Inferencia

Grado en Ingeniería Biomédica

Jorge Calero Sanz (jorge.calero@urjc.es)

20 de febrero de 2023

Introducción

Estadística descriptiva

Introducción

Estadística descriptiva

Podemos hacer una división dentro de la Estadística en dos partes:

Introducción

Estadística descriptiva

Podemos hacer una división dentro de la Estadística en dos partes:

- Estadística descriptiva: recolección, organización y descripción de los datos

Introducción

Estadística descriptiva

Podemos hacer una división dentro de la Estadística en dos partes:

- Estadística descriptiva: recolección, organización y descripción de los datos
- Inferencia Estadística: Obtener resultados generales (población) a través de observaciones aleatorias (inferencia)

Introducción

Estadística descriptiva

Introducción

Estadística descriptiva

- También se puede hacer una distinción entre **Matemática Estadística**

Introducción

Estadística descriptiva

- También se puede hacer una distinción entre **Matemática Estadística** y **Estadística Aplicada**. Dentro de la Estadística Aplicada, se encuentra la Bioestadística, encargada de estudiar métodos y modelos en problemas biológicos o médicos.
- La Estadística Descriptiva es el conjunto de técnicas que obtienen, organizan, presentan y describen un conjunto de datos, valiéndose principalmente de tablas y gráficas.

Introducción

Estadística descriptiva

- También se puede hacer una distinción entre **Matemática Estadística** y **Estadística Aplicada**. Dentro de la Estadística Aplicada, se encuentra la Bioestadística, encargada de estudiar métodos y modelos en problemas biológicos o médicos.
- La Estadística Descriptiva es el conjunto de técnicas que obtienen, organizan, presentan y describen un conjunto de datos, valiéndose principalmente de tablas y gráficas.
- Partimos de una población P de la que extraemos una muestra de datos: $x_1, x_2, ..., x_n$.

Introducción

Estadística descriptiva

Introducción

Estadística descriptiva

- En primer lugar, hay que tener en cuenta cuántas variables de interés tiene nuestra población, y de qué tipo.
- Variables cualitativas:
  - Nominal: color de ojos, barrio donde vive, ...
  - Ordinal: Calificaciones, valoraciones del 1 al 5, ...
- Variables cuantitativas:
  - Discreta: Años, número de hermanos, ...
  - Continua: Altura, presión sanguínea, ...

Frecuencias

Distribución de frecuencias

Frecuencias

Distribución de frecuencias

- La primera información que podemos sacar es lo que conocemos como **frecuencia absoluta**: el número de veces que aparece un determinado valor en la muestra.

Frecuencias

Distribución de frecuencias

- La primera información que podemos sacar es lo que conocemos como **frecuencia absoluta**: el número de veces que aparece un determinado valor en la muestra. Si la muestra tiene tamaño N, la frecuencia absoluta es
$F_i =$ número de veces que aparece el evento i).

Frecuencias

Distribución de frecuencias

- La primera información que podemos sacar es lo que conocemos como **frecuencia absoluta**: el número de veces que aparece un determinado valor en la muestra. Si la muestra tiene tamaño N, la frecuencia absoluta es $F_i =$ número de veces que aparece el evento i). Debe cumplirse entonces $N = \sum_i F_i$.

Frecuencias

Distribución de frecuencias

- La primera información que podemos sacar es lo que conocemos como **frecuencia absoluta**: el número de veces que aparece un determinado valor en la muestra. Si la muestra tiene tamaño N, la frecuencia absoluta es $F_i =$ número de veces que aparece el evento i). Debe cumplirse entonces $N = \sum_i F_i$.
- Significa que esta información depende del tamaño de la muestra, por lo que es más útil...

