Introducción a la Estadística Computacional 2024

Juan Piccini

LPE/IMERL

Descripción de Datos

Juan Piccini

LPE/IMERL

Indice

Introducción

- La estadística actúa como disciplina puente entre los modelos matemáticos y el mundo real.
- Un modelo matemático es una abstracción simplificada de una realidad más compleja, y siempre existirá cierta discrepancia entre lo observado y lo previsto por el modelo.
- Una buena descripción de esto es la frase **"all models are wrong, but some are useful"**.
- La estadística proporciona los medios para evaluar y juzgar estas discrepancias.

Introducción

- Los modelos estadísticos pueden clasificarse en función de la información que utilizan y el objetivo que persiguen.
- Cuando la información utilizada es de una única variable, hablaremos de **modelos univariados**, mientras que si participan varias variables hablaremos de **modelos mulitvariados**.
- Si el objetivo es investigar las variables en un instante temporal dado, hablaremos de **modelos de corte transversal o modelos estáticos** (p.ej. la relación entre el salario y el nivel educativo en las familias uruguayas en el año 2019).
- Cuando se desea estudiar una evolución temporal hablaremos de **modelos longitudinales o modelos dinámicos**.

Introducción

- Supondremos que el orden en que se recogen los datos es irrelevante.
- Cuando los datos se observan con una pauta temporal fija (cada mes, año, etc.), constituyen una **Serie de Tiempo**, cuyo análisis requiere métodos especiales que tengan en cuenta que el orden de los datos es informativo.
- Nos referiremos a datos donde el orden temporal no juega papel alguno.

Tipos de variables

- Dado un conjunto de datos (mediciones) de una variable X, buscamos sintetizar la información contenida en dicho conjunto.
- Las variables pueden subdividirse en dos tipos:
  1. **Variables Cualitativas** (Categóricas o Atributos): No toman valores numéricos y describen cualidades. Por ejemplo, color de cabello, sexo, etc.
  2. **Variables Cuantitativas**: Toman valores numéricos. Pueden subdividirse a su vez en dos tipos:
     - **Variables Cuantitativas Discretas**: Toman valores en un conjunto discreto, el cual puede ser finito o infinito numerable. Por ejemplo, número de veces que ocurre un evento (número de llamadas en un cierto lapso de tiempo, número de hijos, etc.)
     - **Variables Cuantitativas Continuas**: Toman valores en un intervalo. Por ejemplo, tiempo entre llegadas de dos vehículos.

- El primer paso es descriptivo: consiste en buscar formas de visualizar los datos y cómo se distribuyen los mismos.
- La información recogida en la etapa descriptiva puede guiar en la búsqueda de modelos que expliquen los datos.
- Tablas, diagramas de tallo y hojas, histogramas, tortas y boxplots son formas de visualizar y sintetizar la información de un conjunto de datos.
- Todas ellas de un modo u otro se basan en el concepto de **frecuencia**.

Distribuciones de Frecuencias

- La presentación de un conjunto de datos suele hacerse indicando los valores de la variable y sus **frecuencias** de aparición, tanto **absolutas (cantidad de veces que aparece cada valor)** como **relativas**.
- La **frecuencia relativa** de un suceso A viene dada por

$$fr(A) = \frac{\text{Cantidad de veces que se observa } A}{\text{Cantidad total de datos}} =$$

$$= \frac{\text{Frecuencia Absoluta de A}}{\text{Cantidad total de datos}}$$

- Por lo general tanto para calcular frecuencias como para su presentación, es de mucha utilidad agrupar los datos.

Agrupamiento

- Cuando tenemos una variable continua o una variable discreta que toma muchos valores, suelen agruparse los datos en clases, como sigue:
- Redondeamos los datos a dos o a lo sumo tres cifras significativas.
- Decidimos el número de clases c a considerar. Habitualmente $5 \leq c \leq 20$, aunque una regla muy utilizada es $c = \sqrt{n}$, donde n es la cantidad de datos (c debe ser entero, redondear $\sqrt{n}$ si hace falta).
- Seleccionamos los límites de clase que definen los intervalos, de modo que las clases sean de la misma longitud y cada dato caiga en una sola clase.
- Contamos la cantidad de datos en cada clase (frecuencia de clase) y dividimos entre el total de datos para tener la frecuencia relativa de cada clase.

Ejemplo

Tenemos 120 mediciones del tiempo que insume realizar una tarea. Los mismos varían entre 20'4" y 43'48".
Redondeamos a minutos y definimos 5 intervalos de igual longitud. La siguiente tabla muestra el resultado.

Table: Tiempos de ejecución

Intervalo	Centro	Frec. Absoluta	Frec. Relativa
20-24	22	36	0.30
25-29	27	48	0.40
30-34	32	24	0.20
35-39	37	6	0.05
40-44	42	6	0.05
Total		120	1

Ejemplo

La siguiente tabla muestra el número X de llamadas que llega a una centralita en períodos de un minuto (monitoreamos 90 minutos).

