathbf{x_i}, \mathbf{y_j})= n_i(x_i,y_j)/n$.
- Frecuencias absoluta de una pareja:
  - $f_r(\mathbf{x_i}, \mathbf{y_j}) \times n$, donde n=20.
    - $n_1(70,175)=0.05 \times 20=1$
    - $n_5(70,165)=0.2 \times 20=4$

Pesos X_i	Altura Y_i	$n_i(x_i,y_i)$	$f_r(x_i,y_i)$
70	175	1	0.05
65	160	3	0.15
85	180	2	0.1
60	155	3	0.15
70	165	4	0.2
75	180	2	0.1
90	185	1	0,05
80	175	1	0.05
60	160	2	0.1
70	170	1	0.05
		n=20	1

Distribuciones de frecuencia multivariantes: Distribución Marginal

- La distribución marginal de frecuencias de una variable se obtiene al estudiar esta aislada, e independiente del resto, así:
  - La distribución marginal de **x** se obtiene como $f_r(\mathbf{x_i}) = \sum_j f_r(\mathbf{x_i}, \mathbf{y_j})$.
  - Y análogamente la de **y** como $f_r(\mathbf{y_j}) = \sum_i f_r(\mathbf{x_i}, \mathbf{y_j})$.
- Ejemplo: Frecuencias relativas del color de ojos de 1.000 personas y de sus madres:

$\boxed{f_r(\mathbf{x_i}, \mathbf{y_j}) = f_r(\text{hijo\_color\_ojos, madre\_color\_ojos}).}$

	Madres		
Hijo	Claros	Oscuros	TOTAL
Claros	0,25	0,08	0,33
Oscuros	0,12	0,55	0,67
TOTAL	0,37	0,63	

0,25x1000=250 parejas de hijos-madre con los ojos claros= $f_r(\mathbf{x_i}=\text{cla}, \mathbf{y_j}=\text{cla}) \times n = 250$

$f_r(\mathbf{x_i}=\text{cla}) = \sum_j f_r(\mathbf{x_i}=\text{cla}, \mathbf{y_j})=$

$= f_r(\mathbf{x_i}=\text{cla}, \mathbf{y_j}=\text{osc}) + f_r(\mathbf{x_i}=\text{cla}, \mathbf{y_j}=\text{cla}) = 0.33$

Distribuciones de frecuencia multivariantes: D. Condicional

- Distribución condicional: la distribución de y para $x=x_i$ es la distribución **univariante** de la variable **y** que se obtiene considerando sólo los elementos que tienen para la variable **x** el valor x_i:
  - $f_r(y_j | x_i) = f_r(x_i, y_j) / f_r(x_i)$.
  - Se normaliza por $f_r(x_i)$: de esta forma se garantiza que la suma de frecuencias relativas para todos los valores de la variable **y** es 1:
    - $\sum_j f_r(y_j | x_i) = \sum_j f_r(x_i, y_j) / f_r(x_i) = f_r(x_i) / f_r(x_i) = 1$.

Distribución marginal de **x**: $f_r(x_i) = \sum_j f_r(x_i, y_j)$

- La distribución condicional de y para $x=x_i$ se interpreta como la distribución de la característica **y** en los elementos de la población que tienen como característica **x** el valor de x_i.