Frecuencias

Distribución de frecuencias

- La primera información que podemos sacar es lo que conocemos como **frecuencia absoluta**: el número de veces que aparece un determinado valor en la muestra. Si la muestra tiene tamaño N, la frecuencia absoluta es $F_i =$ número de veces que aparece el evento i). Debe cumplirse entonces $N = \sum_i F_i$.
- Significa que esta información depende del tamaño de la muestra, por lo que es más útil...
- La **frecuencia relativa** es la frecuencia absoluta dividida entre el tamaño de la muestra:

Frecuencias

Distribución de frecuencias

- La primera información que podemos sacar es lo que conocemos como **frecuencia absoluta**: el número de veces que aparece un determinado valor en la muestra. Si la muestra tiene tamaño N, la frecuencia absoluta es $F_i =$ número de veces que aparece el evento i). Debe cumplirse entonces $N = \sum_i F_i$.
- Significa que esta información depende del tamaño de la muestra, por lo que es más útil...
- La **frecuencia relativa** es la frecuencia absoluta dividida entre el tamaño de la muestra:

$$f_i = \frac{F_i}{N} = \frac{F_i}{\sum_i F_i}$$

Frecuencias

Diagrama de barras de frecuencias absolutas y relativas

En el eje de abscisas X, colocamos los distintos valores posibles que toman la muestra. Si la cantidad de valores posibles es muy grande, se toman agrupados.
En el eje de ordenadas Y, podemos colocar tanto las frecuencias absolutas (F_i) como las frecuencias relativas (f_i).

Frecuencias

Diagrama de barras de frecuencias absolutas y relativas

En el eje de abscisas X, colocamos los distintos valores posibles que toman la muestra. Si la cantidad de valores posibles es muy grande, se toman agrupados.
En el eje de ordenadas Y, podemos colocar tanto las frecuencias absolutas (F_i) como las frecuencias relativas (f_i).

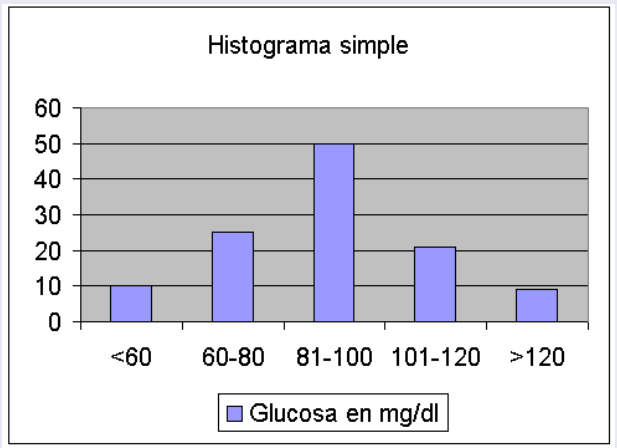


Figura: www.hrc.es

Frecuencias

Histogramas

Los histogramas son el equivalente al diagrama de barras cuando los datos son continuos. Esta variante de los ejemplos anteriores se diferencia en que las barras están juntas, y la frecuencia relativa no está representada por la altura de la barra, si no por su superficie

Frecuencias

Diagrama de tallo y hoja

- Este diagrama es más sencillo de hacer que el anterior para muestras no muy grandes, y tiene la ventaja de aportar más información sobre la distribución dentro de los grupos.

Frecuencias

Diagrama de tallo y hoja

- Este diagrama es más sencillo de hacer que el anterior para muestras no muy grandes, y tiene la ventaja de aportar más información sobre la distribución dentro de los grupos.
- Empezamos separando cada dato numérico en dos partes:
- Primero, se selecciona las cifras significativas (según los datos) por un lado, que forman el **tallo**.
- Trazamos una línea vertical, y se incluyen al otro lado el resto de las cifras del número. Repetimos este proceso para todos los datos. Ejemplo:
  $348 = 34 \mid 8$; $2523 = 25 \mid 23$; $341 = 34 \mid 1$;

Frecuencias

Diagrama de tallo y hoja

- Este diagrama es más sencillo de hacer que el anterior para muestras no muy grandes, y tiene la ventaja de aportar más información sobre la distribución dentro de los grupos.
- Empezamos separando cada dato numérico en dos partes:
- Primero, se selecciona las cifras significativas (según los datos) por un lado, que forman el **tallo**.
- Trazamos una línea vertical, y se incluyen al otro lado el resto de las cifras del número. Repetimos este proceso para todos los datos. Ejemplo:
  $348 = 34 \mid 8$; $2523 = 25 \mid 23$; $341 = 34 \mid 1$;
- Colocamos en la parte izquierda de la recta la parte izquierda de los datos, de modo que el menor dato esté arriba, y el mayor abajo.

