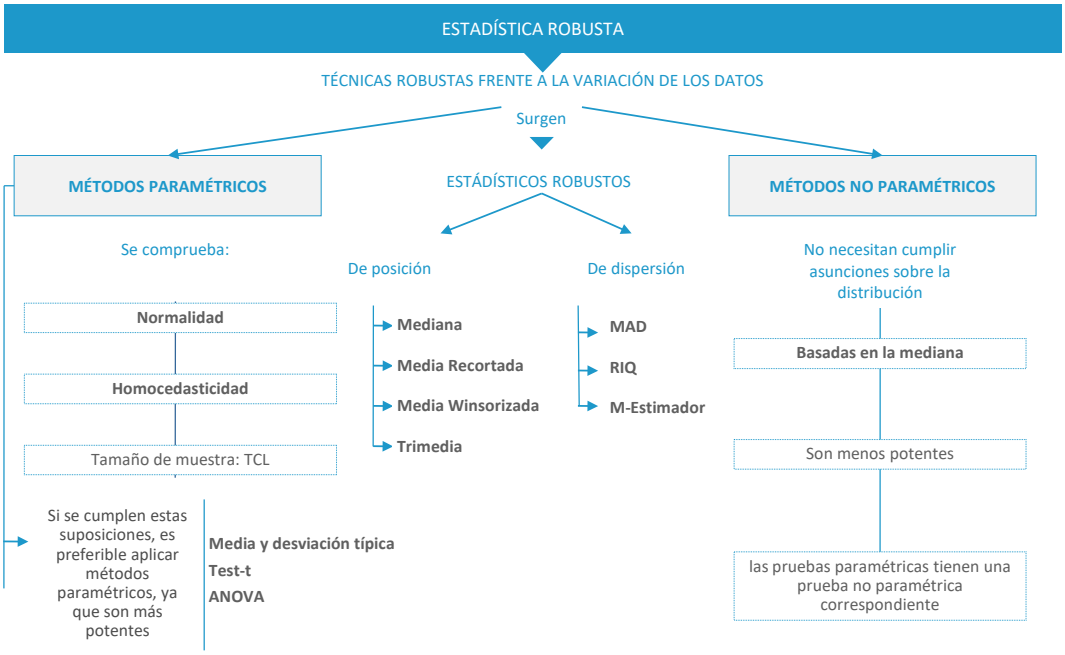
Técnicas Multivariantes

Estadística robusta

Índice

Esquema



Ideas clave

2.1 Introducción y objetivos

En este tema se explicará en qué consiste la estadística robusta, cuál es la diferencia entre métodos paramétricos y no paramétricos, y qué técnicas permiten tratar con datos atípicos, muy comunes en el análisis estadístico de datos.

La estadística clásica se asienta sobre la base de que los métodos estadísticos cumplen una serie de supuestos que permiten dar fiabilidad a los resultados obtenidos por dichos modelos. Estos supuestos son, entre otros, la normalidad de los datos, la homocedasticidad de la muestra, la independencia de los datos y el tamaño suficiente de muestra, que se explican con más detalle en el próximo subapartado. Por lo tanto, antes de comenzar a realizar un estudio multivariante de los datos según la estadística clásica, sería necesario comprobar el comportamiento de cada una de las variables por separado y comprobar si dichas variables cumplen los supuestos establecidos (aunque también puede resultar un problema hacer un test para cada variable, ya que se pierde potencia al final).

No obstante, en muchas ocasiones, estos supuestos o bien no se pueden cumplir, o bien se obvian por parte de los investigadores, dando lugar a resultados que no son del todo fiables. Esto ocurre, por ejemplo, cuando la muestra presenta valores atípicos que impiden la normalidad de los datos, y a su vez sesgan los parámetros de la misma en que se basan los modelos estadísticos (por ejemplo, la media). Supongamos las siguientes distribuciones que hacen referencia a las notas de dos alumnos distintos:

- Distribución A = {6, 6, 7, 5, 6}.

- Distribución B = {8, 7, 8, 7, 0}.

Aunque tengan la misma media ($\mu = 5$), no son iguales las distribuciones de las notas, por lo que atendiendo únicamente a la media para definir estas distribuciones, las conclusiones que se alcancen pueden ser poco apropiadas.

En cambio, el uso de estadística robusta proporciona técnicas que permiten establecer conclusiones adecuadas a pesar de no cumplir con todos los supuestos estadísticos anteriores. Su nombre hace referencia a que no se ve afectada por las anomalías de las muestras, siendo, por tanto, más "robusta".

2.2 Métodos paramétricos

En estadística se suele hacer referencia a dos tipos de métodos según los supuestos que deben cumplir las poblaciones que se muestrean: los métodos paramétricos y los métodos no paramétricos.

Los métodos paramétricos forman parte de la estadística paramétrica, la cual es una rama de la estadística inferencial que comprende los procedimientos estadísticos y de decisión que están basados en distribuciones conocidas. Estas son determinadas usando un número finito de parámetros: el estimador clásico de localización, empleado en los métodos paramétricos es la media, mientras que el estimador clásico de dispersión es la desviación típica.

Así pues, por ejemplo, si conocemos que la altura de las personas sigue una distribución normal, pero desconocemos cuál es la media y la desviación de dicha normal. La media y la desviación típica de la distribución normal son los dos parámetros que queremos estimar. Cuando desconocemos totalmente qué distribución siguen nuestros datos entonces deberemos aplicar primero un test no paramétrico, que nos ayude a conocer primero la distribución.

La mayoría de procedimientos paramétricos requiere conocer la forma de distribución para las mediciones resultantes de la población estudiada. Para la inferencia paramétrica es requerida como mínimo una escala de intervalo, esto quiere decir que nuestros datos deben tener un orden y una numeración del intervalo. Por ejemplo, si estamos midiendo la variable edad, podríamos emplear los siguientes intervalos: menores de 10 años, de 10 a 30 años, de 30 a 50, de 50 a 70, etc. ya que hay números con los cuales realizar cálculos estadísticos. Sin embargo, datos categorizados en: niños, jóvenes, adultos y ancianos no pueden ser interpretados mediante la estadística paramétrica ya que no se puede hallar un parámetro numérico (como por ejemplo la media de edad) cuando los datos no son numéricos.

En el caso de que la muestra o muestras de estudio no cumplan con los supuestos específicos para realizar el modelo (por ejemplo, los supuestos del modelo de regresión), o cuando se sospeche que la muestra contiene datos anómalos, o cuando no proceda de una única población (por ejemplo, en muchos casos de big data) no se deberían aplicar métodos paramétricos.

A continuación, se definen los supuestos más relevantes que van a permitir diferenciar entre los métodos paramétricos y los no paramétricos.