Table: Llamadas por minuto

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

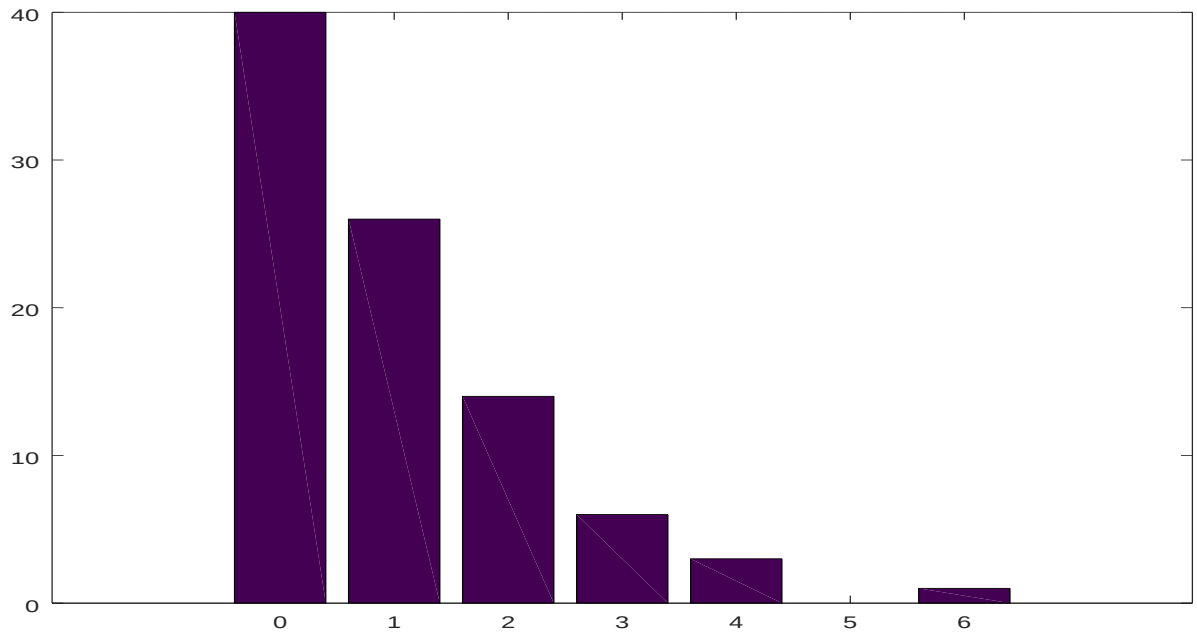
En 40 ocasiones tuvimos un minuto sin llamadas, en 26 ocasiones un minuto con solo una llamada, etc.

Observación

- Las tablas suelen ser la primera forma de organizar y visualizar datos, pero todavía no son suficientemente descriptivas.
- Podemos mostrar la misma información de la tabla 2 mediante un diagrama de barras o barplot.
- Para ello (Octave) creamos un vector con las frecuencias absolutas del número de llamadas: data=[40,26,14,6,3,0,1], obteniendo data = 40 26 14 6 3 0 1
- Luego indicamos el lugar donde irá cada valor: c=[0,1,2,3,4,5,6], obtenemos c =
  0 1 2 3 4 5 6
- Luego graficamos mediante: bar(c,data)

Ejemplo

Figure: Cantidad de llamadas en un minuto



Ejemplo: Control de calidad

- La siguiente tabla representa un ejemplo para una variable cualitativa. Se indican las clases o atributos y sus frecuencias observadas (ordenadas de mayor a menor).

Table: Distribución de defectos en libros en una imprenta

Clases	Frecuencia	Frecuencia Relativa
Corte de las hojas	60	0.43
Mala Impresión	40	0.29
Tinta Irregular	20	0.14
Encuadernación	12	0.09
Portada	6	0.04
Lomo	2	0.01
Total	140	1

Diagrama de Pareto

- Si representamos cada clase por un rectángulo cuya altura es su frecuencia relativa, obtenemos el *Diagrama de Pareto* que es muy útil para datos cualitativos. Para ello hacemos:
- datos=[0.43,0.29,0.14,0.09,0.04,0.01]
- nombres= {'Corte',Impresion','Tinta','Encuadernacion','Portada','Lomo'}
- bar(datos)
- set(gca,'xticklabel',nombres)

