

Etapas de una investigación

La Estadística nos permite realizar inferencias y sacar conclusiones a partir de los datos. Extrayendo la información que contienen, podremos comprender mejor las situaciones que ellos representan.

Los métodos estadísticos abarcan todas las etapas de la investigación, desde el diseño de la investigación hasta el análisis final de los datos.

Podemos distinguir tres grandes etapas:

1. **Diseño:** Planeamiento y desarrollo de las investigaciones
2. **Descripción:** Resumen y exploración de los datos
3. **Inferencia:** Predicciones y toma de decisiones sobre las características de una población en base a la información recogida en una muestra de la población.

En la etapa de **Diseño** se define cómo se desarrollará la investigación con el fin de responder las preguntas que le dieron origen. Un diseño bien realizado puede ahorrar esfuerzos en etapas posteriores y puede redundar en un análisis más sencillo. Esta etapa es crucial, pues un estudio pobremente diseñado o con datos incorrectamente recolectados o registrados puede ser incapaz de responder las preguntas que originaron el estudio.

Una vez formulado el problema, en la etapa de Diseño se definirá, entre otras cosas, la población objetivo, los tamaños de muestra, los mecanismos de selección de individuos, los criterios de inclusión y exclusión de sujetos, los métodos de asignación de tratamientos, las variables que se medirán y cómo se entrenará al equipo de trabajo para el cumplimiento del protocolo.

Los métodos de **Análisis Exploratorio** o **Estadística Descriptiva** ayudan a comprender la estructura de los datos, de manera de detectar tanto un patrón de comportamiento general como apartamientos del mismo. Una forma de realizar esto es mediante gráficos de sencilla elaboración e interpretación. Otra forma de describir los datos es resumiéndolos en uno, dos o más números que caractericen al conjunto de datos con fidelidad. Explorar los datos permitirá detectar datos erróneos o inesperados y nos ayudará a decidir qué métodos estadísticos pueden ser empleados en etapas posteriores del análisis de manera de obtener conclusiones válidas.

Finalmente, la **Inferencia Estadística** nos permite tanto hacer predicciones y estimaciones como decidir entre dos hipótesis opuestas relativas a la población de la cual provienen los datos (test de hipótesis).

La calidad de las estimaciones puede ser muy variada y está afectadas por errores. La ventaja de los métodos estadísticos es que, aplicados sobre datos obtenidos a partir de muestras aleatorias, permiten cuantificar el error que podemos cometer en una estimación o calcular la probabilidad de cometer un error al tomar una decisión en un test de hipótesis.

Para entender qué tipo de problemas consideraremos en Estadística tomemos, por ejemplo, las siguientes mediciones de la proporción de la masa de la Tierra con respecto a la Luna

Mariner II	81.3001
Mariner IV	81.3015
Mariner V	81.3006
Mariner VI	81.3011
Mariner VII	81.2997
Pioneer VI	81.3005
Pioneer VII	81.3021

En *Probabilidad* podríamos suponer que las posibles mediciones se distribuyen alrededor del verdadero valor 81.3035 siguiendo una distribución determinada y nos preguntaríamos

¿Cuál es la probabilidad de que se obtengan 7 mediciones menores que el verdadero valor?

En *Estadística*, a partir de los 7 observaciones nos preguntaríamos:

¿Son consistentes los datos con la hipótesis de que el verdadero valor es 81.3035?

¿Cuán confiable es decir que el verdadero valor está en el intervalo (81.2998, 81.3038)?

Las técnicas del análisis exploratorio nos ayudan a organizar la información que proveen los datos, de manera de detectar algún patrón de comportamiento así como también apartamientos importantes al modelo subyacente. Nos guían a la estructura subyacente en los datos de manera rápida y simple.

Estadística Descriptiva

Examinaremos los datos en forma descriptiva con el fin de:

- Organizar la información
- Sintetizar la información
- Ver sus características más relevantes
- Presentar la información

Definimos:

Población: conjunto total de los sujetos o unidades de análisis de interés en el estudio

Muestra: cualquier subconjunto de sujetos o unidades de análisis de la población en estudio.

Unidad de análisis o de observación: objeto bajo estudio. Puede ser una persona, una familia, un país, una institución o en general, cualquier objeto.

Variable: cualquier característica de la unidad de observación que interese registrar y que en el momento de ser registrada puede ser transformada en un número.

Valor de una variable, Dato, Observación o Medición: número que describe a la característica de interés en una unidad de observación particular.

Caso o Registro: conjunto de mediciones realizadas sobre una unidad de observación.

Datos cuantitativos

Esquema de Tallo y Hoja

Nos da una primera aproximación rápida a la distribución de los datos sin perder de vista las observaciones.

Ejemplo: La siguiente tabla contiene 45 observaciones correspondientes a la fuerza de compresión de cierta aleación de Aluminio-Litio.

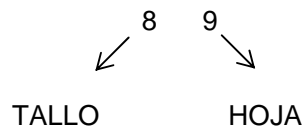
96	93	88	117	127	95	113	96
108	94	148	156	139	142	94	107
125	155	155	103	112	127	117	120
112	135	132	111	125	104	106	139
134	119	97	89	118	136	125	143
120	103	113	124	138			

- Ordenamos los datos de menor a mayor

88	89	93	94	94	95	96	96
97	103	103	104	106	107	108	111
112	112	113	113	117	117	118	119
120	120	124	125	125	125	127	127
132	134	135	136	138	139	139	142
143	148	155	155	156			

- Separamos a cada observación en dos partes: **tallo** y **hoja**
- Listamos en forma vertical y creciente los tallos y agregamos las hojas a la derecha del tallo correspondiente.

Ejemplo. Consideremos el segundo dato :



Elegimos un número de dígitos a la derecha de cada número que corresponderán a las hojas: 1 en este caso.

Separamos esos dígitos de los restantes, que constituirán los tallos. En este caso obtendremos 8 tallos, de 8 a 15.

9	3445667
10	334678
11	122337789
12	00455577
13	2456899
14	238
15	556

¿Qué podemos ver en este tipo de diagrama?

- Rango de las observaciones, valores máximo y mínimo.
- Forma de la distribución: simetría, asimetría a derecha, asimetría a izquierda y cuántos picos tiene la distribución.
- Posición del centro de la distribución y concentración de los datos.
- Desviaciones marcadas respecto al comportamiento general: outliers o valores atípicos.

Ejemplo: Los siguientes datos corresponden a tiempos de falla de cables Kevlar 49/epoxy sometidos a una presión del 90%:

TIEMPOS DE FALLA															
0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.04	0.05	0.06	0.07	0.07	0.08	0.09	0.09	0.10
0.10	0.11	0.11	0.12	0.13	0.18	0.19	0.20	0.23	0.80	0.80	0.83	0.85	0.90	0.92	0.95
0.99	1.00	1.01	1.02	1.03	1.05	1.10	1.10	1.11	1.15	1.18	1.20	1.29	1.31	1.33	1.34
1.40	1.43	1.45	1.50	1.51	1.52	1.53	1.54	1.54	1.55	1.58	1.60	1.63	1.64	1.80	1.80
1.81	2.02	2.05	2.14	2.17	2.33	3.03	3.03	3.24	4.20	4.69	7.89				

El correspondiente esquema de tallo y hoja resulta:

0	0000000000000001111111122
0	88889999
1	000001111122333444
1	55555555666888
2	00113
2	
3	002
3	
4	2
4	6
5	
5	
6	
6	
7	
7	8

En este caso cada tallo ha sido dividido en 2 líneas: en la primera se listan las hojas 0 a 4 y en la segunda las hojas 5 a 9.

Se observa asimetría a derecha y un valor alejado del resto: 7.8

El número de tallos debe ser tal que permita mostrar una imagen general de la estructura del conjunto de datos. Aunque existen algunos criterios para definir el número de tallos, la decisión depende fundamentalmente del sentido común. Demasiados detalles en general serán poco informativos, demasiado agrupamiento puede distorsionar la imagen del conjunto.

Cuando el volumen de datos es muy grande conviene usar otro tipo de gráficos que también son de fácil interpretación .

Ejemplo: Consideremos el siguiente ejemplo con datos sobre consumo diario *per cápita* de proteínas en 32 países desarrollados. Los datos se presentan ordenados de menor a mayor por simplicidad.