Normalidad

La normalidad es una de las hipótesis más frecuentes en estadística paramétrica. Este supuesto hace referencia a que las variables o las muestras utilizadas de las mismas deben seguir una distribución normal de los datos. Ciertos modelos estadísticos, como los contrastes de medias o el ANOVA (análisis de varianzas) requieren de la comprobación de este supuesto para poder realizarse. Para comprobar si la distribución teórica se ajusta a una distribución normal, se realizan diferentes tipos de pruebas. Las más conocidas son la *Prueba de Kolmogorov-Smirnov*, utilizada con tamaños de muestra grandes, y la *Prueba de Shapiro-Wilk*, válida para tamaños de muestra no muy grandes ($n < 5000$). Hay que tener en cuenta que en la *Prueba de Kolmogorov-Smirnov* se com-

paran dos distribuciones (habitualmente se compara la muestra sobre la que queremos comprobar que proviene de una distribución normal frente a una distribución normal), mientras que en la *Prueba de Shapiro-Wilk* sólo se evalua la muestra de interés.

Tanto en la *Prueba de Kolmogorov-Smirnov* como en la de *Shapiro-Wilk*, la hipotesis nula consiste en la normalidad de los datos, por tanto, el p-valor que debe aparecer para no rechazar la normalidad es $\alpha > 0.05$. Si es así, se da por válida la hipótesis de que los datos se distribuyen de manera normal.

Por otro lado, también es adecuado analizar la distribución de los valores mediante técnicas de visualización, lo que permitirá comprobar la existencia de datos anómalos (también llamados datos aberrantes u *outliers*) y comprobar si se trata de posibles errores. Una de las opciones es utilizar un gráfico q-q (*q-q plot*) donde se representan los valores reales de los percentiles de la distribución frente a los teóricos de una distribución normal. Cuanto más se aproximen los puntos a la recta, más similitud se tendrá con la distribución normal. En la Figura 1 se muestra una distribución normal de media $\mu = 0$ y desviación típica $\sigma = 1$ (obtenida mediante la simulación de una muestra de tamaño 1000 de dicha distribución) y en la Figura 2 se muestra una distribución uniforme [0, 1] (obtenida mediante la simulación de una muestra de tamaño 1000 de dicha distribución).

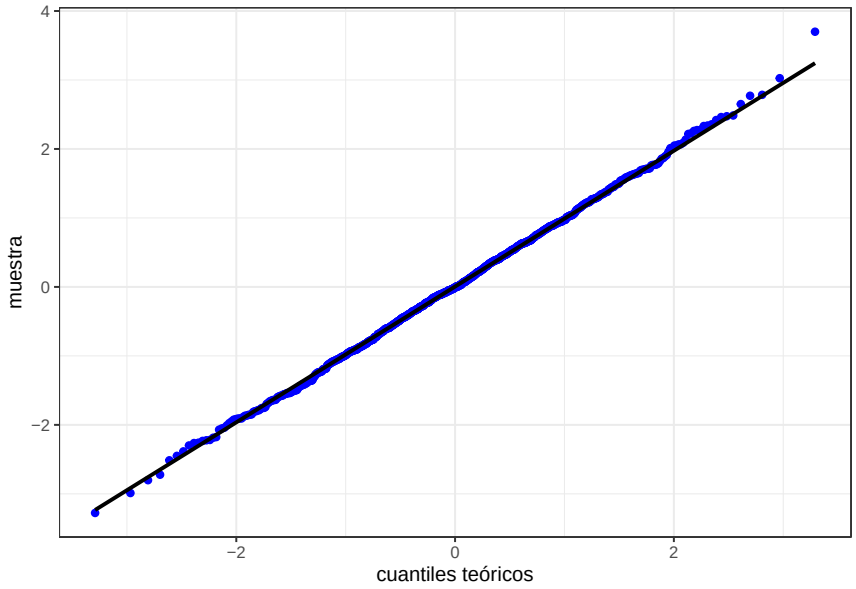


Figura 1: Gráfico q-q de una muestra normal

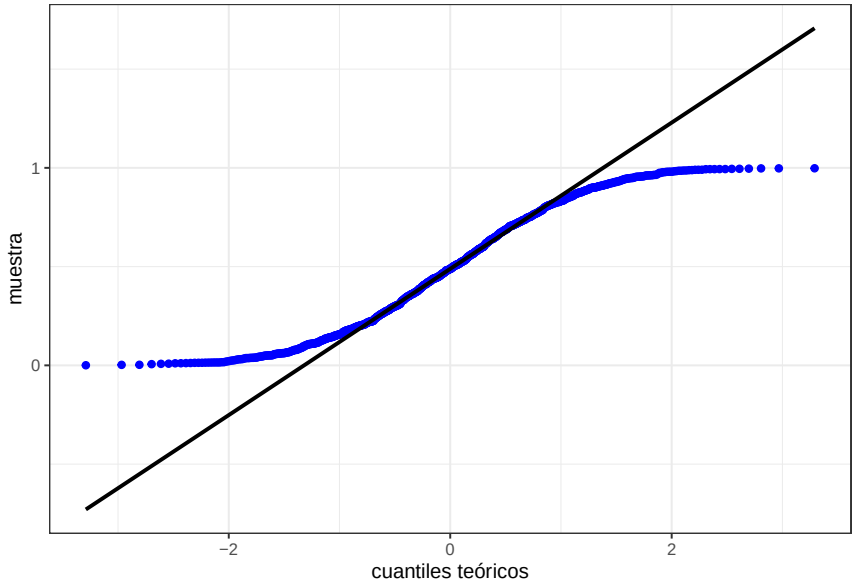


Figura 2: Gráfico q-q de una muestra uniforme

En el primer caso, se puede observar como la distribución de los cuantiles se ajusta a la recta, mientras que en el segundo caso se observa que no. Además, cuantas más variables haya, más potencia perderemos comprobando la normalidad de cada una (es decir, aumentaremos la probabilidad de error II). También cabe destacar que aunque el *qq plot* sirve para dar una idea intuitiva y gráfica sobre la normalidad, al no dar un p-valor no es un método tan fiable como los test comentados anteriormente.

Homocedasticidad.

La homocedasticidad, o igualdad de varianzas, hace referencia a que la varianza de los errores es constante a lo largo del tiempo. Por tanto, será importante cuando tengamos que enfrentarnos a problemas de regresión o series temporales, ya que cuando se verifica esta hipótesis el modelo resultante es mucho más sencillo. En el caso de que los errores presenten una varianza no constante, estaríamos ante heterocedasticidad, lo cual impide la estabilidad de los datos y la aplicación de métodos estadísticos clásicos. En generalnnna vgeneral, la varianza no constante surge en presencia de valores atípicoos valores tienen demasiado peso, por lo que influyen desproporcionadamente en el rendimiento del modelo. De nuevo, la visualización gráfica puede resultar muy útil.