Diagrama de Pareto

Figure: Diagrama de Pareto de los defectos en libros en una imprenta

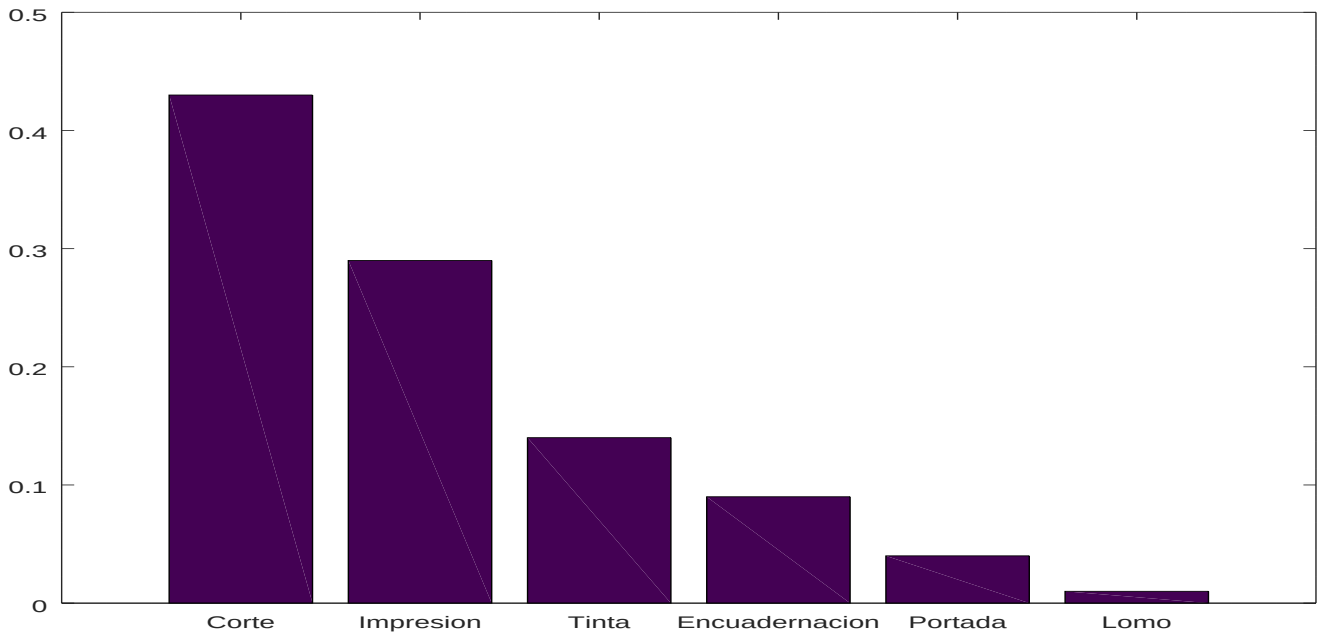


Diagrama de Pareto

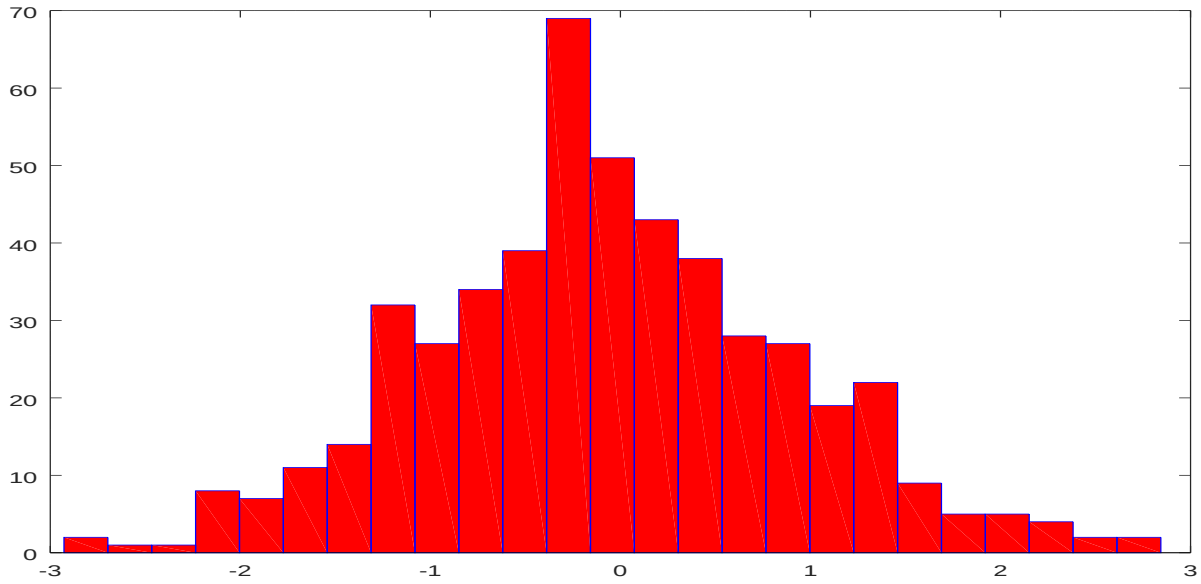
- El área de cada rectángulo representa la frecuencia relativa con la que aparece cada defecto (por tanto el área total es $=1$).
- Al ordenar por frecuencias decrecientes, los defectos mayores aparecen primero, lo que nos da una idea de su importancia.
- Se observa que el 86% de los defectos corresponden a Corte de Hojas, Mala Impresión y Tinta irregular, lo que sugiere cuales defectos hay que atacar primero.
- Esto convierte a los diagramas de Pareto en una poderosa herramienta en problemas de control de calidad y distribución de recursos.

Histogramas

- Es una de las representaciones más utilizadas.
- Es un conjunto de rectángulos, cada uno de los cuales representa una clase o un intervalo en el que se agrupan los datos.
- La altura de los rectángulos se determina de modo tal que el área de cada rectángulo sea proporcional a la frecuencia de cada clase.
- Para el caso de variables continuas, se divide el intervalo definido por los valores mínimo y máximo en subintervalos de igual longitud.