Distribuciones de frecuencia multivariantes: D. Condicional

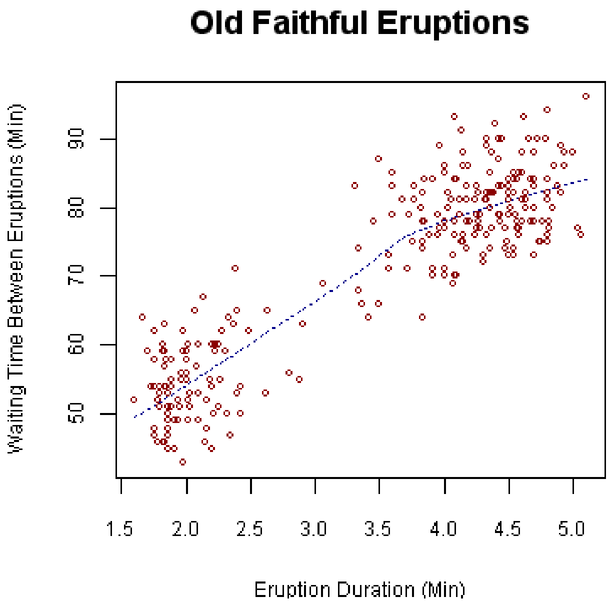
- La diferencia con la distribución marginal es clara: La distribución marginal en **y** tiene en cuenta la distribución de **y** en todos los elementos con independencia del valor que en ellos tenga la característica **x.** En cambio la condicional fija un conjunto de valores de $\mathbf{x}=\mathbf{x}_i$.
- Se puede deducir la distribución marginal de la condicional:
  - $f_r(\mathbf{y}) = \sum_i f_r(\mathbf{y} \mid \mathbf{x}_i) \times f_r(\mathbf{x}_i)$, i.e. la frecuencia de **y** en la población total se obtiene ponderando su frecuencia en las subpoblaciones definidas por los distintos valores de x.
- De igual forma la distribución conjunta se obtiene de la condicional, si conocemos todas las distribuciones marginales y condicionales para cada variable:
  - $f_r(\mathbf{x}_i, \mathbf{y}_j) = f_r(\mathbf{y}_j \mid \mathbf{x}_i) \times f_r(\mathbf{x}_i)$.

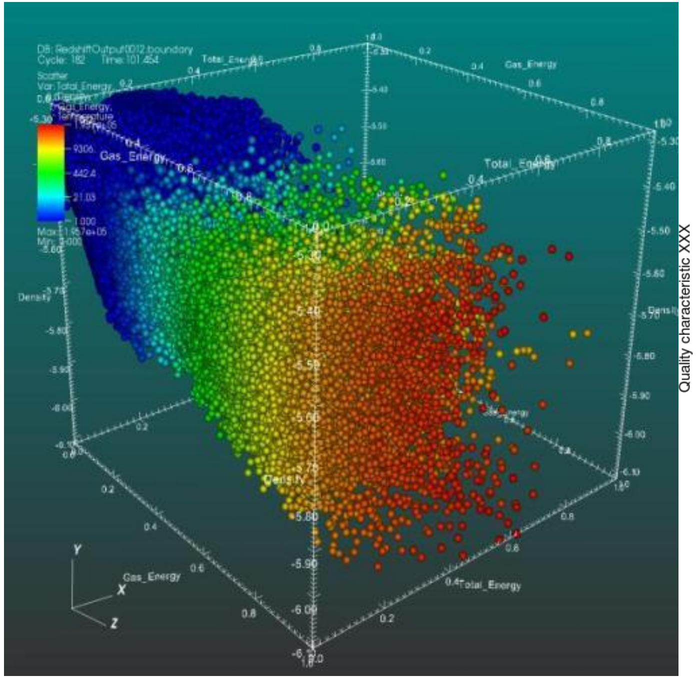
Distribuciones de frecuencia multivariantes: Representaciones Gráficas

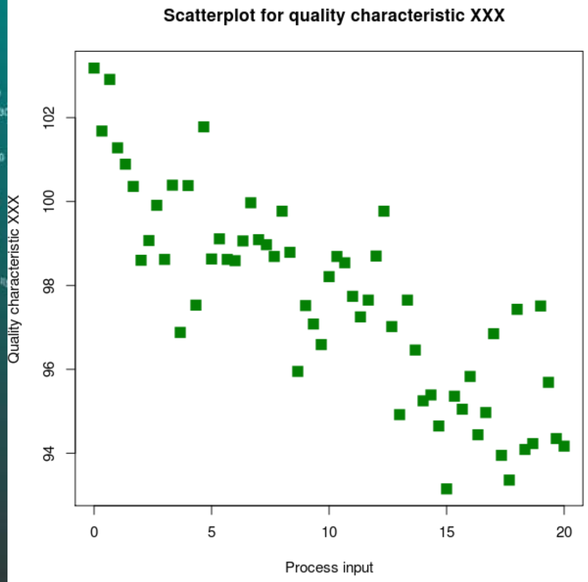
- La representación de las observaciones de dos o tres variables enfrentadas es lo que se llaman diagramas de dispersión.
- En el se puede observar muy rápidamente la posible existencia de relaciones entre las variables:
  - Relación lineal positiva.
  - Relación lineal negativa.
  - Falta de relación.
  - Otros tipos de relación.
- Cuando tenemos varias variables se pueden replantear como una matriz de figuras de diagramas de dispersión y en la diagonal los histogramas de las variables.