Para ello, se puede realizar un diagrama de dispersión de los residuos. En la Figura 3 se muestran los residuos de un modelo de regresión lineal donde el error del modelo (y por tanto los residuos del ajuste) son proporcionales a X.

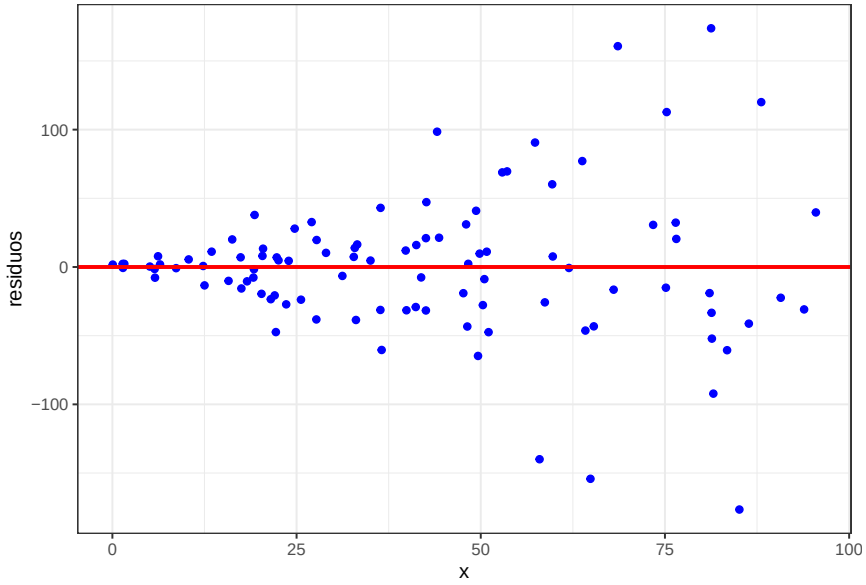


Figura 3: Distribución de los residuos

```
##
##  studentized Breusch-Pagan test
##
## data:  y_het ~ muestra_unif
## BP = 18.625, df = 1, p-value = 1.591e-05
```

Se comprueba, por tanto, que no se cumple la homocedasticidad, ya que:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon \cdot X.$$

Si no se quiere depender del análisis visual, se puede emplear una prueba estadística. Es decir, se puede comprobar estadísticamente la presencia o ausencia de heterocedasticidad. Para ello, la prueba de Breusch-Pagan es la más empleada. Si el test muestra un p-valor > 0.05, entonces podemos asumir la hipótesis de homocedásticidad. En el ejemplo de la Figura 3 se obtiene es la prueba de Breusch-Pagan un p-valor = 0.00002,

por lo que no podemos rechazar la hipótesis nula de presencia de heterocedasticidad.

Tamaño de la muestra

Los métodos paramétricos exigen un tamaño muestral mínimo, se utilice una distribución normal o no. Además, como se ha visto anteriormente, la distribución normal se utiliza en muchas ocasiones y presenta muy buenas propiedades: los errores son aleatorios, simétricos, etc. ¿Y qué pasa si algo no es normal?

En algunas ocasiones, cuando el tamaño de la muestra es elevado, es posible considerar que se cumple el Teorema Central del Límite. Este teorema establece que la suma o el promedio de casi cualquier conjunto de variables independientes generadas al azar se aproximan a la distribución normal. El Teorema central del límite explica por qué la distribución normal surge tan comúnmente y por qué es generalmente una aproximación excelente para definir el comportamiento de la media muestral de casi cualquier colección de datos.

Este hallazgo se mantiene sin importar la forma que adopte la distribución de datos que tomemos. Para ilustrar este teorema, se puede considerar diferentes distribuciones, como la binomial, la exponencial, la geométrica o la distribución de Poisson, para las cuales se han simulado una distribución de tamaño muestral 1000 tal y como se muestra en la Figura 4.

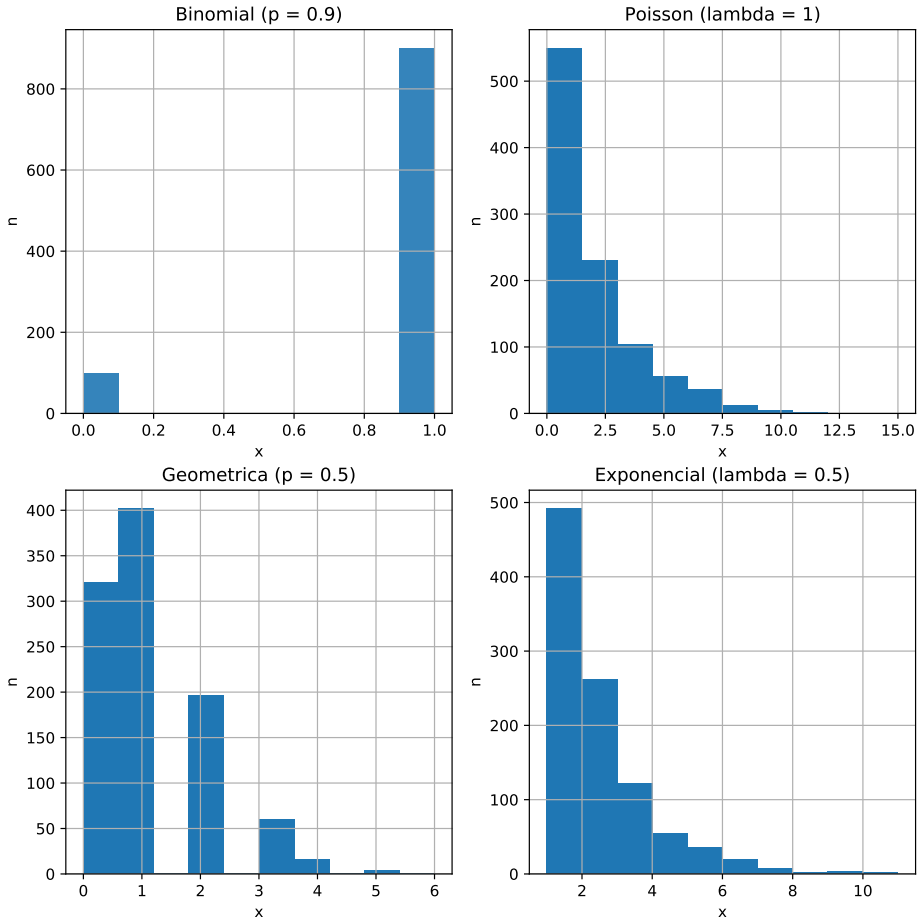


Figura 4: Distribución simulada (n=1000) para diferentes distribuciones

Para cada una de las distribuciones de la Figura 4 se calcula la media muestral. Este proceso, se repite 20000 veces para obtener la distribución en el muestreo de la de la media muestral. La distribución de esta media muestral se puede comprobar empirícamente (mediante simulación) que se distribuye como una distribución normal. A continuación, en la Figura 5 se muestra gráficamente, el resultado de la distribución de la media muestral de las distintas distribuciones.