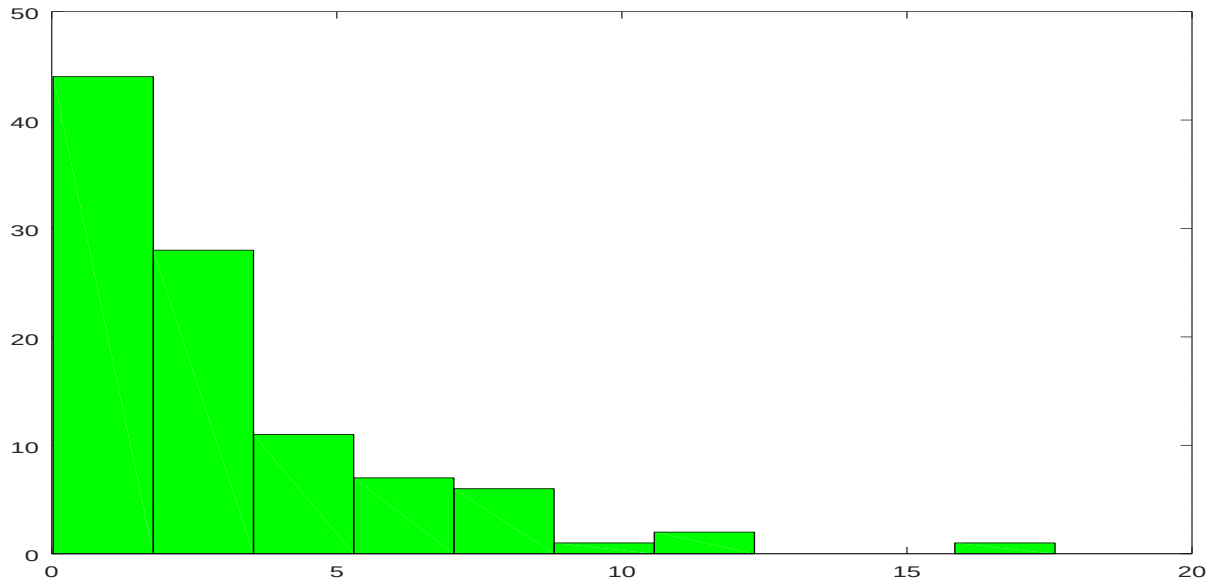
Histogramas: Ejemplos

- Generamos 500 datos con distribución normal estándar y ploteamos el histograma:
- hist (randn (1, 500), 25, "facecolor", "r", "edgecolor", "b")



Histogramas: Ejemplos

- Generamos 100 datos con distribución exponencial de parámetro $\lambda = 3$: x=exprnd(3,100,1)
- Hacemos el histograma: hist(x,"facecolor","g")



Diagramas de Tallo y Hojas

- Un diagrama de tallo-hoja (Tukey, 1977) puede verse como un histograma que conserva información numérica. De manera similar al histograma permite ver el lote como un todo y advertir aspectos como:
- Cuán aproximadamente simétricos son los datos.
- Cuán dispersos están los valores.
- La aparición de valores inesperadamente más frecuentes.
- Si algunos valores están alejados del resto.
- Si hay concentraciones de valores.
- Si hay grupos separados.

Diagramas de Tallo y Hojas

- Primero redondeamos los datos a dos o tres cifras significativas, expresándolas en unidades convenientes.
- Luego los disponemos en una tabla con dos columnas separadas por una línea vertical como sigue:
- Para datos de dos dígitos escribimos a la izquierda de la línea los dígitos de las decenas (el tallo), y a la derecha las unidades (las hojas). Por ejemplo 87 se escribe $8|7$.
- Para datos de tres dígitos el tallo estará formado por los dígitos de las centenas y decenas, mientras que las unidades serán las hojas. Por ejemplo 127 será $12|7$
- Cada tallo define una clase. El número de hojas representa la frecuencia de dicha clase.

Diagramas de Tallo y Hojas: Ejemplo

- Datos (en cm.):

11.357	12.542	11.384	12.431
14.212	15.213	13.300	11.300
17.206	12.710	13.455	16.143
12.162	12.721	13.420	14.698

Datos redondeados (en mm.) y ordenados:

113	114	114	122	124	125	127	127
133	134	135	142	147	152	161	172

x=[113,114,114,122,124,125,127,127,133,134,135,142,147,152,161,172]

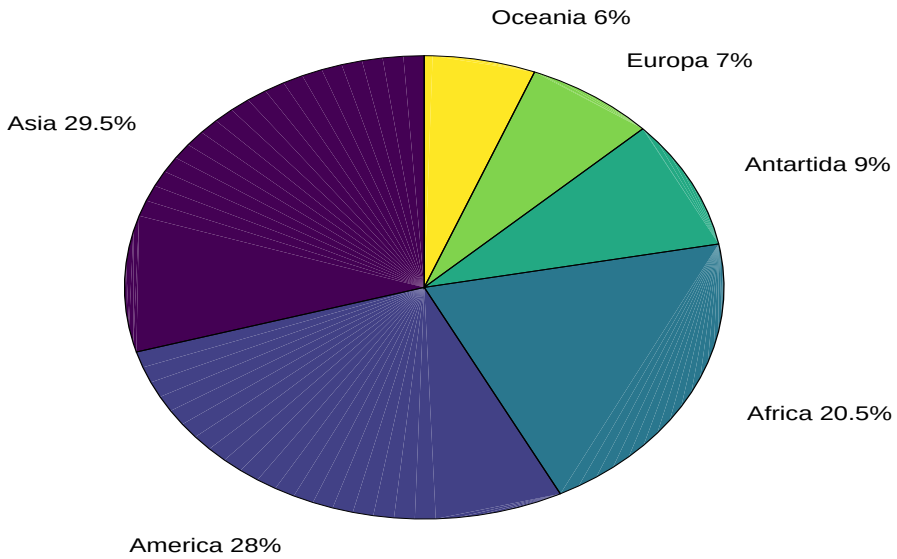
- Diagrama de tallo y hojas: stemleaf (sort(x), "Tallo y hojas")