Distribuciones de frecuencia multivariantes: Representaciones Gráficas

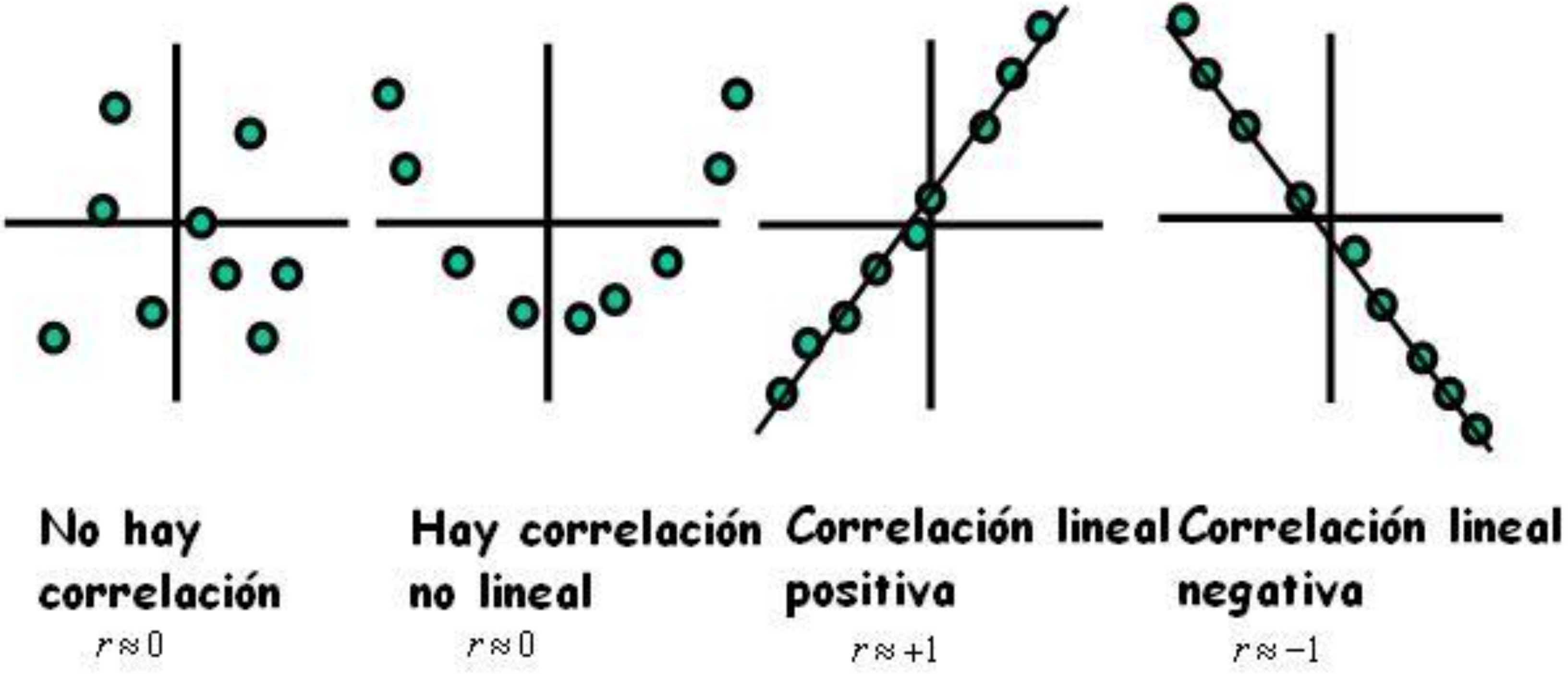
Imágenes extraídas de https://en.wikipedia.org/wiki/Scatter_plot