Consumo de proteínas per cápita en países desarrollados.

7.83	9.03	10.56
8.06	9.16	10.52
8.45	9.23	10.75
8.49	9.34	10.86
8.53	9.39	10.89
8.60	9.42	11.07
8.64	9.56	11.27
8.70	9.89	11.36
8.75	10.00	11.58
8.92	10.28	11.76
8.93	10.41	

Seleccionando como tallo la unidad obtenemos el gráfico de tallo-hojas de la izquierda de la figura. En este gráfico se acumula un número importante de hojas en cada tallo, por lo que podríamos estar perdiendo información acerca de la estructura de los datos. En el gráfico de la derecha, cada tallo ha sido dividido en dos líneas, en la primera se listan las hojas 0 a 4 y en la segunda as hojas 5 a 9.

Como puede observarse, al expandir la escala se observan más detalles y parece haber dos “grupos” de países, uno con mayor consumo per cápita de proteínas y otro con menor consumo, ya que la distribución de la variable tiene dos picos.

Variación del número de tallos. Datos de consumo de proteínas per cápita.

7		8		7		8
8		0 4 4 5 6 6 7 7 9 9		8		0 4 4
9		0 1 2 3 3 4 5 8		8		5 6 6 7 7 9 9
10		0 2 4 5 5 7 8 8		9		0 1 2 3 3 4
11		0 2 3 5 7		9		5 8
				10		0 2 4
				10		5 5 7 8 8
				11		0 2 3
				11		5 7

El problema de expandir la escala es que podrían comenzar a aparecer detalles superfluos, o simplemente atribuibles al azar.

Gráfico de tallo-hojas espalda con espalda. Comparación de grupos.

Los gráficos de tallo-hojas son útiles para comparar la distribución de una variable en dos condiciones o grupos. El gráfico se denomina tallo-hojas espalda con espalda porque ambos grupos comparten los tallos.

A continuación se muestra un gráfico de la presión arterial sistólica (PAS) a los 30 minutos de comenzada la anestesia en pacientes sometidos a dos técnicas anestésicas diferentes a las que nos referiremos como T1 y T2.

Comparación de la presión arterial sistólica en pacientes sometidos a dos técnicas anestésicas (30 minutos del inicio de la anestesia).

T1	T2
5	47
6	2
74	7 37
963	8 778999
660	9 0358
9662	10 222
821	11 37
70	12
2	13
	14
	15
4	16

El gráfico nos muestra las siguientes características de la PAS en los dos grupos de pacientes.

- La distribución de PAS tiene *forma* similar en ambos grupos: Un pico o moda y forma simétrica y aproximadamente acampanada.
- Diferencias en *posición*. Los pacientes del grupo T1 tienen niveles de PAS levemente mayores que los pacientes del grupo T2.
- Similar *dispersión*. Los valores de PAS de los pacientes de ambos grupos se encuentran en rangos aproximadamente iguales, salvo por el valor atípico (*outlier*) que se observa en el grupo T1.

Histograma

- Se divide el rango de los datos en **intervalos o clases**, que no se superpongan. Las clases deben ser **excluyentes y exhaustivas**.
- Se cuenta la cantidad de datos en cada intervalo o clase, es decir la **frecuencia**. También se puede usar para cada intervalo la

$$\text{frecuencia relativa} = \frac{\text{frecuencia}}{\text{cantidad total de datos}}$$

- Se grafica el histograma en un par de ejes coordenados representando en las abscisas los intervalos y sobre cada uno de ellos un rectángulo cuya área sea proporcional a la frecuencia relativa de dicho intervalo.

Observaciones:

- No existen criterios óptimos para elegir la cantidad de intervalos. En general, entre 8 y 15 intervalos deberían ser suficientes. Utilizar muchos o muy pocos intervalos puede ser poco informativo. Se debe buscar un equilibrio entre un histograma muy irregular y uno demasiado suavizado.
- No es necesario que todos los intervalos tengan la misma longitud, pero es recomendable que así sea. Ésto facilita su interpretación.
- El histograma representa la frecuencia o la frecuencia relativa a través del **área** y no a través de la altura.
- Es recomendable tomar

$$\text{altura del rectángulo} = \frac{\text{frecuencia relativa}}{\text{longitud del intervalo}}$$

De esta manera el área es 1 y dos histogramas son fácilmente comparables independientemente de la cantidad de observaciones en las que se basa cada uno.

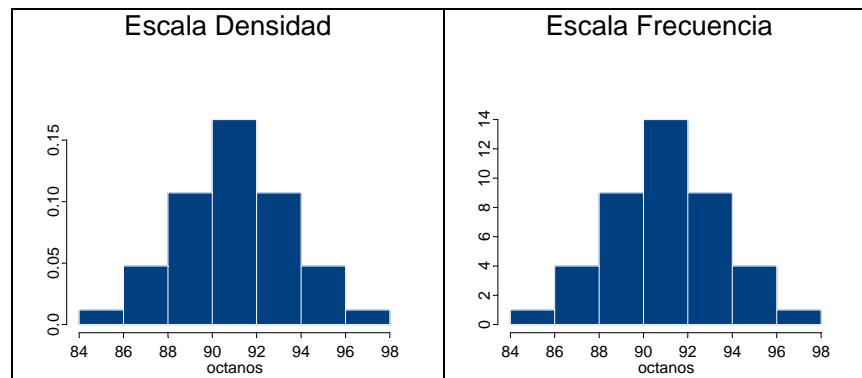
Ejemplo: Los siguientes datos corresponden a Porcentajes de Octanos en Naftas:

85.3	87.5	87.8	88.5	89.9	90.4	91.8	92.7
86.7	87.8	88.2	88.6	90.3	91.0	91.8	93.2
88.3	88.3	89.0	89.2	90.4	91.0	92.3	93.3
89.9	90.1	90.1	90.8	90.9	91.1	92.7	93.4
91.2	91.5	92.6	92.7	93.3	94.2	94.7	94.2
95.6	96.1						

Los agrupamos en 7 clases:

Clase	Frecuencia f_i	Frecuencia relativa fr_i
[84, 86]	1	0.02380952
(86, 88]	4	0.09523810
(88, 90]	9	0.21428571
(90,92]	14	0.33333333
(92,94]	9	0.21428571
(94,96]	4	0.09523810
(96,98]	1	0.02380952
Total	42	1

Histogramas para datos de OCTANOS

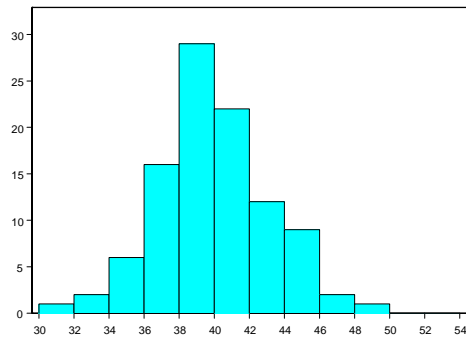


En general, si el histograma es muy irregular puede ser imposible descubrir la forma. En ese caso es conveniente tomar intervalos más anchos.

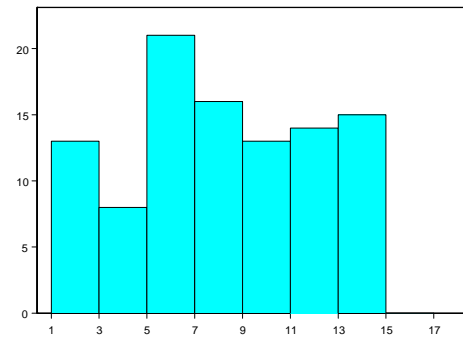
¿Qué formas puede tener un histograma?