11|344
12|24577
13|345
14|27
15|2
16|1

Diagramas de torta

- El objetivo de un gráfico es describir de manera simple y fiel la información contenida en los datos. Por tanto la naturaleza de los datos puede sugerir una representación gráfica específica, distinta de las anteriores.
- Por ejemplo, para reflejar la idea de división de un conjunto de datos en clases excluyentes podemos utilizar como alternativa los diagramas de torta. La torta es tal que el área de cada porción es proporcional a la frecuencia relativa.
- A modo de ejemplo usemos la proporción de superficie ocupada por los distintos continentes.
- x=[29.5,28,20.5,9,7,6]; labels={'Asia 29.5%','America 28%','Africa 20.5%','Antartida 9%','Europa 7%','Oceanía 6%'}
- pie(x,labels), obtenemos el siguiente diagrama:

Diagramas de torta: Ejemplo



Medidas-resumen

- Cuando disponemos de un conjunto de datos homogéneos de una variable cuantitativa, resulta conveniente complementar la información de la distribución de frecuencias con ciertas medidas-resumen.
- Las más importantes son las de **tendencia central** o **centralización**, que indican el valor medio de los datos, el "centro "de la nube de puntos.
- Y las medidas de **dispersión**, que miden la variabilidad de los datos, su dispersión respecto del centro antes mencionado.
- También hay medidas que dan información sobre la forma de la distribución, como su grado de simetría o de concentración.
- Estas medidas-resumen son informativas cuando los datos son homogéneos, pero pueden ser engañosas cuando los datos provienen de una mezcla de distintas poblaciones.

Medidas de Centralización

- **Media:** Dado un conjunto de datos numéricos $\{x_1, x_2, \ldots, x_n\}$, se define la **media aritmética** o **promedio**

$$\overline{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Otra forma equivalente es agrupando primero aquellos valores iguales: sea n_i a la cantidad de veces que aparece el valor x_i, entonces

$$\overline{X} = \frac{n_1 x_1 + \cdots + n_k x_k}{n} = \frac{n1}{n}x_1 + \cdots + \frac{n_k}{n}x_k = \sum_i x_i fr(x_i) \qquad (1)$$

- Para datos discretos agrupados (por ejemplo cantidad de llamadas que llegan a una centralita), si llamamos x_j a los distintos valores que la variable X toma y $fr(x_j)$ a la frecuencia relativa de dicho valor, la media se define como

$$\overline{X} = \sum_j x_j fr(x_j)$$

.

Medidas de Centralización

- Para datos agrupados en clases, suponemos que todos los datos de cada clase son idénticos al centro de la clase. Sea m_j al centro de la clase j y $fr(m_j)$ la frecuencia relativa de la clase j, tendremos

$$\overline{x} = \sum_j m_j fr(m_j)$$

P.ej. agrupamos el tiempo (en minutos) para realizar una cierta tarea,

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

la media nos da

$$\overline{x} = 0.30 \times 22 + 0.40 \times 27 + 0.20 \times 32 + 0.07 \times 37 + 0.03 \times 42 = 27.65$$

Propiedades de la media

- La media aritmética minimiza la suma de los desvíos: si sumamos las diferencias por exceso y por defecto de los datos respecto a la media, dicha suma es cero.
- En efecto, $\sum_{i=1}^{n}(x_i - \overline{x}) = \left(\sum_{i=1}^{n} x_i\right) - n\overline{x} = 0$. La media puede verse como el "centro de masa" de la nube de datos.
- La media también minimiza la suma de los cuadrados de los desvíos: si deseamos hallar el valor a que minimiza $\sum_{i=1}^{n}(x_i - a)^2$, derivando respecto de a obtenemos $2\sum_{i=1}^{n}(x_i - a) = 0$, de donde $a = \overline{x}$.

La Mediana

- La **mediana** x_m es el valor que está en el medio de los datos ordenados. Esto es, si ordenamos los datos de menor a mayor, el 50% de los datos es $\leq x_m$ y el otro 50% es mayor.
- Si tenemos un número impar de datos, la mediana es el dato central en la muestra ordenada, sino se define como el primer dato donde se alcanza o supera el 50% de los datos. En ocasiones se define como el promedio de los dos datos centrales (cuando n es par).
- Para datos agrupados discretos x_m es el menor valor tal que $fr(x < x_m) < 0.5$ y $fr(x \leq x_m) \geq 0.5$.
- Es decir, si ordenamos los datos, antes de x_m tenemos menos de la mitad de los datos, y al incluír a x_m tenemos al menos la mitad de los datos.
- Para datos continuos agrupados en intervalos se toma como x_m al centro del "intervalo central" (x_a, x_b), que es aquel que verifica $fr(x \leq x_a) < 0.5$ y $fr(x \leq x_b) > 0.5$.