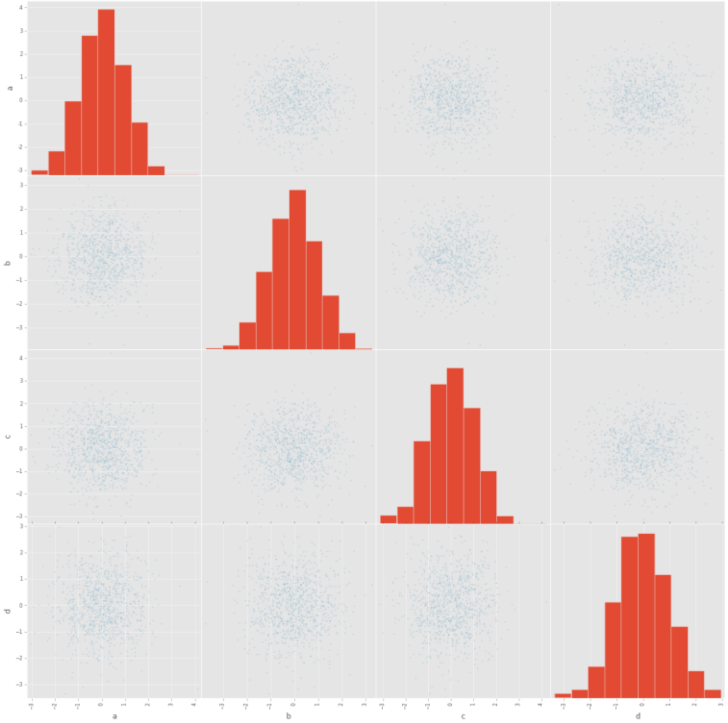


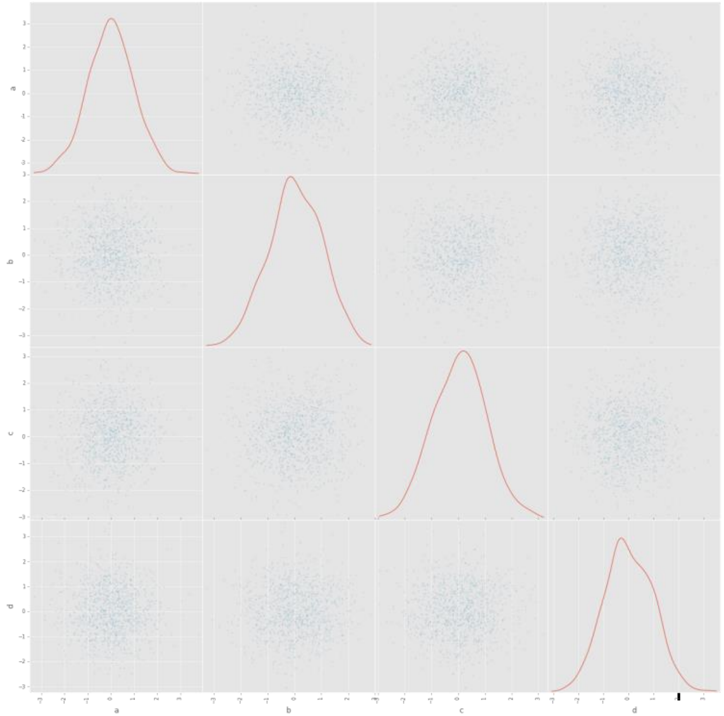
Distribuciones de frecuencia multivariantes: Representaciones Gráficas