Un aspecto a tener en cuenta en la distribución de los datos es la simetría. Un conjunto de datos que no se distribuye simétricamente, se dice que es **asimétrico**. La asimetría puede verse en el esquema de Tallo y Hoja o en el Histograma y también puede apreciarse a través de la posición relativa entre media y mediana. Más adelante, en un boxplot lo veremos a través de la posición relativa entre la mediana y los cuartos. En los siguientes gráficos mostramos algunas de las formas posibles que puede tener un histograma:

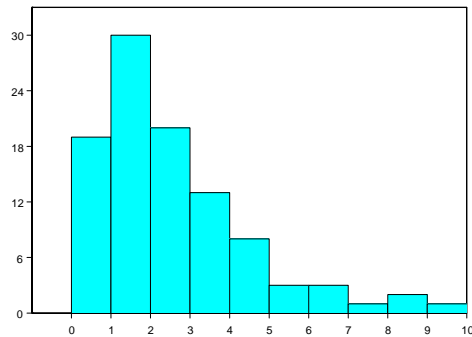
Distribución acampanada



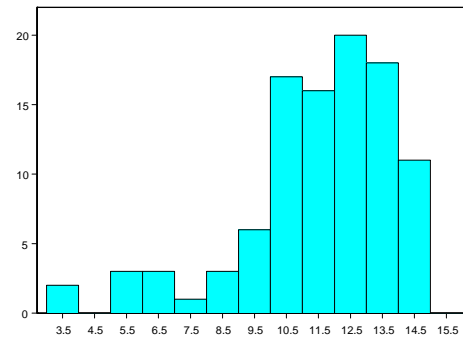
Distribución uniforme



Asimetría a derecha



Asimetría a izquierda



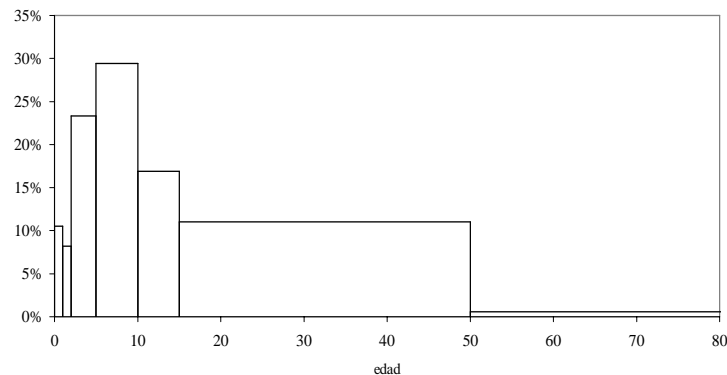
Histograma con intervalos de distinta longitud

Los datos de la siguiente tabla presentan los casos de rubéola notificados al SINAVE durante el año 2000 según grupos de edad. Notemos que los intervalos de edad tienen diferente longitud.

Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE

Intervalo (años)	Frecuencia (f_i)	Frecuencia relativa (f_r)
[0, 1)	497	10.5%
[1, 2)	387	8.2%
[2, 5)	1100	23.3%
[5, 10)	1389	29.4%
[10, 15)	798	16.9%
[15, 50)	521	11.0%
≥ 50	28	0.6%
Total	4720	100.00%

Si **erróneamente** se construye un histograma considerando como altura de la barra la frecuencia relativa se obtiene la gráfica siguiente. La última categoría de edad se truncó arbitrariamente en 80 años para poder representarla.



A partir de este gráfico concluiríamos que la proporción de casos es notablemente mayor en los grupos de 2 a 5 años, de 5 a 10 años o de 10 a 15 años que en los grupos de menores de 1 año o de 1 a 2 años. Además, la proporción de casos en el grupo de 15 a 50 años impresiona como notable.

El problema es que en la imagen visual asociamos la frecuencia de casos con el área de la barra, por ello parece haber más notificaciones de gente de 15 a 50 que de cualquier otro grupo de edad.

Recordemos que la barra debe tener una altura tal que el área (base x altura) sea igual a la frecuencia (o a la frecuencia relativa). Es decir,

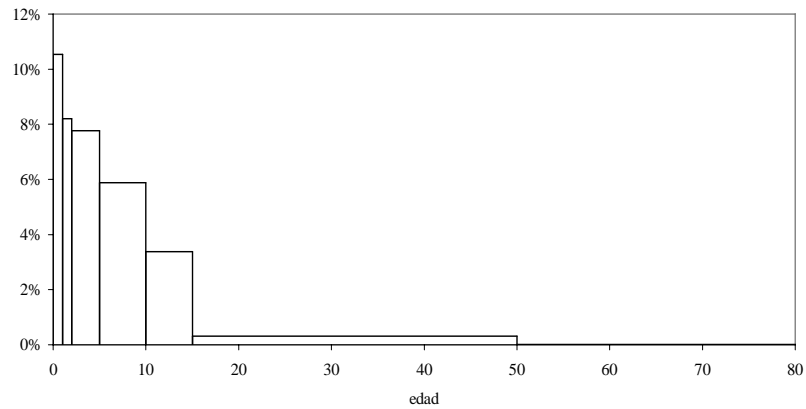
$$\text{altura de la barra} = \frac{\text{frecuencia en el intervalo}}{\text{longitud del intervalo}}.$$

De este modo el área de la barra coincide con la frecuencia en el intervalo. La altura de la barra definida de este modo se denomina *escala densidad* porque indica el número de datos por unidad de la variable. La última columna de la siguiente tabla muestra la escala densidad para los datos de rubéola y la figura siguiente presenta el histograma que se obtiene usando la escala densidad.

*Escala densidad. Notificaciones de casos de rubéola. Argentina, año 2000.
 Fuente: SINAVE.*

Categoría (años)	Frecuencia (f _i)	Frecuencia relativa (f _r)	Escala densidad
[0, 1)	497	10.5%	10.53%
[1, 2)	387	8.2%	8.20%
[2, 5)	1100	23.3%	7.77%
[5, 10)	1389	29.4%	5.89%
[10, 15)	798	16.9%	3.38%
[15, 50)	521	11.0%	0.32%
≥ 50	28	0.6%	0.01%
Total	4720	100.00%	--

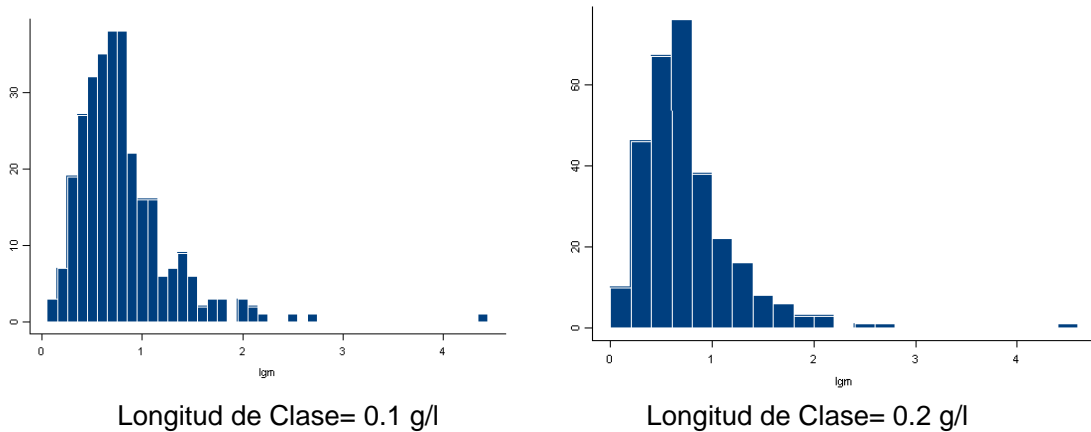
Histograma usando escala densidad. Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE

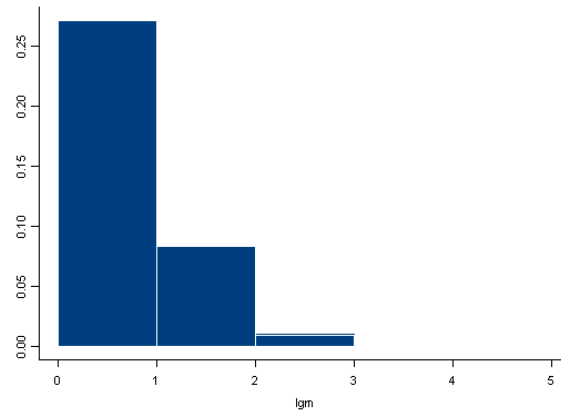


En este gráfico, el porcentaje de casos de rubéola notificados para cada grupo está representado en el *área de la barra*. El histograma muestra que una alta proporción de casos ocurre en menores de 5 años y que la proporción desciende a medida que aumenta la edad. En este gráfico estamos representando la “densidad de notificaciones” por cada año de edad.