Media y Mediana

- Cuando los datos son homogéneos (parecidos entre sí), tanto la media como la mediana darán valores similares.
- La media es muy sensible a datos atípicos. Un solo dato alejado del resto de la nube de puntos afectará sensiblemente a la media.
- Por ejemplo, supongamos datos D=3.2, 4.6, 3.6, 2.7, 4.1, 5.6, 5.0, 3.1, 3.9, 3.9. La media nos da $\overline{x}_D$ =3.97. Si agregamos un nuevo dato, por ejemplo 50,y llamamos $D' = D \cup \{50\}$, tendremos que $\overline{x}_{D'}$ =8.15 (se corrió más del 100%).
- La mediana utiliza menos información que la media, solamente le importa el *orden* de los datos, no su magnitud.
- Por ello la mediana es mucho menos sensible que la media a la presencia de datos atípicos. Para los datos del ejemplo anterior, la mediana es 3.9 en ambos casos.

Media y Mediana

- Es recomendable calcular media y mediana: si ambas son similares, esto es señal de que los datos son homogéneos, la distribución es simétrica; si difieren mucho entonces la distribución será muy asimétrica, lo que sugiere heterogeneidad en los datos.
- A modo de ejemplo, supongamos que se calcula el tiempo medio que un estudiante requiere para completar una carrera universitaria en dos universidades U_1 y U_2, obteniendo 5.5 años para ambas. Podemos concluír que son igualmente difíciles?
- Supongamos que la universidad U_1 es muy homogénea, solamente con títulos de 5 años y de dificultad similar, que los alumnos completan promedialmente en 5.5 años.
- U_2 es más heterogénea, con carreras de 4 años que requieren un promedio de 5 años, y carreras de 6 años que promedialmente insumen 7.5 años.

Media y Mediana

- Supongamos que el 80% de los estudiantes de U_2 cursan carreras de 4 años y el otro 20% cursan las de 6 años. Entonces la duración media será $0.8 \times 5 + 0.2 \times 7.5 = 5.5$
- En U_1 los estudiantes demoran medio año más de lo previsto, mientras que en U_2 demoran entre un año y un año y medio más de lo previsto (para carreras de 5 y 6 años respectivamente), por lo que U_2 es más difícil.
- El problema es que estamos comparando una población homogénea con otra que no lo es tanto.

Varianza y Desviación Típica

- A cada medida de centralización podemos asociarle una medida de la variabilidad o dispersión de los datos respecto del centro.
- Para el caso de la media, lo que hacemos es calcular el promedio de los cuadrados de los desvíos, obteniendo la **Varianza**

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

- el factor $\frac{1}{n}$ es el peso de cada dato (todos pesan lo mismo). Para obtener una medida que tenga las mismas unidades que los datos, se suele usar el **Desvío Estándar** o **Desviación Típica**

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

Varianza y Desviación Típica

- Para datos agrupados,

$$s = \sqrt{\sum_j (x_j - \overline{x})^2 fr(x_j)}$$

  La suma es sobre la cantidad de clases o de valores distintos que toma la variable. Las frecuencias relativas sustituyen al peso $\frac{1}{n}$, y como no tienen porqué ser iguales, ya no podemos sacarlas de factor común.

- Para datos continuos agrupados en intervalos, sustituímos x_j por el centro del intervalo, m_j, obteniendo

$$s = \sqrt{\sum_j (m_j - \overline{x})^2 fr(m_j)}$$

Interpretación de la Desviación Típica

- La **Desigualdad de Tchebychev** establece que

$$fr(|x_i - \overline{x}| > ks) < \frac{1}{k^2}$$

- De donde

$$fr(|x_i - \overline{x}| \leq ks) \geq 1 - \frac{1}{k^2}$$

- Una consecuencia es que para cualquier distribución, entre la media y dos desvíos típicos (k=2) se encuentran al menos el 75% de los datos.
- Entre la media y tres desvíos típicos (k=3), se encuentran al menos el 89% de los datos.

Coeficiente de Variación

- Se llama **Coeficiente de Variación** al cociente

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

  Donde supondremos $\overline{x} \neq 0$.

- **El inverso del CV** es

$$\frac{\overline{x}}{s}$$

  Conocido como **Coeficiente señal/ruido**.

- Cuando el CV es mayor a 1.5 conviene investigar posibles fuentes de heterogeneidad en los datos (medidas con distintos instrumentos, en personas de distinto sexo, en distintos momentos temporales, etc.).

Rango, Percentiles

- El **Rango o Recorrido** de una variable es la diferencia entre su valor máximo y su valor mínimo.
  $rango = \max\{x_1, \ldots, x_n\} - \min\{x_1, \ldots, x_n\}$
- Llamaremos **Percentil p** al menor valor $\geq$ al p% de los datos. Por ejemplo, la mediana es el percentil 50.
- Llamaremos **cuartiles** a aquellos valores que dividen la distribución en 4 partes iguales.
- El primer cuartil Q_1 es por definición el percentil 25, el segundo cuartil Q_2 es la mediana, el tercer cuartil Q_3 es el percentil 75.
- Los percentiles y cuartiles se utilizan para construir medidas de dispersión basadas en los datos ordenados, como el **rango intercuartílico o distancia intercuartílica**, que se define como $Q_3 - Q_1$.
- A menudo se utiliza como resumen de un conjunto de datos al conjunto $\{\min, Q_1, Q_2, Q_3, \max\}$ (resumen de 5 números).