Imagenes extraídas de https://es.wikipedia.org/wiki/Medidas_de_dispersi%C3%B3n

Distribuciones de frecuencia multivariantes: Representaciones Gráficas





Medidas de dependencia lineal: Covarianza

- La covariancia es la medida más utilizada como una medida descriptiva de la posible relación lineal que puede existir entre un par de variables:
  - $\mathrm{Cov}(x, y) = \sum_i ((x_i - \langle x \rangle)(y_i - \langle y \rangle)) / n$, siendo este sumatorio sobre todas las posibles n parejas de valores (x,y).
  - Se puede deducir que una expresión equivalente es:
    - $\mathrm{Cov}(x, y) = \{\sum_i x_i y_i / n\} - \langle x \rangle \langle y \rangle$.
  - Si los datos están agrupados por clases, con frecuencia relativa de cada clase $f_r(\mathbf{x_i}, \mathbf{y_i})$, entonces:
    - $\mathrm{Cov}(x, y) = \sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)\, f_r(\mathbf{x_i}, \mathbf{y_i})$.

Medidas de dependencia lineal: Covariancia

- La covarianza fue introducida por K. Pearson para medir la relación lineal entre las variables, como en la figura anterior.
- El **signo positivo** de la covarianza indica que cuando una variable está por encima de su media, es esperable que la otra variable este por encima de la suya también (producto positivo en el sumatorio).
- El **signo negativo** de la covarianza indica que cuando una variable está por encima de su media, es esperable que la otra variable este por debajo de la suya (producto negativo en el sumatorio).

Medidas de dependencia lineal: Correlación

- La covariancia depende de las unidades de las variables a analizar:
  - p. ej. si calculamos la covarianza entre la estatura medida en centímetros y el peso en gramos y nos sale de valor cov, esta misma covarianza aparecería divida por 10^5, si las medidas utilizadas en la variables son metros y kilogramos.
- La correlación es una medida adimensional que mide la relación lineal entre dos variables, coeficiente de correlación r:
  - $r = Cov(x, y)/s_x s_y$, con s_x y s_y las desviaciones típicas de las variables x e y.
  - Como hemos dividido por un termino que tiene las mismas dimensiones que la Cov, se hace adimensional.

Medidas de dependencia lineal: Correlación

- La elección de las desviaciones típicas como factor dividendo dota a r de una serie de propiedades interesantes:
  - r tiene el mismo signo que Cov.
  - r es adimensional: r no varia si multiplicamos a **x** por k_1 y a **y** por k_2, siempre que k_1 y k_2 sean no nulos y del mismo signo.
  - Si existe una relación pura lineal entre **x** e **y**, i.e. $y=a+bx$, entonces:
    - $r=1$ sii $b>0$ o $r=-1$ sii $b<0$.
  - Si no existe un relación lineal pura entonces r tiene un valor comprendido ente 1 y -1.
- Así el coeficiente de correlación puede resumir un diagrama de dispersión, como hemos visto en la figura anterior.

Recta de regresión

- Cuando existe una relación lineal entre variables, los puntos se agrupan en una línea en un diagrama de dispersión y la forma natural de describir el sistema es mediante la mejor recta que ajusta el diagrama de dispersión.
- De la misma forma que describimos **una variable** con su **media** y **dispersión**, podemos describir **dos variables** con una **recta**, y su **dispersión**: la que existe entre los puntos y la recta.
- Supongamos que queremos minimizar los errores de la variable **y** cuando conocemos el valor de **x**, entonces la recta será de la forma:
  - $h(x)=a+bx$

Recta de regresión

- Así si minimizamos $\sum (\mathbf{y}_i - h(\mathbf{x}))^2 = \sum (\mathbf{y}_i - a - b\mathbf{x}_i)^2$, y obtenemos la recta que mejor se ajusta al diagrama de dispersión lineal.
- Minimizar significa derivar respecto los parámetros a optimizar, i.e. respecto a y b, e igualar a 0 la derivada.
- Por tanto derivamos respecto a ambos coeficientes e igualando a 0, y resolvemos el sistema de 2 ecuaciones con dos incógnitas:
  - $d\,[\sum (\mathbf{y}_i - a - b\mathbf{x}_i)^2]/da = 2\sum (\mathbf{y}_i - a - b\mathbf{x}_i)\,(-1) = 0$,
  - $d\,[\sum (\mathbf{y}_i - a - b\mathbf{x}_i)^2]/db = 2\sum (\mathbf{y}_i - a - b\mathbf{x}_i)\,(-\mathbf{x}_i) = 0$