El siguiente ejemplo nos muestra cómo varía el aspecto del histograma según la longitud de las clases.

Ejemplo: Concentración de I_{mg}





Longitud de clase=1g/l

Medidas de Resumen

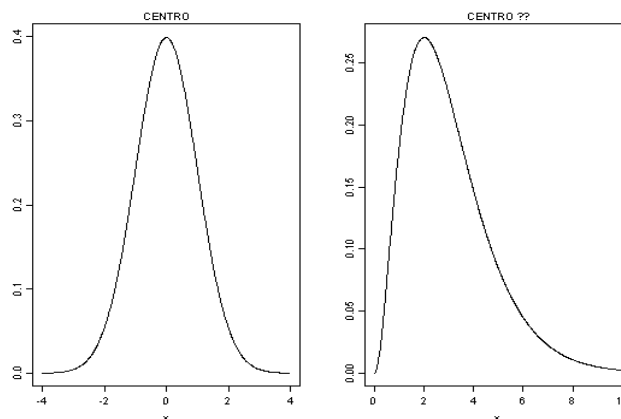
Resumiremos la información de los datos provenientes de variables numéricas mediante medidas de fácil interpretación que reflejen sus características más relevantes. La medida a elegir dependerá de cada problema.

Medidas de Posición o Centrado

Un modo de resumir un conjunto de datos numéricos es a través de un número que represente a todos, en el sentido de ser un valor *típico* para el conjunto.

La pregunta que intentamos responder es: *¿Cuál es el valor central o que mejor representa a los datos?*

Si la distribución es simétrica diferentes medidas darán resultados similares. Si es asimétrica no existe un centro evidente y diferentes criterios para resumir los datos pueden diferir considerablemente, en tanto tratan de captar diferentes aspectos de los mismos.



Supongamos que tenemos un conjunto de n datos que genéricamente representaremos por:

$$x_1, x_2, \dots, x_n$$

Promedio o Media Muestral:

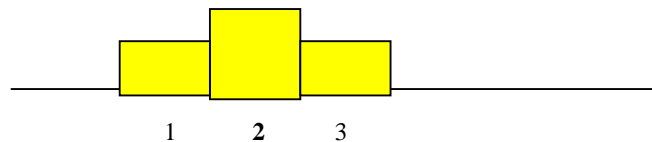
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Es el punto de equilibrio del conjunto de datos.

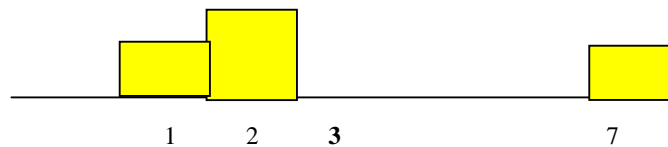
Ejemplo: Fuerza de compresión de cierta aleación de Aluminio-Litio

$$\bar{x} = \frac{\sum_{i=1}^{45} x_i}{45} = \frac{5350}{45} = 118.89$$

Ejemplo: Supongamos que las observaciones son: 1, 2, 2, 3. En este caso $\bar{x} = 2$.



Si reemplazamos el valor 3 por 7, las observaciones son: 1, 2, 2, 7 y $\bar{x} = 3$.



La media muestral es una medida muy sensible a la presencia de datos anómalos (outliers).

Mediana Muestral: Es una medida del centro de los datos en tanto divide a la muestra ordenada en dos partes de igual tamaño. Deja la mitad de los datos a cada lado.

Sean los estadísticos de orden muestrales:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

Definimos como mediana

$$\tilde{x} = \begin{cases} x_{(k+1)} & \text{si } n = 2k + 1 \\ \frac{x_{(k)} + x_{(k+1)}}{2} & \text{si } n = 2k \end{cases}$$

La mediana es resistente a la presencia de datos atípicos. También puede ser útil cuando algunos datos han sido censurados.


Ejemplos:

1) Supongamos que los datos son: 3, 5, 2, 4, 6, 8, 7, 7, 6 . Como $n = 9$, $(n+1)/2 = 5$.

Ordenamos la muestra: 2 3 4 5 6 6 7 7 8


 $\tilde{x} = 6$

2) Supongamos que los datos son: 3, 5, 2, 4, 6, 8, 7, 7. Como $n = 8$, $(n+1)/2 = 4.5$ y por lo tanto la mediana muestral es el promedio de las observaciones que ocupan las posiciones 4 y 5 en la muestra ordenada.

2 3 4 5 6 7 7 8

 $\tilde{x} = 5.5$

Ejercicios: 1) Consideremos los dos conjuntos de datos siguientes:

x's: 1,2,2,3	$\bar{x} = 2$	$\tilde{x} = 2$
y's: 1,2,2,7	$\bar{y} = 3$	$\tilde{y} = 2$

¿Qué pasa si, en el segundo caso, se registra 70 en lugar de 7?

2) Dada una muestra de salarios de cierta población, ¿sería más adecuado tomar la media o la mediana muestral para representarla?

Media α - Podada: Es un promedio calculado sobre los datos una vez que se han eliminado $\alpha \cdot 100\%$ de los datos más pequeños y $\alpha \cdot 100\%$ de los datos más grandes. Es una medida intermedia entre la media y la mediana. Formalmente podemos definirla como:

$$\bar{x}_\alpha = \frac{x_{([n\alpha]+1)} + \dots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

es decir, se obtiene promediando los datos luego de eliminar un número de observaciones en cada extremo de la muestra ordenada igual a la parte entera de $(n\alpha)$.

Otra posible manera de definirla es eliminando $(n\alpha)$ datos en cada extremo si $(n\alpha)$ es entero y, cuando no lo es, interpolando entre dos medias α -podadas, una en la cual se podan $[n\alpha]$ en cada extremo y otra en la que se podan $[n\alpha]+1$ datos en cada extremo.

Ejemplos: 1) Sea el siguiente conjunto de 10 observaciones, ya ordenadas

2 5 8 10 14 17 21 25 28 40

y calculemos la media 0.10-podada. Debemos podar 1 dato en cada extremo y calcular el promedio de los 8 datos restantes, es decir

$$\bar{x}_{0.10} = \frac{5 + 8 + 10 + 14 + 17 + 21 + 25 + 28}{8} = \frac{128}{8} = 16$$

2) Sea el siguiente conjunto de 12 observaciones, ya ordenadas

1 2 5 8 10 14 17 21 25 28 40 45

y calculemos la media 0.10-podada. Usando la definición dada inicialmente, debemos podar $[12 \cdot 0.10] = [1.2] = 1$ dato en cada extremo y calcular el promedio de los 10 datos restantes, es decir

$$\bar{x}_{0.10} = \frac{2 + 5 + 8 + 10 + 14 + 17 + 21 + 25 + 28 + 40}{10} = \frac{170}{10} = 17$$

Con la segunda definición, deberíamos calcular dos medias, una podando una observación en cada extremo de la muestra ordenada y otra podando dos observaciones en cada extremo, e interpolar linealmente entre ambas medias. Es decir, calculamos

$$\begin{aligned}\bar{x}_1 &= \frac{2 + 5 + 8 + 10 + 14 + 17 + 21 + 25 + 28 + 40}{10} = \frac{170}{10} = 17 \\ \bar{x}_2 &= \frac{5 + 8 + 10 + 14 + 17 + 21 + 25 + 28}{8} = \frac{128}{8} = 16\end{aligned}$$

y la media podada se obtiene como la ordenada correspondiente a $x = 1.2$ en la recta que pasa por $(1, 17)$ y $(2, 16)$:

$$\bar{x}_{0.10} = 16.8$$

Observemos que la media es una media α -podada con $\alpha = 0$ y la mediana una media podada con α tan próximo a 0.5 como sea posible. En ese sentido, la media podada es una medida intermedia entre la media y la mediana. Es más resistente a datos atípicos que la media.

¿Cómo elegimos α ?