Boxplot

- **El Diagrama de Caja** o **Boxplot** es una representación semigráfica de una distribución para mostrar sus características principales y señalar los posibles datos atípicos (**outliers**).
- Se ordenan los datos y se obtiene el mínimo, el máximo y los cuartiles Q_1, Q_2, Q_3.
- Se dibuja un rectángulo cuyos extremos son Q_1 y Q_3, indicando la posición de la mediana (Q_2) mediante una línea.
- Se calculan los límites admisibles superior e inferior que servirán para identificar posibles *outliers*. Dichos límites son $LI = Q_1 - 1.5 \times (Q_3 - Q_1)$ y $LS = Q_3 + 1.5 \times (Q_3 - Q_1)$.

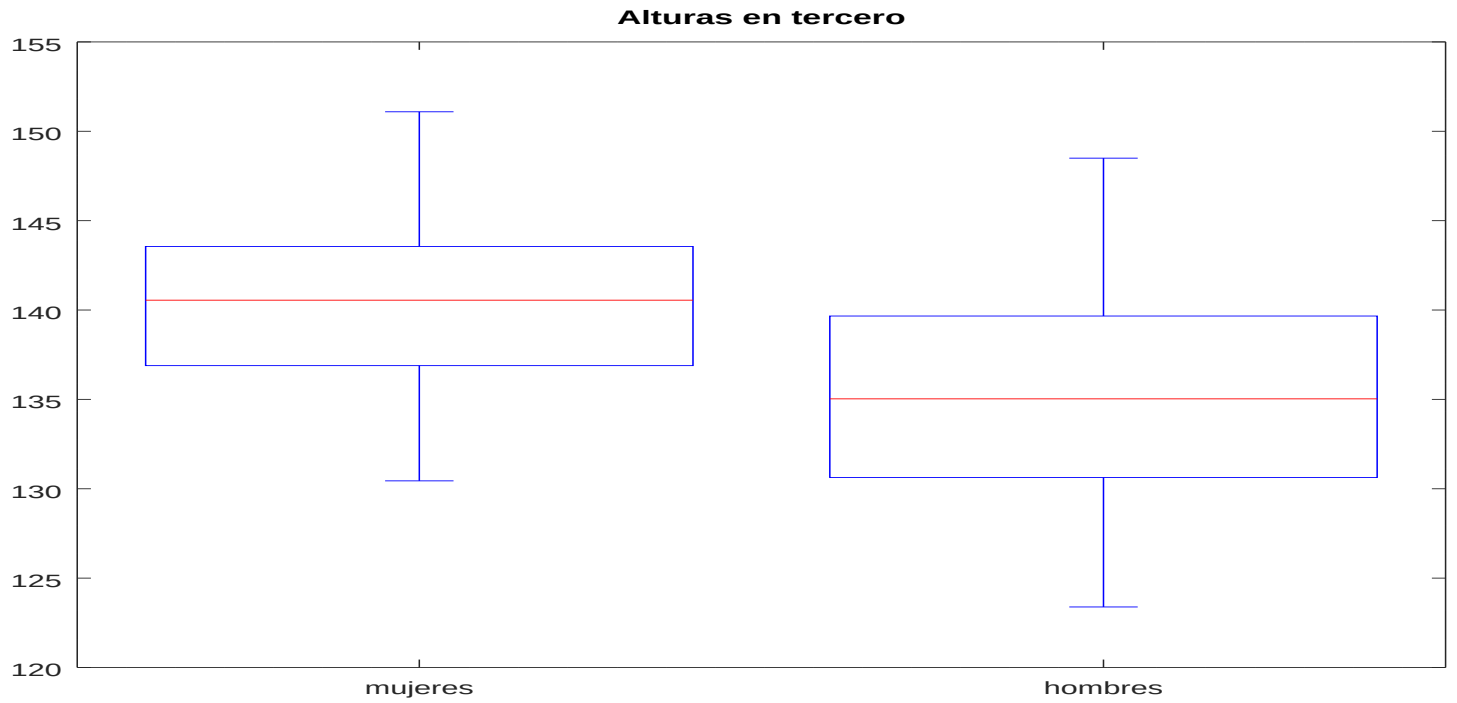
Boxplot

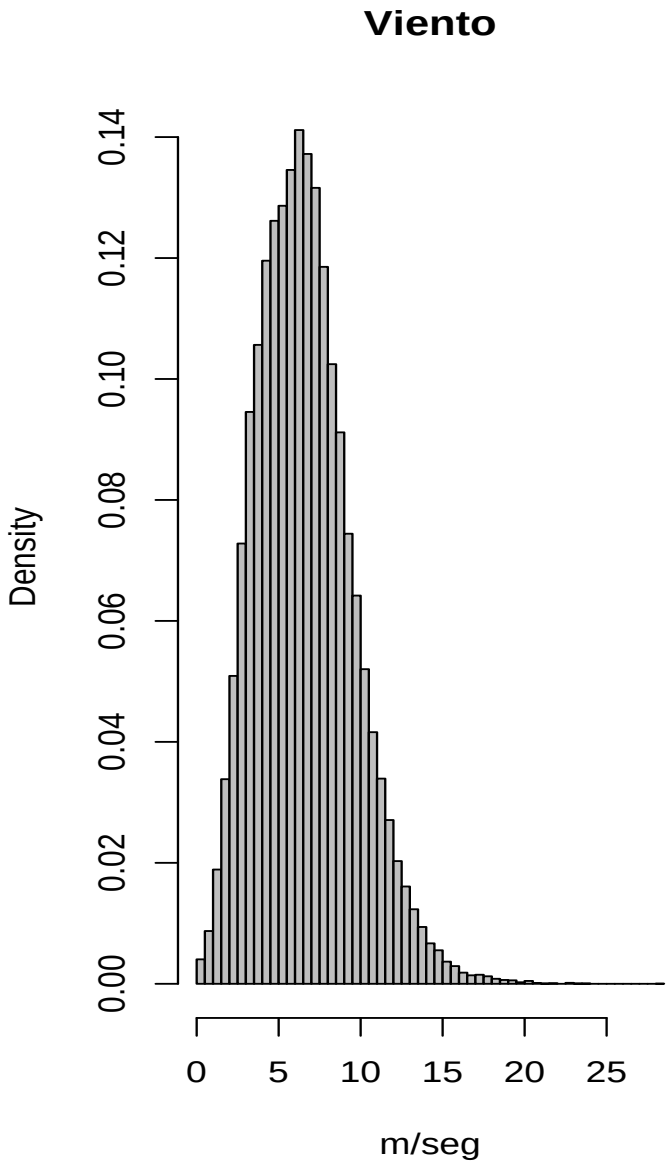
- Todo dato que se aleje en más de 1.5 veces la distancia intercuartílica de los cuartiles 1 y 3 será considerado atípico.
- Dibujar una línea ("bigote") que vaya desde cada extremo del rectángulo central hasta el valor más alejado no atípico, es decir, que está dentro del intervalo (LI, LS).