Frecuencias

Diagrama de tallo y hoja

- Este diagrama es más sencillo de hacer que el anterior para muestras no muy grandes, y tiene la ventaja de aportar más información sobre la distribución dentro de los grupos.
- Empezamos separando cada dato numérico en dos partes:
- Primero, se selecciona las cifras significativas (según los datos) por un lado, que forman el **tallo**.
- Trazamos una línea vertical, y se incluyen al otro lado el resto de las cifras del número. Repetimos este proceso para todos los datos. Ejemplo:
  $348 = 34 \mid 8$; $2523 = 25 \mid 23$; $341 = 34 \mid 1$;
- Colocamos en la parte izquierda de la recta la parte izquierda de los datos, de modo que el menor dato esté arriba, y el mayor abajo.
- Por último, añadimos a la derecha de la recta, la unidad que

Frecuencias

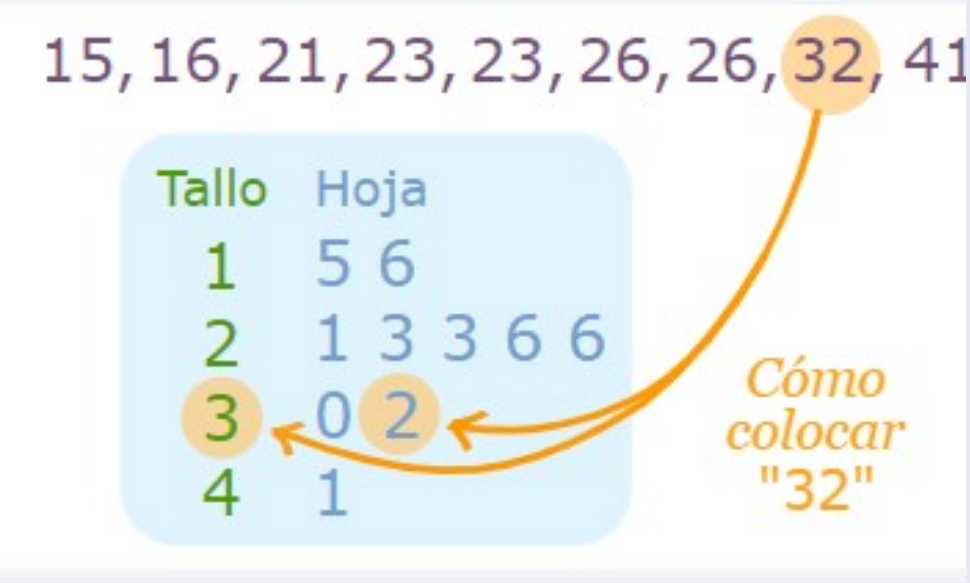
Diagrama de tallo y hoja

- Este diagrama es más sencillo de hacer que el anterior para muestras no muy grandes, y tiene la ventaja de aportar más información sobre la distribución dentro de los grupos.
- Empezamos separando cada dato numérico en dos partes:
- Primero, se selecciona las cifras significativas (según los datos) por un lado, que forman el **tallo**.
- Trazamos una línea vertical, y se incluyen al otro lado el resto de las cifras del número. Repetimos este proceso para todos los datos. Ejemplo:
  $348 = 34 \mid 8$; $2523 = 25 \mid 23$; $341 = 34 \mid 1$;
- Colocamos en la parte izquierda de la recta la parte izquierda de los datos, de modo que el menor dato esté arriba, y el mayor abajo.
- Por último, añadimos a la derecha de la recta, la unidad que

Frecuencias

Diagrama de tallo y hoja

- Opcionalmente, se puede colocar a la izquierda del todo la frecuencia total de los datos agrupados.
- Si volteamos el diagrama de tallo y hoja, el resultado se parece a un histograma.