Dependiendo de cuantos outliers se pretende excluir y de cuán robusta queremos que sea la medida de posición. Como dijimos, cuando seleccionamos $\alpha = 0$ tenemos la media, si elegimos el máximo valor posible para α (lo más cercano posible a 0.5) obtenemos la mediana. Cualquier poda intermedia representa un compromiso entre ambas. Una elección bastante común es $\alpha = 0.10$, que excluye un 20% de los datos.

Ejemplo: En este ejemplo calcularemos las tres medidas resumen. Los datos siguientes, ya ordenados, corresponden al número de pulsaciones por minuto en pacientes con asma durante un espasmo:

40 120 120 125 136 150 150 150 150 167

Las correspondientes medidas son:

$$\bar{x} = 130.8 \quad \tilde{x} = 143 \quad \bar{x}_{0.10} = 137.625$$

Si la distribución es simétrica la mediana y la media identifican al mismo punto. Sin embargo, si la distribución de los datos es asimétrica, esperamos que la relación entre ambas siga el siguiente patrón:

$$\text{Asimetría derecha (cola larga hacia la derecha)} \quad \Rightarrow \quad \bar{x} > \tilde{x}$$

$$\text{Asimetría izquierda (cola larga hacia la izquierda)} \quad \Rightarrow \quad \bar{x} < \tilde{x}$$

La mediana puede ser útil cuando algunos datos son censurados. En estos casos es imposible calcular la media muestral, sin embargo suele ser posible computar la mediana.

Ejemplos: a) Tiempo de supervivencia (en meses) de pacientes con cierta patología. Los datos que se indican entre paréntesis tienen censura a derecha, es decir, se sabe que el paciente sobrevivió ese tiempo, pero no se conoce el tiempo real de supervivencia.

1 5 10 12 18 24 25 28 39 45 (45) 48 50 51 (84) n = 15

Como $n = 15$ la mediana es el octavo dato, por lo tanto $\tilde{X} = 28$. Es posible calcularla aunque haya datos censurados, porque los mismos no participan en el cálculo de la mediana. Por ejemplo, aunque no conocemos exactamente el tiempo que sobrevivió el paciente cuyo dato es (45) sabemos que en esta muestra ese dato ocupará el lugar 11 o uno superior.

b) Si, en cambio, los datos son:

1 5 10 (12) 18 24 25 28 39 45 (45) 48 50 51 (84) n = 15

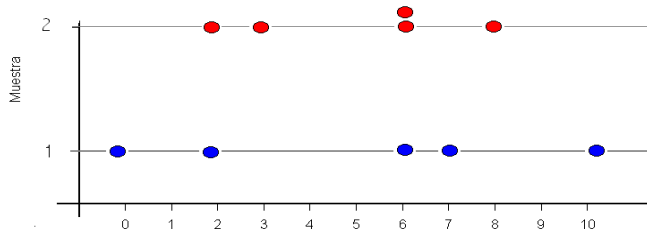
no es posible calcular la mediana debido al dato indicado como (12). Sabemos que este paciente sobrevivió por lo menos 12 meses, pero desconocemos el verdadero valor, el que puede ocupar cualquier posición entre la cuarta y la última.

Medidas de Dispersión o Variabilidad

¿Cuán dispersos están los datos? ¿Cuán cercanos son los datos al valor típico?

Grafiquemos los dos conjuntos de datos siguientes y calculemos para cada uno de ellos su media y su mediana:

x's: 0 2 6 7 10
y's: 2 3 6 6 8



$$\bar{x} = \bar{y} = 5$$
$$\tilde{x} = \tilde{y} = 6$$

A pesar de tener igual media e igual mediana, los conjuntos de datos difieren ¿Cómo medir la diferencia observada?

Rango Muestral: Es la diferencia entre el valor más grande y el más pequeño de los datos:

$$\text{Rango} = \text{máx}(X_i) - \text{mín}(X_i)$$

Ejemplo: en nuestros conjuntos de datos:

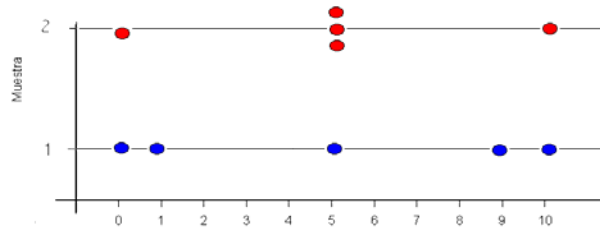
$$\text{Rango}(X) = 10 \quad \text{Rango}(Y) = 6$$

Esta medida es muy sensible a la presencia de outliers. Además no capta la dispersión interna del conjunto de datos.

Veamos otro ejemplo: Sean los siguientes conjuntos de datos

x 's: 0 1 5 9 10

y 's: 0 0 5 5 10



Si calculamos la media, la mediana y el rango muestral de ambos conjuntos, obtenemos:

$$\bar{x} = \bar{y} \quad \tilde{x} = \tilde{y} \quad \text{Rango}(x) = \text{Rango}(y).$$

Es decir, que las 3 medidas coinciden, pero la dispersión no es la misma. Propondremos otra medida de variabilidad.

Varianza Muestral: Mide la variabilidad de los datos alrededor de la media muestral.

$$\text{Varianza muestral} = S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$\text{Desvío Estándar Muestral} = S = \sqrt{S^2}$$

Ejemplo: En los dos conjuntos de datos anteriores obtenemos:

$$S^2_x = 20.5 \quad S_x = 4.258$$

$$S^2_y = 12.5 \quad S_y = 3.536$$

- El desvío estándar tiene las mismas unidades que los datos, mientras que la varianza no.
- Al basarse en promedios, estas medidas son sensibles a la presencia de datos atípicos. Por ejemplo, si en la muestra de los y 's cambiamos el 10 por un 15 obtenemos $S^2_y = 30$ y $S_y = 5.477$, mientras que si lo cambiamos por un 20 obtenemos $S^2_y = 57.5$ y $S_y = 7.583$.

Coefficiente de Variación: Es una medida que relaciona el desvío standard con la media de una muestra.

$$CV = \frac{S}{\bar{x}}$$

Es una medida que está en desuso, ya que no tiene propiedades estadísticas muy interesantes. Sin embargo no depende de las unidades y si lo multiplicamos por 100 nos da una idea de la variabilidad relativa.

Distancia Intercuartil: Es una medida más resistente que el desvío estándar, basada en el rango de los datos centrales de la muestra.

Comenzaremos por definir los **percentiles**. El percentil $\alpha \cdot 100\%$ de la muestra ($0 < \alpha < 1$) es el valor por debajo del cual se encuentra el $\alpha \cdot 100\%$ de los datos en la muestra ordenada.

Para calcularlo:

- Ordenamos la muestra de menor a mayor
- Buscamos el dato que ocupa la posición $\alpha \cdot (n + 1)$. Si este número no es entero se interpolan los dos adyacentes.

Ejemplo: Consideremos los siguientes 19 datos ordenados:

1 1 2 2 3 4 4 5 5 6 7 7 8 8 9 9 10 10 11

↑ ↑ ↑

Percentil	Posición	Valor	
10%	$0.10(19+1) = 2$	1	
25%	$0.25(19+1) = 5$	3	Cuartil Inferior
50%	$0.50(19+1) = 10$	6	Mediana
75%	$0.75(19+1) = 15$	9	Cuartil Superior
95%	$0.95(19+1) = 19$	11	

Notemos que el percentil 50% (o segundo cuartil) coincide con la mediana. Llamaremos cuartil inferior (o primer cuartil) al percentil 25% y cuartil superior (o tercer cuartil) al percentil 75%.

Los cuartiles y la mediana dividen a la muestra ordenada en cuatro partes igualmente pobladas (aproximadamente un 25 % de los datos en cada una de ellas). Entre los cuartiles se halla aproximadamente el 50% central de los datos y el rango de éstos es:

$$d_i = \text{distancia intercuartil} = \text{cuartil superior} - \text{cuartil inferior}.$$

Observación: Si en el ejemplo cambiáramos el último dato por 110, la distancia intercuartil no cambiaría, mientras que el desvío pasaría de 3.2 a 24.13!!!!