Recta de regresión

- Dividiendo por n, i.e. el número de parejas (x_i, y_i), obtenemos:
  - $<y>=a+b<x>$, que indica que la recta debe pasar por el centro de la nube,
  - $\sum(\mathbf{x}_i\ \mathbf{y}_i)/n = a<x> + b \sum \mathbf{x}_i^2/n$.
- Si eliminamos el término a de la segunda ecuación restando la primera multiplicada por $<x>$ obtenemos:
  - $\sum_i x_i y_i / n - <x><y> = b\,(\sum(x_i^2/n) - <x>^2)$

$s=[\sum(x_i^2/n) - <x>^2]^{1/2}$

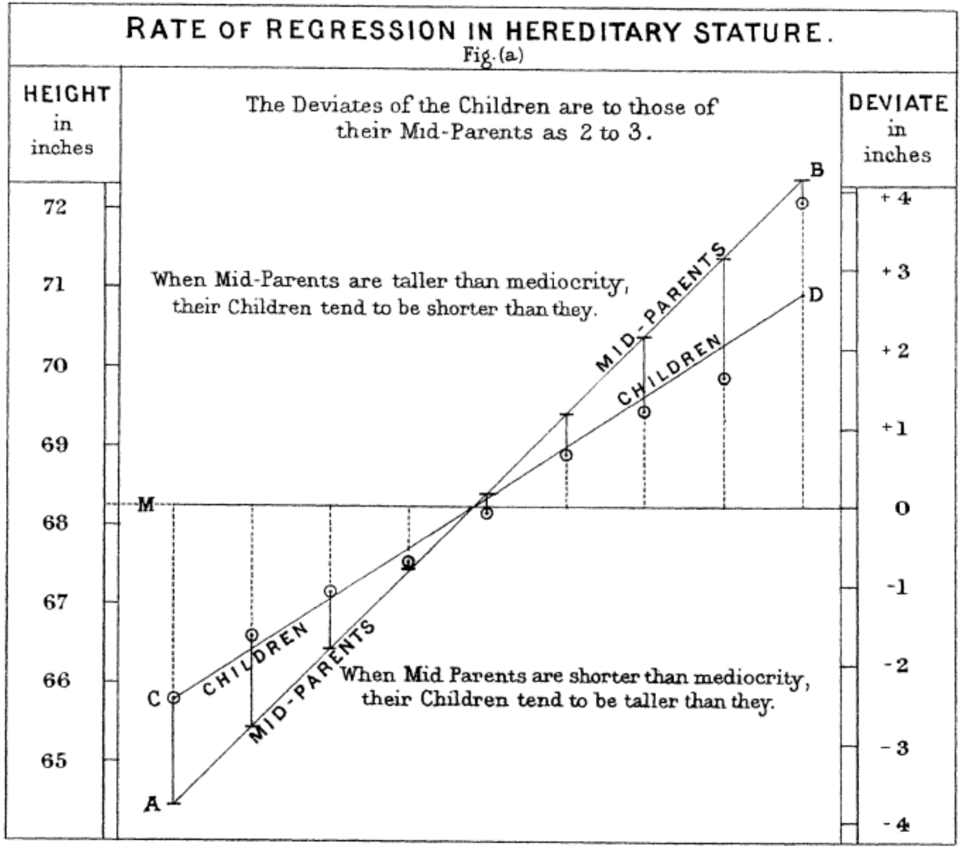
$\mathrm{Cov}(x, y)= \{\sum_i x_i y_i / n\} - <x><y>$

- Así $b = \mathrm{Cov}(x,y) / s_x^2$, indicando que la pendiente de la recta es la covarianza estandarizada para que tenga unidades de x/y como corresponde a la pendiente de una recta.