Medidas de posición

Medidas de centralización

Un tipo de medida útil son conocidas como medidas de centralización.

Medidas de posición

Medidas de centralización

Un tipo de medida útil son conocidas como medidas de centralización.

La más famosa es la **media aritmética**: la suma de todos los datos dividida por la cantidad de datos.

Medidas de posición

Medidas de centralización

Un tipo de medida útil son conocidas como medidas de centralización.

La más famosa es la **media aritmética**: la suma de todos los datos dividida por la cantidad de datos.

Matemáticamente, se denota por $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Ejemplo:

Medidas de posición

Medidas de centralización

Un tipo de medida útil son conocidas como medidas de centralización.

La más famosa es la **media aritmética**: la suma de todos los datos dividida por la cantidad de datos.

Matemáticamente, se denota por $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Ejemplo: Para una muestra de datos: $4, 7, 5.5, 6, 5, 4.5, 6, 7, 5, 6$, la media resultaría:

Medidas de posición

Medidas de centralización

Un tipo de medida útil son conocidas como medidas de centralización.

La más famosa es la **media aritmética**: la suma de todos los datos dividida por la cantidad de datos.

Matemáticamente, se denota por $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Ejemplo: Para una muestra de datos: $4, 7, 5.5, 6, 5, 4.5, 6, 7, 5, 6$, la media resultaría:

$$\overline{x} = \frac{4+7+5.5+6+5+4.5+6+7+5+6}{10} = \frac{56}{10} = 5.6$$

Medidas de posición

Medidas de centralización

Un tipo de medida útil son conocidas como medidas de centralización.

La más famosa es la **media aritmética**: la suma de todos los datos dividida por la cantidad de datos.

Matemáticamente, se denota por $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Ejemplo: Para una muestra de datos: $4, 7, 5.5, 6, 5, 4.5, 6, 7, 5, 6$, la media resultaría:

$$\overline{x} = \frac{4+7+5.5+6+5+4.5+6+7+5+6}{10} = \frac{56}{10} = 5.6$$

La media es problemática cuando en la muestra aparecen valores extremos (outliers), ya que es muy sensible a éstos.

Medidas de posición

Medidas de centralización

Otra medida de centralización es la **mediana**:

Medidas de posición

Medidas de centralización

Otra medida de centralización es la **mediana**: el valor que deja a la mitad de la muestra por debajo (y a la otra mitad por encima).

Medidas de posición

Medidas de centralización

Otra medida de centralización es la **mediana**: el valor que deja a la mitad de la muestra por debajo (y a la otra mitad por encima).
Si ordenamos la muestra de menor a mayor, la mediana será:

- $\frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}$ si n es par.
- $x_{\frac{n+1}{2}}$ si n es impar.

Medidas de posición

Medidas de centralización

Otra medida de centralización es la **mediana**: el valor que deja a la mitad de la muestra por debajo (y a la otra mitad por encima).
Si ordenamos la muestra de menor a mayor, la mediana será:

- $\frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}$ si n es par.
- $x_{\frac{n+1}{2}}$ si n es impar.

A diferencia de la media, la mediana no se ve afectada por valores extremos, pero esto conlleva que esté determinada principalmente por los valores medios.

Medidas de posición

Media vs mediana

- Si la distribución es simétrica, la media y la mediana tendrán valores similares.

Medidas de posición

Media vs mediana

- Si la distribución es simétrica, la media y la mediana tendrán valores similares.
- Si la distribución está sesgada positivamente,la media será más grande que la mediana.