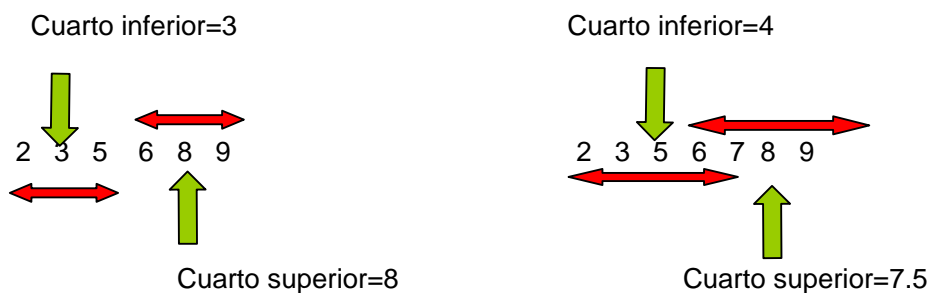
Cuartos y Distancia entre Cuartos: Medidas muy cercanas a los cuartiles inferior y superior son el cuarto inferior y el cuarto superior. Se calculan de la siguiente manera:

- Se ordena la muestra y se calcula la mediana de los datos.
- Dividimos a la muestra ordenada en dos partes: la **primera** corresponde a los datos más pequeños que la mediana y la **segunda** parte a los datos más grandes que la mediana
- Si el tamaño de la muestra es *par*, el **cuarto inferior** es la mediana de la primera mitad, mientras que el **cuarto superior** es la mediana de la segunda mitad.
- Si el tamaño de la muestra es *impar*, a la primera y a la segunda parte se las expande agregándose a cada una de ellas la mediana de todos los datos. El **cuarto inferior** es la mediana de la primera parte expandida y el **cuarto superior** es la mediana de la segunda parte expandida. Es decir, en el caso impar, la mediana interviene en el cómputo de los dos cuartos.

Definimos la distancia entre cuartos como:

$$d_c = \text{distancia entre cuartos} = \text{cuarto superior} - \text{cuarto inferior}.$$

Ejemplo: Sean las siguientes muestras ordenadas



Desvío Absoluto Mediano (Desviación absoluta respecto de la Mediana): Es una versión robusta del desvío estándar basada en la mediana. Definimos la MAD como:

$$MAD = \text{mediana}(|x_i - \tilde{x}|)$$

¿Cómo calculamos la MAD?

- Ordenamos los datos de menor a mayor.
- Calculamos la mediana.
- Calculamos la distancia de cada dato a la mediana.
- Despreciamos el signo de las distancias y las ordenamos de menor a mayor.
- Buscamos la mediana de las distancias sin signo.

Observación: Si deseamos comparar la distancia intercuartil y la MAD con el desvío standard es conveniente dividir las por constantes adecuadas. En ese caso se compara a S con

$$\frac{MAD}{0.675} \quad \text{ó} \quad \frac{d_i}{1.35}$$

Números de Resumen: Los 5 números de resumen de la distribución de un conjunto de datos consisten en el **mínimo**, el **cuartil inferior**, la **mediana**, el **cuartil superior** y el **máximo**.

Ejemplo: Los siguientes datos corresponden a tiempos de CPU (en segundos) de 25 trabajos enviados a un server y seleccionados al azar.

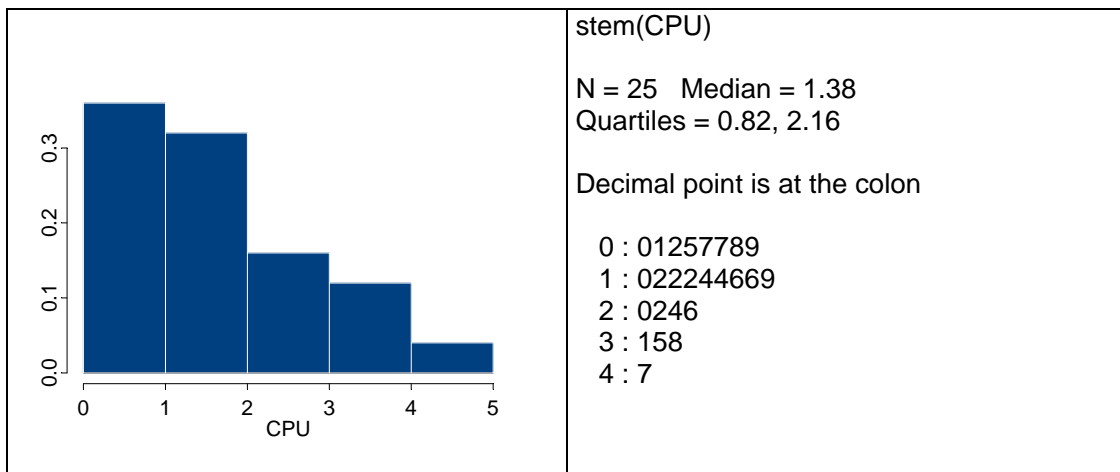
CPU				
1.17	1.23	0.15	0.19	0.92
1.61	3.76	2.41	0.82	0.75
1.16	1.94	0.71	0.47	2.59
1.38	0.96	0.02	2.16	3.07
3.53	4.75	1.59	2.01	1.40

Calculamos los 5 números resumen y la media muestral para este conjunto de datos, utilizando el software R.

```
> summary(server1)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.02 0.82 1.38 1.63 2.16 4.75
```

Realizamos un esquema de Tallo y Hoja y graficamos un histograma para este conjunto de datos:



Todas las medidas y los gráficos muestran que se trata de una distribución asimétrica con cola a derecha.

Box-Plots

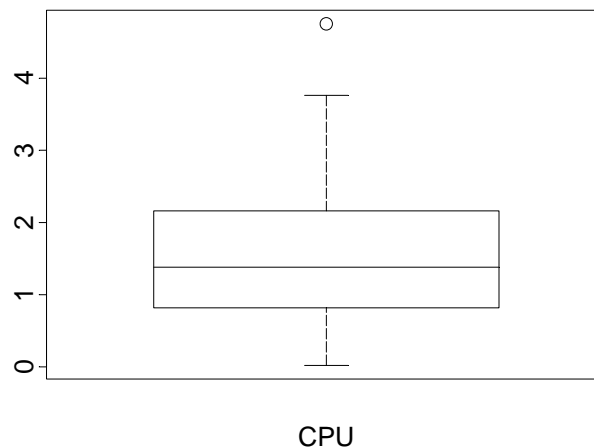
Con las medidas anteriores podemos construir un gráfico de fácil realización y lectura.

¿Cómo lo hacemos? Vamos a dar una versión, pero vale la pena advertir que hay variaciones de un programa a otro.

1. Representamos una escala vertical u horizontal
2. Dibujamos una caja cuyos extremos son los cuartiles y dentro de ella un segmento que corresponde a la mediana.
3. A partir de cada extremo dibujamos un segmento hasta el dato más alejado que está a lo sumo $1.5 d_I$ del extremo de la caja. Estos segmentos se llaman bigotes.
4. Marcamos con * a aquellos datos que están entre $1.5 d_I$ y $3 d_I$ de cada extremo y con \circ a aquellos que están a más de $3 d_I$ de cada extremo. Algunos paquetes, como el R, indican a todos los outliers de la misma forma.

Observación: Muchos paquetes estadísticos realizan el boxplot usando los cuartos y la distancia entre cuartos en lugar de la distancia intercuartil. Como estas medidas son muy próximas, en general los resultados son análogos. Lo importante es que entre los cuartos o entre los cuartiles yace aproximadamente el 50% central de los datos.

Ejemplo: El box-plot correspondiente a los tiempos de CPU es el siguiente



Es interesante observar que en el boxplot se indica a uno de los datos como outlier, mientras que en el análisis anterior esto no parecía evidente.

A partir de un box-plot podemos apreciar los siguientes aspectos de la distribución de un conjunto de datos:

- posición
- dispersión
- asimetría
- longitud de las colas
- puntos anómalos o outliers.

Los box-plots son especialmente útiles para comparar varios conjuntos de datos, pues nos dan una rápida impresión visual de sus características.