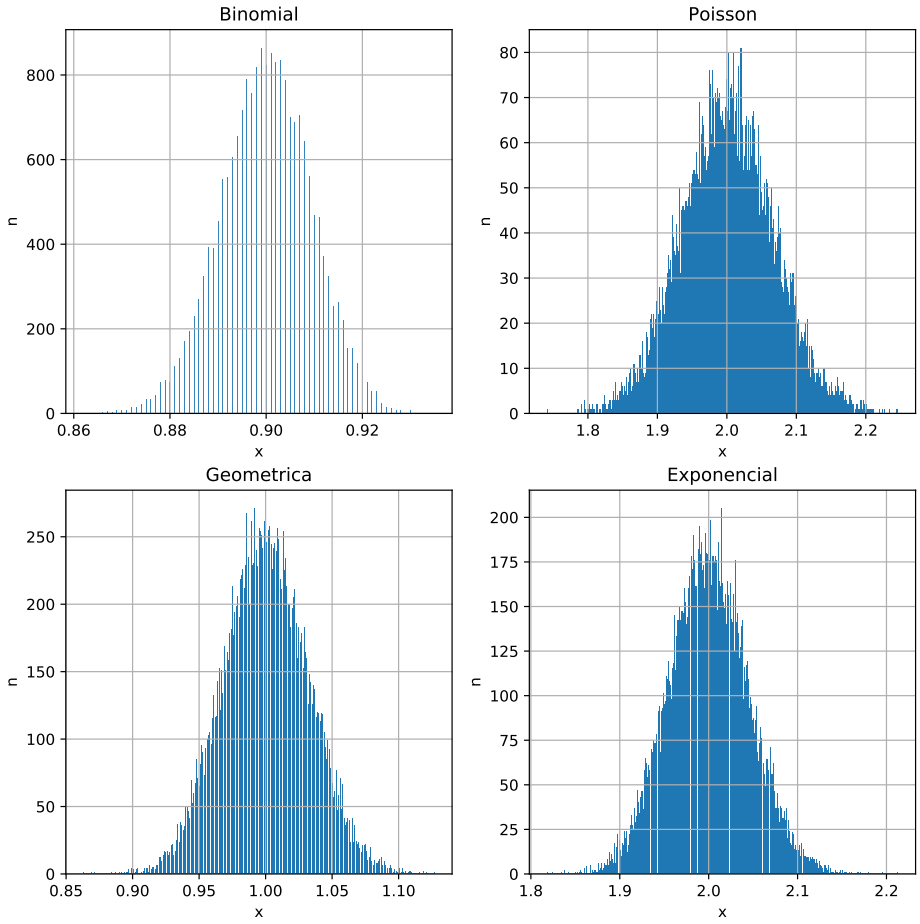


Figura 5: Distribución en el muestreo de la media muestral (r=20000) para diferentes distribuciones de tamaño n=1000

Se comprueba en este ejemplo, que al graficar la distribución en el muestreo de las medias muestrales de las distribuciones Binomial, Poisson, Geométrica y Exponencial, todas ellas responden a la famosa forma de campana de la Distribución Normal.

2.3 Métodos no paramétricos

Una prueba no paramétrica es una prueba de hipótesis que no requiere que la distribución de la población sea caracterizada por ciertos parámetros. Por ejemplo, muchas

pruebas de hipótesis parten del supuesto de que la población sigue una distribución normal con los parámetros μ y σ. En cambio, las pruebas no paramétricas no parten de este supuesto, de modo que son útiles cuando los datos son no normales, o cuando el tamaño muestral es demasiado pequeño, o cuando hay una presencia considerable de outliers o cuando sospechamos que los datos de la muestra pueden provenir de más de una población.

Sin embargo, las pruebas no paramétricas tienen algunas limitaciones:

- En general, **son menos potentes** que la prueba paramétrica correspondiente cuando se cumple el supuesto de normalidad. Por lo tanto, es más difícil que se rechace la hipótesis nula cuando no sea cierta si los datos provienen de la distribución normal.

- Requieren que se modifiquen las hipótesis. Por ejemplo, la mayoría de las pruebas no paramétricas acerca del centro de la población en lugar de emplear la media, emplean **la mediana**.

A continuación, en la Tabla 1 se muestra la correspondencia entre las distintas pruebas paramétricas y no paramétricas.

Tabla 1: Pruebas estadísticas

Pruebas paramétricas y no paramétricas		
Tipo de prueba	Paramétrica	No Paramétrica
Describir una muestra	Media y varianza	Mediana, rango intercuartil
Comparar una muestra con un valor	Test t de una muestra (media)	Prueba Wilcoxon (mediana)
Comparar dos muestras independientes	Test t de igualdad de medias	Prueba Mann-Whitney (mediana)
Comparar dos muestras apareadas	Test t de igualdad de medias apareadas	Prueba Wilcoxon (mediana)
Para *n* muestras	ANOVA	Prueba Kruskal-Wallis

2.4 Tratamiento de outliers

Uno de los elementos que más se reproducen en los análisis de datos es la aparición de *outliers* o elementos anómalos. Un *outlier* es una observación anormal en una muestra estadística o serie temporal de datos que puede afectar potencialmente a la estimación de los parámetros del modelo.

Imaginemos un problema en el cual queremos estimar la media poblacional de la altura de los alumnos de segundo de bachillerato. Para ello realizamos muestras de 10 alumnos. Supongamos que extraemos una muestra $X_1 = \{1.65, 1.80, 1.71, 1.69, 1.75, 1.85, 1.62, 1.79, 1.80, 1.71\}$. La media de altura de la muestra sería de 1.73. Si tenemos en cuenta la altura máxima (1.85) y la altura mínima (1.62) y la distancia entre estas a la media, vemos que es de 0.113 y 0.117 respectivamente. Como podemos observar la media se sitúa aproximadamente en la

mitad de intervalo y se podría considerar como una estimación bastante buena.

Ahora pensemos en otra muestra X_2 de otros 10 alumnos, siendo sus alturas las siguientes: $X2 = \{1.65, 1.80, 1.71, 1.69, 2.18, 2.20, 1.62, 1.79, 1.74, 1.70\}$

En este caso, la altura media de la clase sería de 1.81. Si ahora nos fijamos en la altura máxima (2.20) y en la altura mínima (1.62) y la distancia entre estas a la media, vemos que es de 0.39 y 0.18 respectivamente. En este caso la media ya no está situada aproximadamente en la mitad del intervalo. El efecto de las dos observaciones más extremas (2.18 y 2.20) ha hecho que la media aritmética se haya desplazado hacia el valor máximo de la distribución. Con este ejemplo, vemos el efecto que tienen los outliers y cómo pueden desvirtuar el cálculo de una media.