Medidas de posición

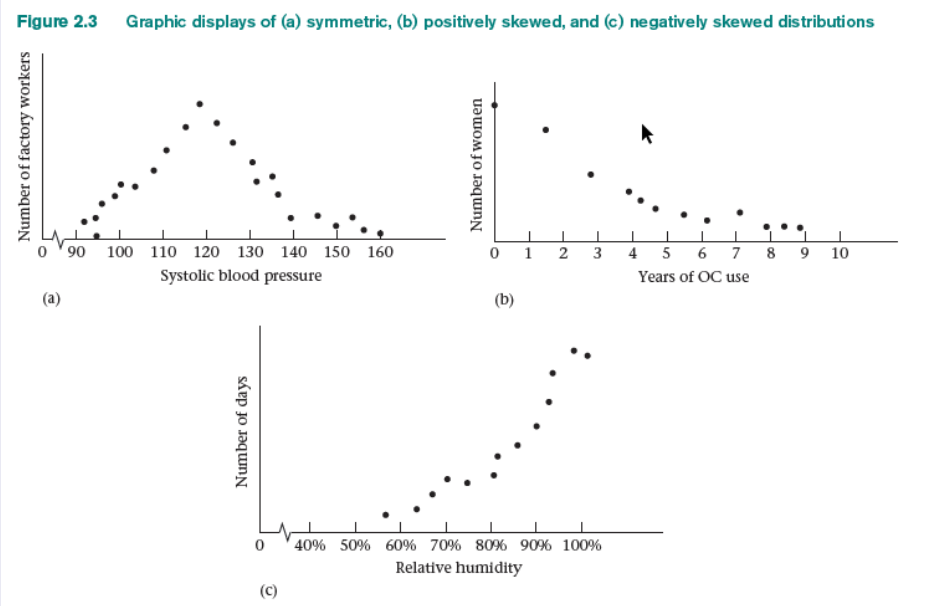
Media vs mediana

- Si la distribución es simétrica, la media y la mediana tendrán valores similares.
- Si la distribución está sesgada positivamente,la media será más grande que la mediana.
- Por último, si la distribución está sesgada negativamente, la mediana será más grande que la media.

Medidas de posición

Media vs mediana

- Si la distribución es simétrica, la media y la mediana tendrán valores similares.
- Si la distribución está sesgada positivamente,la media será más grande que la mediana.
- Por último, si la distribución está sesgada negativamente, la mediana será más grande que la media.



Medidas de posición

Propiedades de la media

Si tenemos una muestra de datos $x_1, x_2, ..., x_n$, podemos obtener su media $\overline{x}$.

Medidas de posición

Propiedades de la media

Si tenemos una muestra de datos $x_1, x_2, ..., x_n$, podemos obtener su media $\overline{x}$. Pero, si desplazamos todos esos datos una cantidad fija, es decir, hacemos $y_i = x_i + c$, tenemos una muestra trasladada.

Medidas de posición

Propiedades de la media

Si tenemos una muestra de datos $x_1, x_2, ..., x_n$, podemos obtener su media $\overline{x}$. Pero, si desplazamos todos esos datos una cantidad fija, es decir, hacemos $y_i = x_i + c$, tenemos una muestra trasladada. ¿Cuál será su media?

Medidas de posición

Propiedades de la media

Si tenemos una muestra de datos $x_1, x_2, ..., x_n$, podemos obtener su media $\overline{x}$. Pero, si desplazamos todos esos datos una cantidad fija, es decir, hacemos $y_i = x_i + c$, tenemos una muestra trasladada.
¿Cuál será su media?
La respuesta es:

Medidas de posición

Propiedades de la media

Si tenemos una muestra de datos $x_1, x_2, ..., x_n$, podemos obtener su media $\overline{x}$. Pero, si desplazamos todos esos datos una cantidad fija, es decir, hacemos $y_i = x_i + c$, tenemos una muestra trasladada.
¿Cuál será su media?
La respuesta es: $\overline{y} = \overline{x} + c$