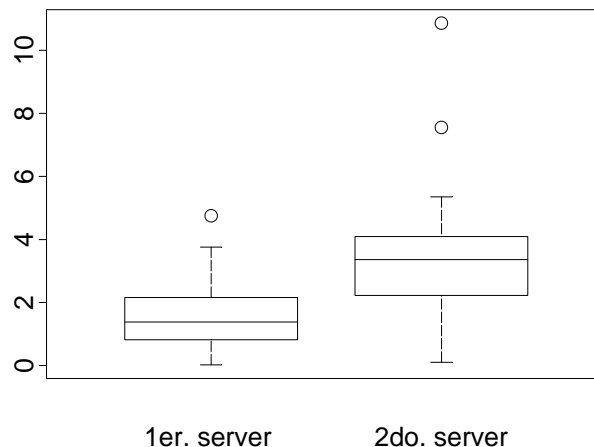
Outliers: Los métodos que hemos visto nos permiten identificar puntos atípicos que pueden aparecer en una o más variables. Su detección es importante pues pueden determinar o influenciar fuertemente los resultados de un análisis estadístico clásico, dado que muchas de las técnicas habitualmente usadas son muy sensibles a la presencia de datos atípicos.

Los outliers deben ser cuidadosamente inspeccionados. Si no hay evidencia de error y su valor es posible no deberían ser eliminados. Asimismo, la presencia de outliers puede indicar que la escala elegida no es la más adecuada.

Boxplots Paralelos

Una aplicación muy útil de los boxplots es la comparación de la distribución de dos o más conjuntos de datos graficando en una escala común los boxplots de cada una de las muestras. En este sentido los boxplots se muestran como un método muy efectivo de presentar y resumir los datos, tal como veremos en el siguiente ejemplo.

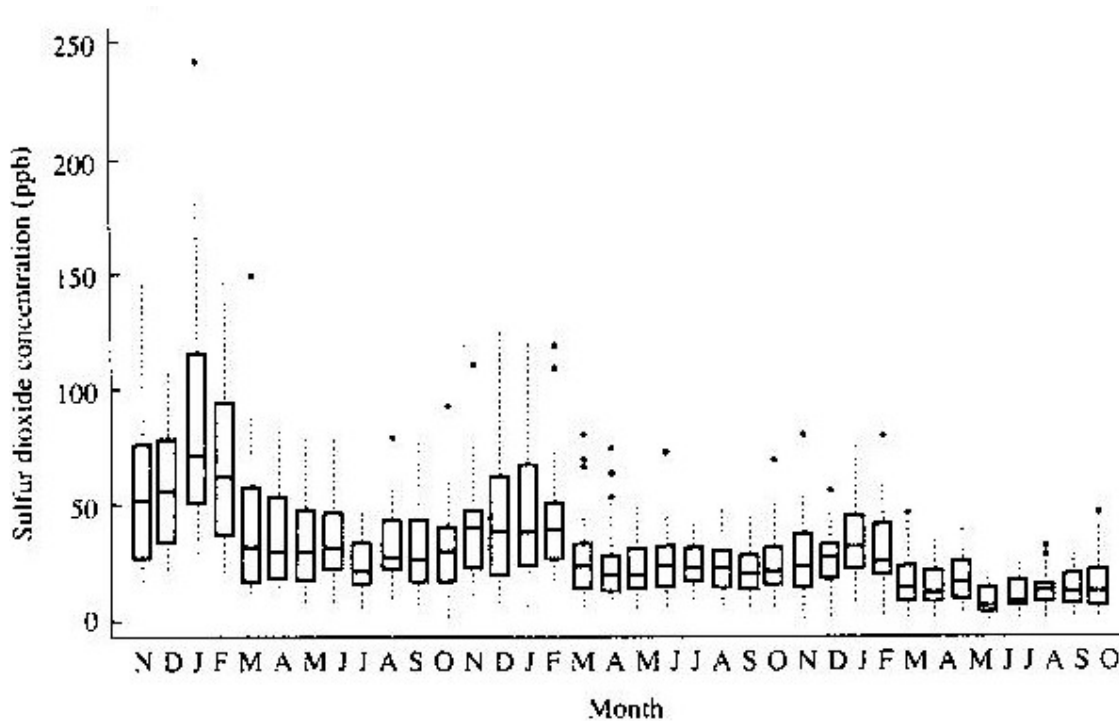
Ejemplo: Supongamos que se dispone de otros 25 datos correspondientes a tiempos de CPU enviados a otro server. Si realizamos boxplots paralelos para ambos conjuntos de datos obtenemos el siguiente gráfico. La simple comparación de los boxplots obtenidos revela que los trabajos enviados al segundo server son más largos. De hecho, el 75% de los trabajos muestreados en el segundo server tienen tiempos de CPU mayores que el cuartil superior de los trabajos muestreados en el primer server.



Ejemplo: Los siguientes boxplots corresponden a datos de concentración máxima diaria, en partes por mil millones de dióxido de azufre en Bayonne, en el estado de Nueva Jersey, desde noviembre de 1969 hasta octubre de 1972 agrupados por meses. Hay 36 grupos de datos, cada uno de tamaño aproximadamente 30.

Los boxplots muestran algunas características de estos datos en forma muy rápida.

Hay una reducción general de la concentración de dióxido de azufre a lo largo del tiempo debida a la conversión gradual en la zona al uso de combustibles con baja concentración de azufre. Esta disminución es más fuerte para los cuartiles superiores. También se muestran concentraciones más elevadas para los meses de invierno debido al uso de calderas a petróleo. Claramente se ve un efecto cíclico y amortiguado. Los boxplots muestran una distribución asimétrica a derecha, con presencia de outliers en algunos meses, y que la dispersión de la distribución es mayor cuando el nivel general de la concentración es más alto.



QQ-plot (Normal Probability Plot): El QQ-plot es un gráfico que nos sirve para evaluar la cercanía a una distribución dada, en particular a la distribución normal.

Consideremos la muestra aleatoria: X_1, X_2, \dots, X_n y los correspondientes estadísticos de orden

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

Observemos que $X_{(1)} = \min(X_1, X_2, \dots, X_n)$, mientras que $X_{(n)} = \max(X_1, X_2, \dots, X_n)$.

En particular, si U_1, U_2, \dots, U_n son v.a. i.i.d tales que $U_i \sim U(0,1)$, se puede demostrar que

$$E(U_{(i)}) = \frac{i}{n+1}.$$

Por lo tanto esperamos que, si la distribución subyacente fuese Uniforme y graficásemos $U_{(1)}, \dots, U_{(n)}$ vs sus valores esperados $\frac{1}{n+1}, \dots, \frac{n}{n+1}$, el gráfico debería parecerse a una recta.

Por otro lado, sabemos que si X es una variable continua con función de distribución F estrictamente creciente, entonces

$$Y = F(X) \sim U(0,1)$$

Esto sugiere que si suponemos que $X_i \sim F$, entonces podemos graficar

$$F(X_{(i)}) \quad \text{vs} \quad \frac{i}{n+1}$$

o equivalentemente

$$X_{(i)} \quad \text{vs} \quad F^{-1}\left(\frac{i}{n+1}\right).$$

Observemos que si F es de la forma

$$F(x) = G\left(\frac{x - \mu}{\sigma}\right),$$

o sea, si depende de un parámetro de posición y uno de escala, como es el caso de la normal, podemos graficar

$$\frac{X_{(i)} - \mu}{\sigma} \quad \text{vs} \quad G^{-1}\left(\frac{i}{n+1}\right)$$

o bien

$$X_{(i)} \quad \text{vs} \quad G^{-1}\left(\frac{i}{n+1}\right)$$

Como,

$$X_{(i)} \cong \sigma \cdot G^{-1}\left(\frac{i}{n+1}\right) + \mu$$

el gráfico será aproximadamente una recta.