Ante la existencia de valores atípicos que pueden distorsionar las estimaciones obtenidas caben dos opciones:

1. Detectar los valores atípicos, u outliers, y eliminarlos de la muestra para, posteriormente, aplicar los métodos tradicionales de estimación.

2. Diseñar métodos robustos de estimación que no se vean muy afectados por la existencia de estos outliers aunque, si la muestra no tuviera valores anómalos, las estimaciones fueran algo menos eficientes que las tradicionales.

La estadística robusta, por tanto, trata de la aplicación de métodos estadísticos que no se vean afectados por la existencia de valores atípicos en la muestra. Eliminando los outliers reducimos el tamaño muestral y eliminamos información que puede ser relevante. En general, solo deberían eliminarse datos anómalos cuando estemos seguros de que son fruto de un error (por ejemplo, altura de 175m). Cuando tenemos datos multivariantes, en general es complicado detectar los outliers. Existen también técnicas clásicas (paramétricas) para hacerlo, pero suelen utilizarse técnicas robustas.

2.5 Estimadores robustos

Estimadores de localización

Como se ha comentado anteriormente, la media es el estadístico de localización por excelencia. Sin embargo, existe otros estimadores de localización, que son robustos (tal y como se podrá comprobar en la sección de ejercicios resueltos), entre los que destaca la mediana.

- **Mediana**: es el valor de la variable que divide a la distribución en dos partes iguales conteniendo cada una de ellas el 50 % de las observaciones. La mediana muestral resulta mucho menos afectada que la media muestral por la existencia de valores atípicos. Evidentemente, la media es un estadístico muy poco resistente a cambios en los datos, dado que se ve influida por todos y cada uno de ellos. La mediana, en cambio, es un estadístico altamente resistente. Como contrapartida, hay que señalar que, si las observaciones X_i procedieran de una distribución $N(\mu, \sigma)$, la varianza de la mediana muestral sería un 57 % mayor que la de la media muestral siendo, por tanto, menos eficiente utilizar la mediana que la media cuando todos los datos muestrales proceden del modelo supuesto.

Además de la mediana muestral, existen otros estimadores alternativos a la media que son robustos frente a la existencia de outliers. Entre ellos, las α-medias recortadas, las cuales eliminan un porcentaje α de las observaciones muestrales en cada extremo.

- **Media recortada**: Consiste en calcular la media aritmética sobre un subconjunto central del conjunto de datos, no considerándose una determinada proporción α por cada extremo.

La media recortada queda definida por la proporción de casos, α, que son excluidos desde cada extremo de la muestra ordenada. Una vez que se han eliminado los valores

indicados de cada extremo, se calcula el promedio de los valores restantes.

- Si α es un múltiplo de $1/n$, se eliminan de cada extremo un número entero de valores $[\alpha \cdot n]$ y la media recortada es el promedio de los $s = n \cdot (1 - 2 \cdot \alpha)$ valores restantes, que se calcula mediante la fórmula:

$$m(\alpha) = \frac{\sum_i^s X_i}{s}.$$

- Si α no es múltiplo de $1/n$, se elimina el menor número entero de valores $[\alpha \cdot n]$ más cercano a $\alpha \cdot n$ de cada extremo, y al mayor y menor valor restante se le pondera con $p = 1 + [\alpha \cdot n] - \alpha \cdot n$.

Para calcular la media ponderada recortada, en este caso, se usa la fórmula:

$$m(\alpha) = \frac{p \cdot X_1 + X_2 + ... + X_{s-1} + p \cdot X_s}{n(1 - 2\alpha)}.$$

Esto es así porque para calcular la media recortada se eliminan las puntuaciones más alejadas del núcleo central de la distribución por ambos lados, y se trabaja solo con el centro de las puntuaciones, así la media no se ve afectada por las puntuaciones extremas.

Por ejemplo, una media recortada al 40 % ($\alpha = 0.4$) en una secuencia de 10 datos implica no tener en cuenta ni los 4 valores menores ni los 4 valores mayores. Es interesante hacer notar que la media recortada al 0 % es la media aritmética, y que la media recortada al 25 % ($\alpha = 0.25$) se la denomina centrimedia.

- **Media winsorizada**: esta media sustituye los casos excluidos del análisis por el último valor, en cada extremo, que sí forme parte del análisis. Cuando se cambian estos valores, se calcula el promedio de las puntuaciones. Se calculan usando la fórmula:

$$W(\alpha) = \frac{\sum_i^n X_i}{n},$$

donde α es la proporción de casos excluidos por cada extremo. En este caso no se recortan las puntuaciones extremas, como en el caso anterior, sino que, para mantener el mismo tamaño muestral, lo que se hace es sustituir las puntuaciones que se encuentran fuera del núcleo central de la distribución por otras que si están incluidas en él (la más próxima a la puntuación o puntuaciones eliminadas) (Palmer 1999).

- **trimedia**: Es un índice de tendencia central que consiste en calcular una media aritmética ponderada de tres medidas: el primer cuartil, la mediana (con peso doble) y el tercer cuartil.

$$\text{trimedia} = \frac{\text{cuartil 1} + 2 \cdot \text{mediana} + \text{cuartil 3}}{4}.$$

Como ya se ha comentado anteriormente, en general las estimaciones robustas (mediana, media recortada, media winsorizada y trimedia) son menos eficientes que los estimadores clásicos, como la media muestral, cuando las observaciones muestrales proceden de distribuciones normales. Sin embargo, son mucho más robustos cuando hay valores anómalos (su función de influencia está acotada mientras que la de la media muestral no lo está). Para diseñar estimadores que sigan siendo robustos y, además, ganen eficiencia en el caso de distribuciones normales, se han propuesto los denominados M-estimadores que constituyen una familia de estimadores robustos basados en una generalización de los estimadores máximo-verosímiles. Aunque su cálculo resulta a veces complejo, en gran parte del software estadístico especializado se pueden obtener sin esfuerzo, especialmente el **estimador de Huber**, lo cual está permitiendo su difusión en los análisis estadísticos habituales.

En general, un modo intuitivo de comprobar si la muestra puede contener outliers que distorsionen los resultados es calcular estimadores clásicos y robustos y realizar

un análisis de sensibilidad analizando las diferencias entre ellos. Si ambos tipos de estimadores difieren poco, probablemente no haya grandes anomalías en la muestra y los estimadores clásicos sean adecuados; en otro caso, puede ser más conveniente utilizar estimadores robustos.

Estimadores de dispersión

Al igual que existen alternativas robustas a la media (el cual es un estimador de localización), también existen alternativas robustas a la desviación típica como estimador de escala.