Medidas de posición

Propiedades de la media

Por otra parte, si para la muestra $x_1, x_2, ..., x_n$ obtenemos la muestra con un reescalado: $y_i = c \cdot x_i$, la media de la nueva muestra es:

Medidas de posición

Propiedades de la media

Por otra parte, si para la muestra $x_1, x_2, ..., x_n$ obtenemos la muestra con un reescalado: $y_i = c \cdot x_i$, la media de la nueva muestra es:
$\overline{y} = c \cdot \overline{x}$

Medidas de posición

Cuartiles

- Los cuartiles son medidas de tendencia no central. Estos son:
  - Q_1: el que deja por debajo el 25 % de los datos.
  - Q_2 el que deja por debajo el 50 % de los datos. Es la mediana.
  - Q_3: el que deja por debajo el 75 % de los datos.

Medidas de dispersión

Medidas de dispersión

Dos muestras distintas pueden tener una media aritmética igual y ser radicalmente distintas:

Medidas de dispersión

Medidas de dispersión

Dos muestras distintas pueden tener una media aritmética igual y ser radicalmente distintas:
Este fenómeno se conoce como dispersión o variabilidad. ¿Cómo medimos la dispersión?

Medidas de dispersión

Medidas de dispersión

- Una primera medida es el **rango**: la distancia del valor más grande al más pequeño.

Medidas de dispersión

Medidas de dispersión

- Una primera medida es el **rango**: la distancia del valor más grande al más pequeño.
- Tiene la ventaja de ser fácil de calcular, pero, al igual que la media, es muy sensible a los valores extremos.

Medidas de dispersión

Medidas de dispersión

- Una primera medida es el **rango**: la distancia del valor más grande al más pequeño.
- Tiene la ventaja de ser fácil de calcular, pero, al igual que la media, es muy sensible a los valores extremos.
- Otra medida interesante es el **rango intercuartil (RIC)**, que se define como la diferencia entre el tercer cuartil (75-percentil) y el primer cuartil (25-percentil), es decir, $Q_3 - Q_1$.

Medidas de dispersión

Medidas de dispersión

- Cuando hablamos de dispersión, nos referimos la diferencia que hay entre los datos y la media, es decir: $x_1 - \overline{x}, ..., x_n - \overline{x}$.

Medidas de dispersión

Medidas de dispersión

- Cuando hablamos de dispersión, nos referimos la diferencia que hay entre los datos y la media, es decir: $x_1 - \overline{x}, ..., x_n - \overline{x}$.
- Una manera de medir la dispersión podría ser la media de todas esas diferencias: $d = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n}$.

Medidas de dispersión

Medidas de dispersión

- Cuando hablamos de dispersión, nos referimos la diferencia que hay entre los datos y la media, es decir: $x_1 - \overline{x}, ..., x_n - \overline{x}$.
- Una manera de medir la dispersión podría ser la media de todas esas diferencias: $d = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n}$.
- ¿Qué ocurre con d?

Medidas de dispersión

Medidas de dispersión

- Cuando hablamos de dispersión, nos referimos la diferencia que hay entre los datos y la media, es decir: $x_1 - \overline{x}, ..., x_n - \overline{x}$.
- Una manera de medir la dispersión podría ser la media de todas esas diferencias: $d = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n}$.
- ¿Qué ocurre con d? Que siempre vale 0.

Medidas de dispersión

Varianza muestral

Para arreglar esto, se puede tomar el valor absoluto de las diferencias (**desviación media**):

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

Medidas de dispersión

Varianza muestral

Para arreglar esto, se puede tomar el valor absoluto de las diferencias (**desviación media**):

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

O la más usada, la **varianza muestral** :

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

(En realidad, el denominador es $n - 1$, como veremos más adelante).