Notemos que si F^{-1} es la inversa de F , entonces el p -ésimo percentil de F , x_p , es tal que

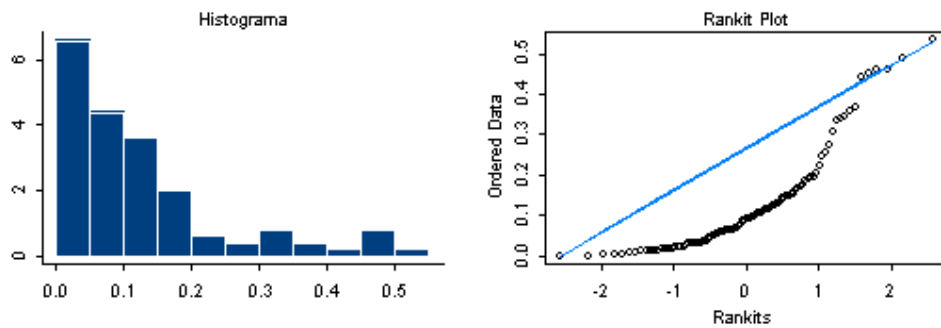
$$F(x_p) = p \Rightarrow x_p = F^{-1}(p)$$

por lo tanto, $F^{-1}\left(\frac{i}{n+1}\right)$ es el $\frac{i}{n+1}$ -percentil de F .

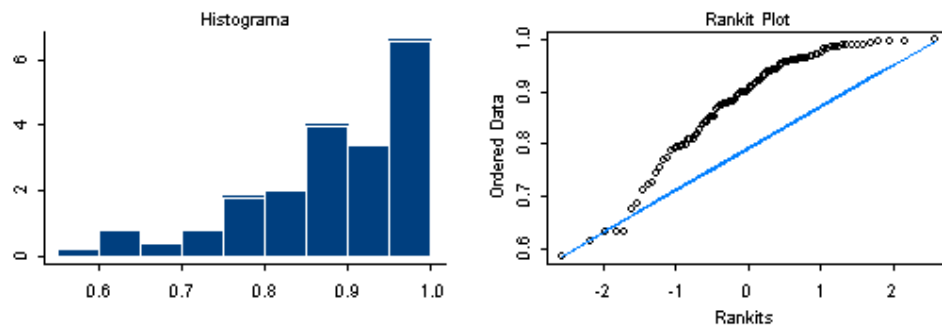
En el QQ-plot se grafican en el eje de abscisas los percentiles de la distribución teórica (en nuestro caso normal) y en el eje de ordenadas las observaciones ordenadas, que pueden ser vistas como percentiles empíricos.

En los siguientes gráficos ilustramos el uso de estas técnicas gráficas con algunos ejemplos. Cabe observar que algunos paquetes estadísticos representan a los percentiles teóricos de la distribución normal en el eje de abscisas y otros en el eje de ordenadas

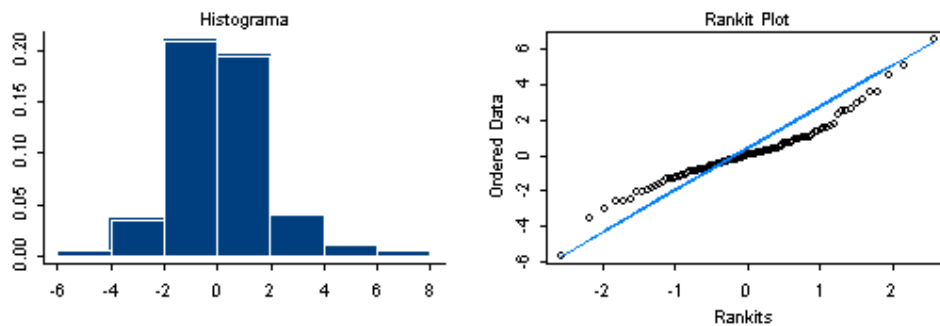
Asimétrica a Derecha



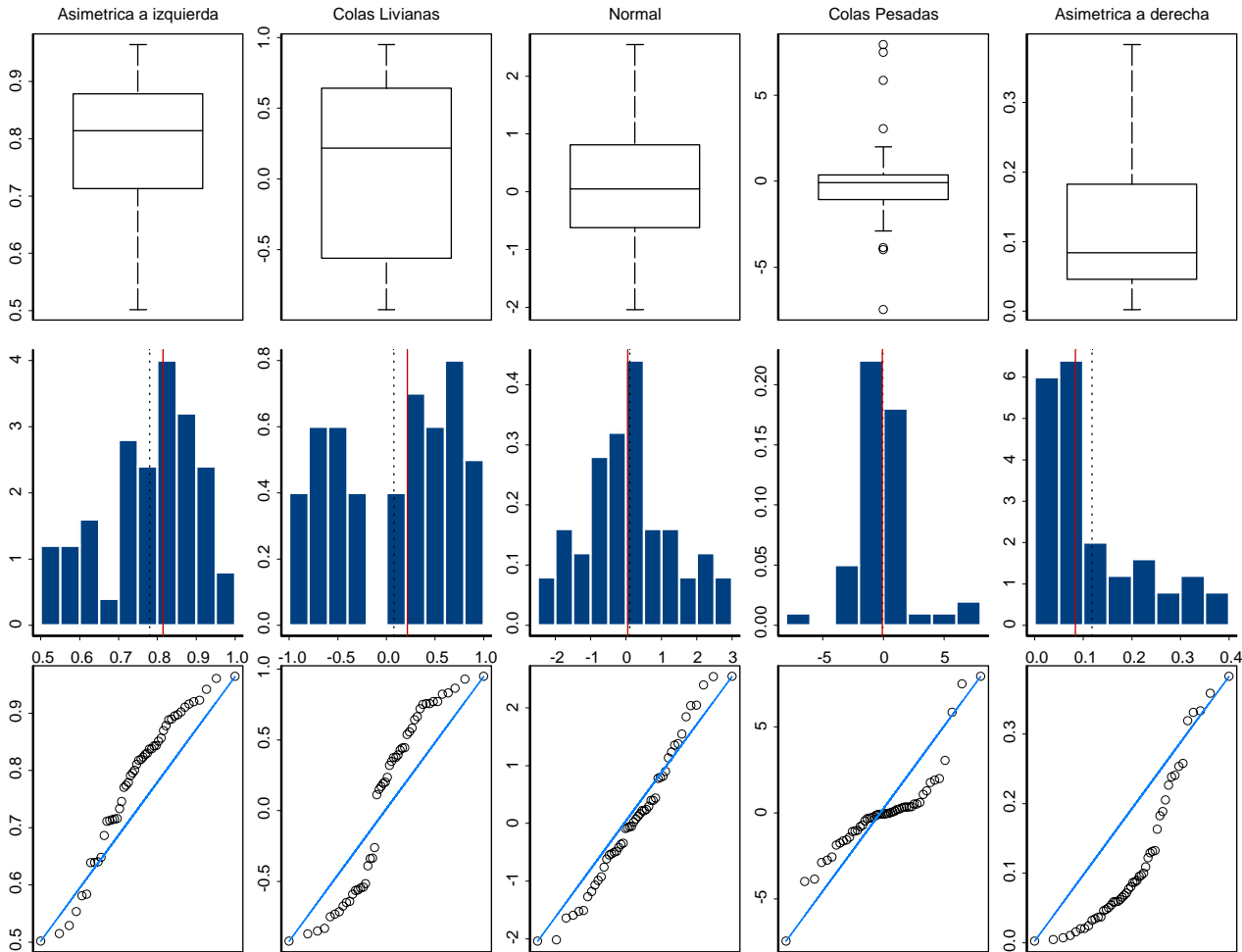
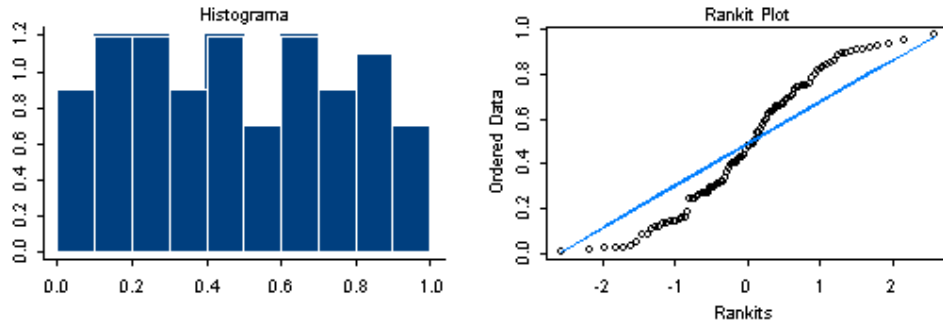
Asimétrica a Izquierda



Simétrica con Colas Pesadas



Simétrica con Colas Livianas



Rojo=Mediana, Negro=Media