- **MAD** (o MEDA): Entre las más habituales se encuentra la desviación absoluta mediana, definida como la mediana de las desviaciones en valor absoluto respecto a la mediana (Me), es decir:

$$MAD = Mediana|i - M| \quad \text{para } i = 1, ..., n.$$

- **Rango intercuartílico**: Otro estimador de dispersión robusto es el rango intercuartílico definido como la distancia entre las cuales está el 50 % central de los valores muestrales, es decir:

$$RIQ = [\text{Percentil } 0.25, \ \text{Percentil } 0.75].$$

- **M-estimador de un paso**: un M-estimador se define como *Maximum Likelihood Estimator* (estimador de máxima verosimilitud). Su objetivo es buscar un índice de localización a partir del conjunto de observaciones, ponderando a éstas en función de lo cerca o lejos que se encuentren del centro de datos. Para el cálculo del M-estimador de una muestra se corta la cola con puntuaciones anormalmente distantes del centro de la misma. En general, conocer el comportamiento de los

estimadores robustos en muestras pequeñas suele ser un problema complejo y, en muchas ocasiones, hay que recurrir a técnicas de remuestreo (bootstrap y jacknife), que se verán en el Tema 3.

Material audiovisual

Accede al vídeo: Distribución de los residuos

2.6 Referencias bibliográficas

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2):133–155.

Evans, M. and Rosenthal, J. S. (2015). *Probabilidad y estadística : la ciencia de la incertidumbre*. Reverté, Barcelona.

Palmer, A. L. (1999). *Análisis de Datos, Etapa Exploratoria*. 01 edition.

Pat Fernández, L. A. (2013). *Introducción a los modelos de regresión*. Plaza y Valdés, México, D.F.

2.7 Ejercicios resueltos

Ejercicio 1

a) Comprueba la normalidad a una muestra obtenida mediante el muestreo aleatorio de una distribución normal con media $\mu = 0.05$ y desviación típica $\sigma = 0.9$ y $n = 1000$.

b) Comprueba la normalidad a una muestra obtenida mediante el muestreo aleatorio de una distribución normal con media $\mu = 0.05$ y desviación típica $\sigma = 0.9$ y $n = 10000$.

Solución

a) Como el tamaño de muestra no es muy grande, se puede emplear la *Prueba de Shapiro-Wilk*. En Python, se realiza utilizando la función *shapiro* de la librería *scipy.stats*. La muestra se obtiene con la función *np.random.normal*, la cual se vió en el tema 1.

El primer paso a realizar cuando trabajamos en Python es cargar los paquetes y funciones que se van a utilizar (recordar que deben haberse instalado previamente).

```
# cargar librerias-----------------------------------------------
import pandas as pd
import numpy as np
from pandas.core.common import flatten
from plotnine import *
from array import *
import scipy.stats as stats
import math
```

```
import matplotlib as mpl
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn import linear_model
import statsmodels.formula.api as smf
import statsmodels.stats.api as sms
```

Una vez cargadas las librerías ya se puede simular la muestra y realizar sobre la misma la prueba de normalidad:

```
# definir semilla para que los resultados sean los mismos--------
np.random.seed(seed = 11)
# se definen los parametros de la distribucion normal------------
mu, sigma = 0.05, 0.9
# se obtiene la muestra estandarizada----------------------------
muestra = np.random.normal(mu, sigma, 1000)
# se realiza la prueba-------------------------------------------
sh_result = stats.shapiro(muestra)
# dar formato a la salida----------------------------------------
print("estadístico: %5.3f, p.valor: %5.3f" %(sh_result.statistic,
sh_result.pvalue))
```

```
## estadístico: 0.999, p.valor: 0.702
```

El resultado sale mayor que 0.05 y por lo tanto se acepta la hipotesis nula de normalidad de la muestra.

b) Como el tamaño de muestra es mayor que 5000, la *Prueba de Shapiro-Wilk* no es tan recomendable. En cambio, se puede realizar las pruebas de normalidad en Python utilizando la función *kstest* de la librería *scipy.stats*. La muestra se obtiene con la función *np.random.normal*, la cual se vió en el tema 1. La *Prueba de Kolmogorov-Smirnov* realmente necesita comparar entre dos distribuciones. Así pues, se compara

la muestra con la distribución normal pasando como argumentos la media y la desviación típica.

```
# definir semilla para que los resultados sean los mismos--------
np.random.seed(seed = 11)
# se definen los parametros de la distribucion normal------------
mu, sigma = 0.05, 0.9
# se obtiene la muestra-------------------------------------------
muestra_2 = np.random.normal(mu, sigma, 10000)
# se realiza la prueba--------------------------------------------
ks_result = stats.kstest(muestra, "norm", args = (mu,sigma))
# dar formato a la salida-----------------------------------------
print("estadístico: %5.3f, p.valor: %5.3f" %(ks_result.statistic,
ks_result.pvalue))
```

```
## estadístico: 0.016, p.valor: 0.962
```

Como el p-valor de la Prueba de Shapiro-Wilk es mucho mayor de 0.05 se considera que la distribución cumple la normalidad de manera estadísticamente significativa. A continuación, en la Figura 6 se muestra un gráfico q-q donde se corrobora visualmente dicha normalidad.

```
# definir dimensiones de la figura-------------------------------
fig = plt.figure(figsize=(5,5))
qqplot = stats.probplot(muestra, dist = "norm", plot = plt)
plt.xlabel("cuantiles teoricos")
plt.ylabel("muestra")
plt.title(" ")
plt.show()
```

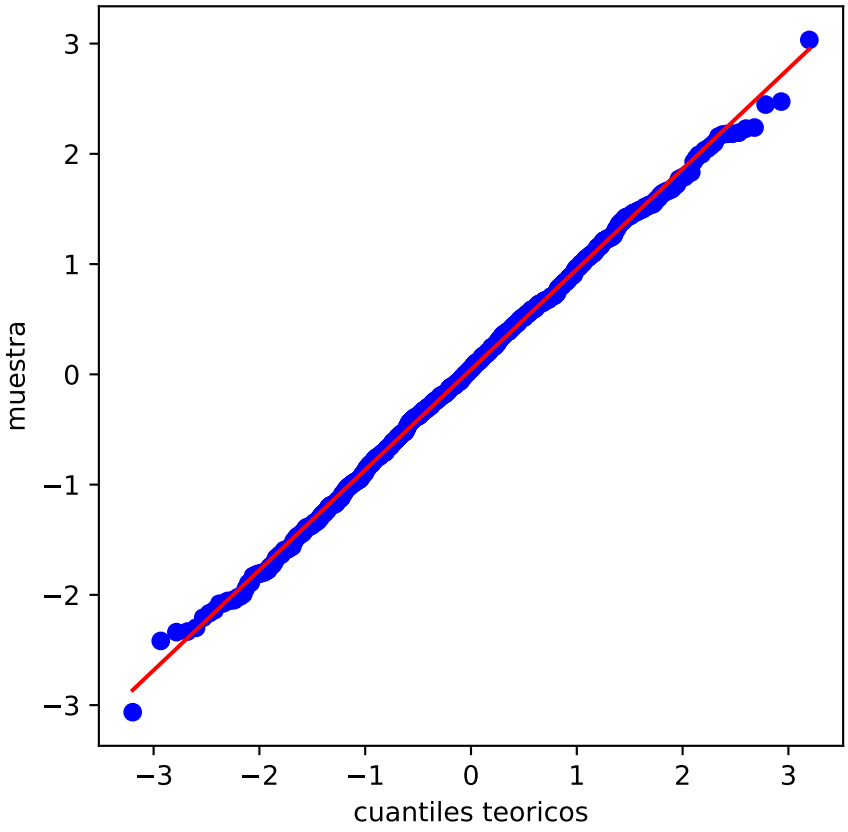


Figura 6: Gráfico q-q de la distribución

Ejercicio 2

Simula:

- Una muestra ε de n = 500 de una distribución normal con media $\mu = 0$ y desviación típica $\sigma = 1$.
- Una muestra x de n = 500 de una distribución uniforme [0, 200].