Medidas de dispersión

Varianza muestral

Para arreglar esto, se puede tomar el valor absoluto de las diferencias (**desviación media**):

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

O la más usada, la **varianza muestral** :

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

(En realidad, el denominador es $n - 1$, como veremos más adelante).
Por último, la raíz cuadrada de varianza muestral se conoce como desviación estándar:

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

Medidas de dispersión

Propiedades de la varianza muestral

¿Qué ocurrirá con la varianza (o la desviación típica) al someter a nuestra muestra a transformaciones del tipo $y_i = x_i + c$ e $y_i = c \cdot x_i$

Medidas de dispersión

Propiedades de la varianza muestral

¿Qué ocurrirá con la varianza (o la desviación típica) al someter a nuestra muestra a transformaciones del tipo $y_i = x_i + c$ e $y_i = c \cdot x_i$
Al trasladar la muestra, la dispersión permanece inalterada, luego $s_y^2 = s_x^2$

Medidas de dispersión

Propiedades de la varianza muestral

¿Qué ocurrirá con la varianza (o la desviación típica) al someter a nuestra muestra a transformaciones del tipo $y_i = x_i + c$ e $y_i = c \cdot x_i$

Al trasladar la muestra, la dispersión permanece inalterada, luego $s_y^2 = s_x^2$

Al reescalar la muestra, la dispersión si se modifica, resultado $s_y^2 = c^2 \cdot s_x^2$

Métodos gráficos

Resumiendo, hemos hallado para una muestra, las siguientes medidas:

Métodos gráficos

Resumiendo, hemos hallado para una muestra, las siguientes medidas:

- Frecuencia
- Media aritmética: $\overline{x}$
- Mediana
- Rango
- p-percentiles
- Rango intercuartil
- Varianza muestral (y desviación estándar)

Métodos gráficos

Diagramas de cajas (*y bigotes*)

Para dibujar este diagrama, necesitamos:

- Mediana
- El primer y tercer cuartil (Q_1 y Q_3)
- RIC ($Q_3 - Q_1$)

Este tipo de diagramas son muy útiles para observar si hay simetría en la muestra

Métodos gráficos

Diagramas de cajas (*y bigotes*)

- Paso 1. Dibujamos una caja. Los extremos de la caja son Q_1 y Q_3

Métodos gráficos

Diagramas de cajas (*y bigotes*)

- Paso 1. Dibujamos una caja. Los extremos de la caja son Q_1 y Q_3
- Paso 2. Dibujamos una recta dentro de la caja para indicar la mediana (Q_2)

Métodos gráficos

Diagramas de cajas (*y bigotes*)

- Paso 1. Dibujamos una caja. Los extremos de la caja son Q_1 y Q_3
- Paso 2. Dibujamos una recta dentro de la caja para indicar la mediana (Q_2)
- Paso 3. Con el RIC, consideramos valores atípicos (outliers) a los valores que sean $x > Q_3 + 1.5 \cdot RIC$ y $x < Q_1 - 1.5 \cdot RIC$ y extremadamente atípicos a aquellos que sean $x > Q_3 + 3 \cdot RIC$ y $x < Q_1 - 3 \cdot RIC$.

Métodos gráficos

Diagramas de cajas (*y bigotes*)

- Paso 1. Dibujamos una caja. Los extremos de la caja son Q_1 y Q_3
- Paso 2. Dibujamos una recta dentro de la caja para indicar la mediana (Q_2)
- Paso 3. Con el RIC, consideramos valores atípicos (outliers) a los valores que sean $x > Q_3 + 1.5 \cdot RIC$ y $x < Q_1 - 1.5 \cdot RIC$ y extremadamente atípicos a aquellos que sean $x > Q_3 + 3 \cdot RIC$ y $x < Q_1 - 3 \cdot RIC$.
- Paso 4. Dibujamos los *bigotes* como los últimos valores que no sean atípicos.
- Paso 5. Por último, dibujamos con 0 los valores atípicos y con $*$ los valores extremadamente atípicos.

Métodos gráficos

Diagramas de cajas (*y bigotes*)