Recta de regresión

- Sustituyendo en la ecuación de la recta $h(x)=a+bx$, los valores $a=<y> - b<x>$ y $b= Cov(x,y)/s_x^2$, se obtiene (ver ej. 3.4):
  - $h(x)=<y> + (Cov(x,y)/s_x^2)(x - <x>)$, i.e. la regresión lineal (su nombre por F. Galton y su famosa regresión de la media).
  - También b se puede expresar como (es mejor para cálculo):
    - $b= Cov(x,y)/s_x^2= \sum_i ((x_i-<x>)(y_i-<y>)) / \sum_i (x_i-<x>)^2$.
- La medida de variabilidad de los datos respecto a la recta calculada (al igual que calculábamos la varianza en la descripción de un sola variable) se denomina desviación típica residual:
  - $S_R=[\sum (\mathbf{y}_i - h(\mathbf{x}_i))^2 /n]^{1/2}$, es el promedio de las desviaciones verticales de la recta con los datos.

Recta de regresión



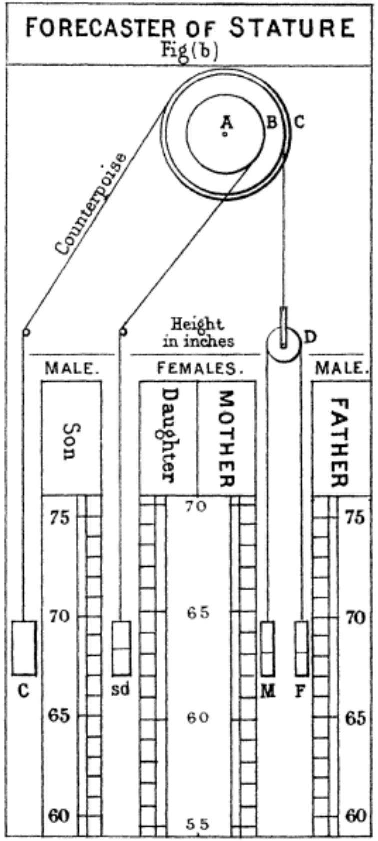
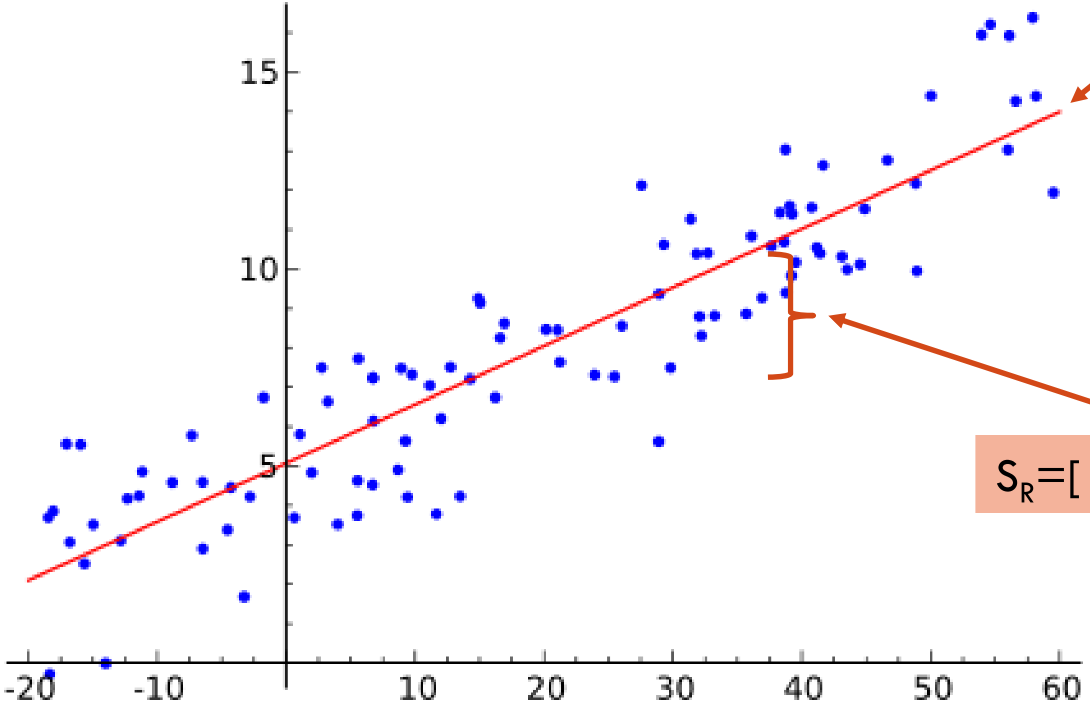