Construye las muestra $Y1$ e $Y2$ que sean igual a:

$$Y1_i = 100 + 200 \cdot X_i + \varepsilon \cdot X_i \text{ para i = 1, ... , n.}$$

$$Y2_i = 100 + 200 \cdot X_i + \varepsilon \text{ para i = 1, ... , n.}$$

a) Obtén un modelo de regresión lineal $Y = \beta_0 + \beta_1 \cdot X + error$ para Y1 e Y2 y representa gráficamente los residuos de los modelos.

b) Realiza la prueba de Breusch-Pagan para cada uno de los modelos.

Solución

a) Se generan las muestras del ejercicio y se construyen las muestras $Y1$ e $Y2$. Posteriormente se realizan los modelos de regresión y se muestran gráficamente los residuos. Por el modo en que están construidas ambas muestras, $Y1$ tendrá unos residuos crecientes con X mientras que para $Y2$ los residuos serán homogéneos para todo el rango de variación de X.

```
# definir semilla para que los resultados sean los mismos--------
np.random.seed(seed = 11)
# se obtiene la muestra normal-----------------------------------
muestra_normal = np.random.normal(0, 1, 500)
muestra_unif = np.random.randint(0, 200, 500)
# para la muestra y1---------------------------------------------
y1 = 100 + 200 * muestra_unif + muestra_normal * muestra_unif
# Se necesita trasponer el vector x------------------------------
x = muestra_unif.reshape((-1, 1))
lm1 = linear_model.LinearRegression()
model1 = lm1.fit(x, y1)
# obtener valores modelo-----------------------------------------
```

```
y_pred1 = model1.predict(x)
# obtener residuos-----------------------------------------------
residuos1 = y1 - y_pred1
# para la muestra y2---------------------------------------------
y2 = 100 + 200 * muestra_unif + muestra_normal
lm2 = linear_model.LinearRegression()
model2 = lm2.fit(x, y2)
# obtener valores modelo-----------------------------------------
y_pred2 = model2.predict(x)
# obtener residuos-----------------------------------------------
residuos2 = y2 - y_pred2
```

En la Figura 7 se muestran los residuos del ajuste de Y1.

```
plt.scatter(x, residuos1)
plt.axhline(0, color = "red", linewidth = 4)
plt.xlabel("x")
plt.ylabel("residuos")
plt.show()
```

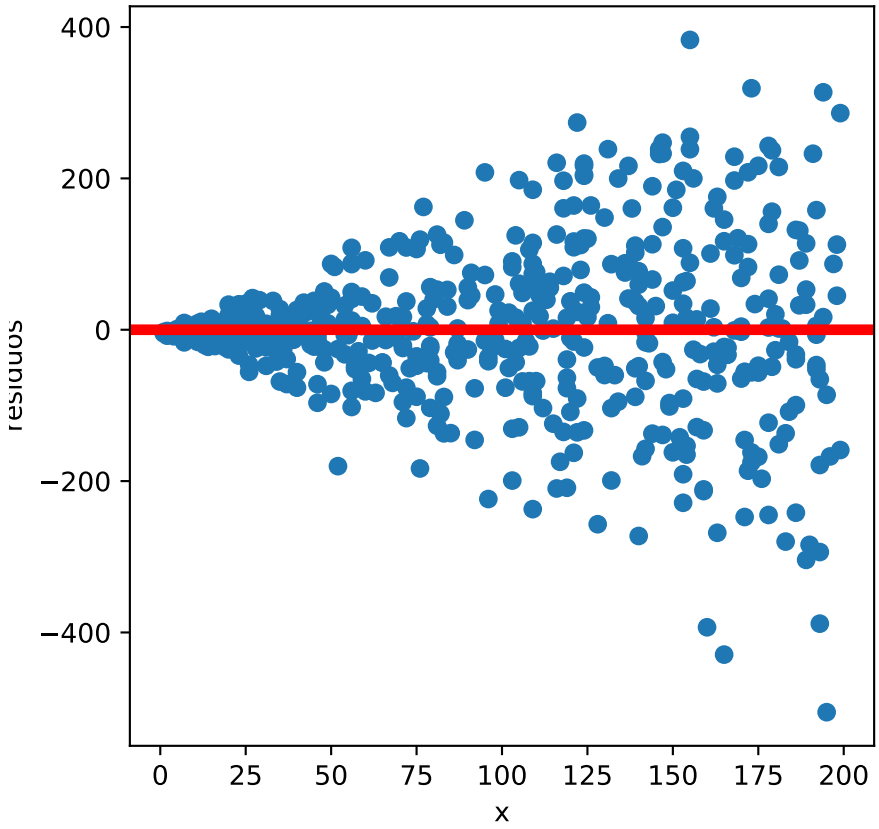


Figura 7: Residuos del modelo Y1

En la Figura 8 se muestran los residuos del ajuste de Y2.

```
plt.scatter(x, residuos2)
plt.axhline(0, color = "red", linewidth = 4)
plt.xlabel("x")
plt.ylabel("residuos")
plt.show()
```