- Identificar todos los datos fuera de dicho intervalo como atípico u **outlier**.
- Los Boxplot son especialmente útiles para comparar la distribución de una variable en distintas poblaciones.

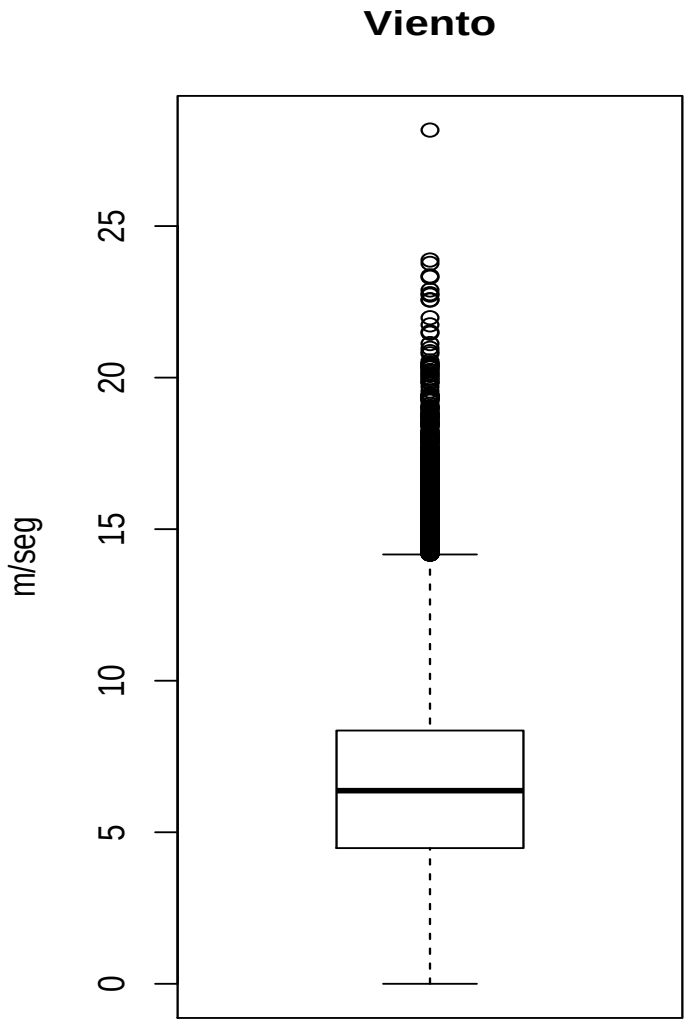
Boxplot: Ejemplos

- Generamos dos conjuntos de 50 datos normales con distinta media y dispersión para simular la estatura de mujeres y hombres en un grupo dado y ploteamos los boxplots de cada grupo. Los comandos son:
- axis([0,3]); Este comando es para posicionar cada boxplot
- boxplot(randn(50,1)*5+140,randn(50,1)*8+135); crea el boxplot y de paso ya creamos cada una de las dos muestras que simulan las estaturas.
- Tenemos 50 datos normales con media 140 y varianza 25, otros 50 datos con media 135 y varianza 64.
- set(gca(),'xtick',[1,2],'xticklabel','mujeres','hombres');
- title('Alturas en tercero');

Boxplot: Ejemplos




Viento
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
5
10
15
20
25
m/seg


Viento
m/seg
0
5
10
15
20
25