Imagen extraída de:
Galton, Francis (1886). "Regression Towards Mediocrity in Hereditary Stature". The Journal of the Anthropological Institute of Great Britain and Ireland. 15: 246–263. doi:10.2307/2841583.

Recta de regresión

$h(x) = \langle y \rangle + Cov(x,y)/s_x^2 \, (x - \langle x \rangle)$



Parte de la foto sacada de https://en.wikipedia.org/wiki/Linear_regression

$S_R = [\, \sum (\mathbf{y}_i - h(\mathbf{x}_i))^2 / n]^{1/2}$

Vector de medias y matriz de varianzas y covarianzas

- Supongamos que tomamos varias medidas $X_1 \dots X_n$ de más de dos variables, por ejemplo 3:
- $X_i = \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix}$
- El vector de medias por ejemplo para una variable tridimensional:
  - $< X > = \begin{bmatrix} < x > \\ < y > \\ < z > \end{bmatrix} = \frac{1}{n} \begin{bmatrix} \sum x_i \\ \sum y_i \\ \sum z_i \end{bmatrix} = \frac{1}{n} \sum X_i$
  - Y en general $<X>=1/n \sum X_i$
- La matriz de covarianzas para dos variables:
  - $M = \begin{bmatrix} S_x^2 & cov(x,y) \\ cov(y,x) & S_y^2 \end{bmatrix}$, es simétrica ya que cov(x,y)=cov(y,x).
- Esta matriz se puede extender a un problema k-dimensional, generando una matriz simétrica de igual forma.

Vector de medias y matriz de varianzas y covarianzas

- Para caso k-dimensional variable, si definimos s_i^2 como la varianza de la variable i y s_{ij} como la covarianza de la variable i con la j, la matriz de covarianzas quedaría:
  - $M = \begin{bmatrix} S_1^2 & \cdots & S_{1k} \\ \vdots & \ddots & \vdots \\ S_{k1} & \cdots & S_k^2 \end{bmatrix}$
- La matriz de covarianzas se calcula conociendo el vector de medias como:
  - $M = \frac{1}{n}\sum(X_i - < X >)(X_j - < X >)$, por ejemplo para tres variables:
  - $M = \frac{1}{n}\sum \begin{bmatrix} x_i - < x > \\ y_i - < y > \\ z_i - < z > \end{bmatrix} [x_i - < x > \quad y_i - < y > \quad z_i - < z >]$, i.e.
  - $M =$

$$\frac{1}{n}\sum \begin{bmatrix} (x_i - < x >)^2 & (x_i - < x >)(y_i - < y >) & (x_i - < x >)(z_i - < z >) \\ (y_i - < y >)(x_i - < x >) & (y_i - < y >)^2 & (y_i - < y >)(z_i - < z >) \\ (z_i - < z >)(x_i - < x >) & (z_i - < z >)(y_i - < y >) & (z_i - < z >)^2 \end{bmatrix}$$

Ejemplos para realizar en casa

- Realizar para casa los ejemplos del libro:
  - 3-1, 3-2, 3-3, 3-4, 3-5, 3-6 y 3-7.

Ejemplos para realizar en casa (ejemplo 3.1, tabla 3.3)

	TRABAJADORES				
Ventas	**1-24**	**25-59**	**50-74**	**75-99**	**Total**
1-100	0,28	0,07	0,01	0,00	**0,36**
101-200	0,10	0,15	0,06	0,02	**0,33**
201-300	0,04	0,10	0,08	0,09	**0,31**
TOTAL	**0,42**	**0,32**	**0,15**	**0,11**	

Frecuencias relativas del volumen de ventas y número de trabajadores para un grupo de 200 empresas pequeñas y medianas