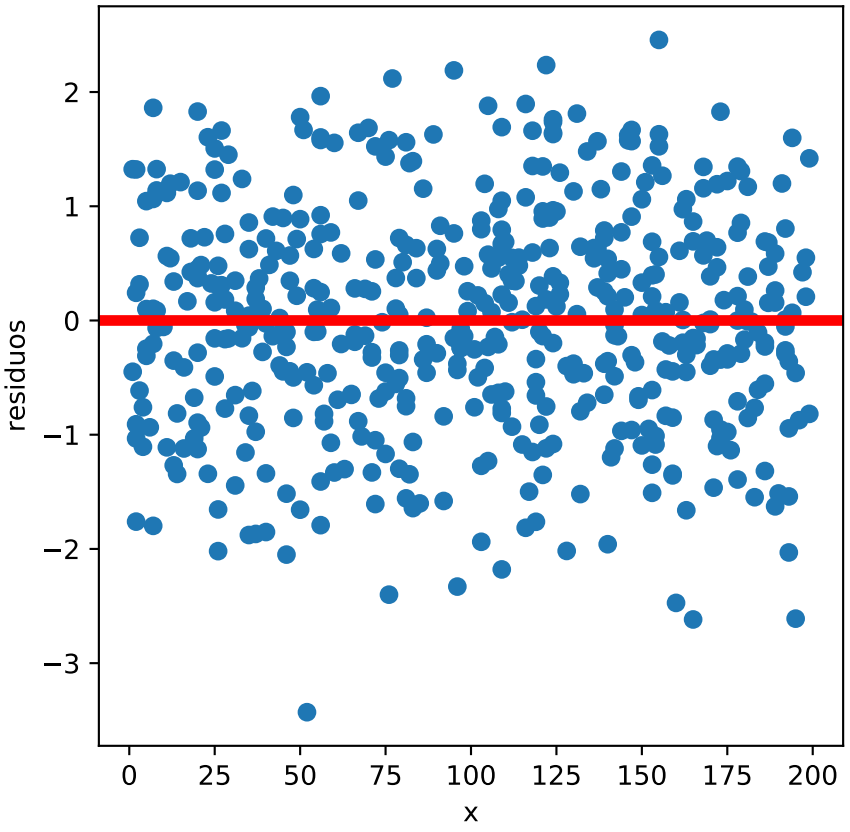


Figura 8: Residuos del modelo Y2

b) Se realiza la prueba de *Breusch-Pagan* para cada uno de los modelos.

```
# y1-------------------------------------------------------------
m1 = sm.OLS(y1, sm.add_constant(x)).fit()
bp1 = sms.het_breuschpagan(resid = m1.resid,
exog_het = m1.model.exog)[1]
print("El resultado del test Breusch-Pagan es: p.valor = %5.3f"
%(bp1))

# y2-------------------------------------------------------------
```

```
## El resultado del test Breusch-Pagan es: p.valor = 0.000
```

```
m2 = sm.OLS(y2, sm.add_constant(x)).fit()
bp2 = sms.het_breuschpagan(resid = m2.resid,
exog_het = m2.model.exog)[1]
print("El resultado del test Breusch-Pagan es: p.valor = %5.3f"
%(bp2))
```

```
## El resultado del test Breusch-Pagan es: p.valor = 0.881
```

Se corrobora, como se preveía, que el modelo de regresión Y1 es heterocedástico, y el modelo de regresión Y2 es homocedástico.

Ejercicio 3

Se tiene la siguiente muestra $X = \{4, 3, 11, 5, 4, 50, 6, 8, 7, 9\}$, en el que el valor 50 es un outlier.

a) Calcula la media recortada de la muestra al 20 %, al 10 % y calcula la centrimedia.

b) Calcula la media winsorizada al 20 %.

c) Calcula la trimedia.

d) Compara los valores obtenidos en los apartados a), b) y c) con la media muestral.

e) Calcula el MAD y comparalo con la desviación típica.

Solución

a) Para calcular las medias recortadas, se eliminan los datos una vez ordenados por cada extremo y se calcula la media. para recortar un 20 % ($\alpha = 0.2$), se recorta un 10 % (un valor) en cada extremo y se calcula la media. Para recortar un 10 % ($\alpha = 0.1$), se pondera los valores de los extremos con 0.5 y se calcula la media considerando que n es igual a 9. Por último, para obtener la centrimedia se eliminan 2 valores de cada extremo, se pondera por 0.5 el tercer y el octavo valor y se obtiene la media considerando $n = 5$.

```
# definir la muestra-----------------------------------------
X = np.array([4, 3, 11, 5, 4, 50, 6, 8, 7, 9])
# ordenar los valores de la muestra-----------------------------
X.sort()
# print(X)
# quitar valores recortados 0.1---------------------------------
X1 = X[1:9]
# print(X1)
print("La media recortada al 20% es", sum(X1)/8)
# quitar valores recortados 0.05--------------------------------
```

```
## La media recortada al 20% es 6.75
```

```
X2 = np.array([3*0.5,  4,  4,  5,  6,  7,  8,  9, 11, 50*0.5])
print("La media recortada al 10% es",round(sum(X2)/9, 2))
# quitar valores recortados 0.25 (centrimedia)------------------
```

```
## La media recortada al 10% es 8.94
```

```
X3 = X[2:8]
# print(X3)
X3[0] = 4*0.5
X3[5] = 8*0.5
```

```
print("La centrimedia es", sum(X3)/5)
```

```
## La centrimedia es 6.4
```

b) Para obtener la media winsorizada, en lugar de eliminar las observaciones, se sustituyen por los últimos valores. Es análogo a las medias recortadas excepto en que las puntuaciones que se eliminaban ahora se sustituyen por los valores menor y mayor que quedan para el cómputo de la media winsorizada.

```
# definir la muestra----------------------------------------
X = np.array([4, 3, 11, 5, 4, 50, 6, 8, 7, 9])
# ordenar los valores de la muestra-----------------------------
X.sort()
# print(X)
# quitar valores recortados 0.1---------------------------------
W1 = np.array([4,  4,  4,  5,  6,  7,  8,  9, 11, 11])
# print(W1)
print("La media winsorizada es", sum(W1)/10)
```

```
## La media winsorizada es 6.9
```

c) La trimedia se obtiene ponderando la mediana y los primer y tercer cuartiles según la fórmula:

$$\text{trimedia} = \frac{\text{cuartil 1} + 2 \cdot \text{mediana} + \text{cuartil 3}}{4}$$

```
# definir la muestra----------------------------------------
X = np.array([4, 3, 11, 5, 4, 50, 6, 8, 7, 9])
# ordenar los valores de la muestra-----------------------------
X.sort()


q1_x = np.quantile(X, 0.25, interpolation='midpoint')
```

```
me_x = np.quantile(X, 0.5, interpolation='midpoint')
q3_x = np.quantile(X, 0.75, interpolation='midpoint')

trimedia = (q1_x + 2 * me_x + q3_x)/4
print("La trimedia es ", trimedia)
```

```
## La trimedia es  6.5
```

d)

```
# definir la muestra-----------------------------------------
X = np.array([4, 3, 11, 5, 4, 50, 6, 8, 7, 9])
# ordenar los valores de la muestra-----------------------------
print("La media muestral es", X.mean())
```

```
## La media muestral es 10.7
```

La media muestral es 10.7, que es mucho más sensible al outlier de 50, que los métodos robustos obtenidos en los apartados a, b y c.

e) Para calcular el MAD es necesario calcular la mediana de la diferencia absoluta de cada valor con la mediana de la muestra.

```
print("La mediana de la muestra vale me_x = ", me_x)
# construir vector diferencias absolutas-----------------------
```

```
## La mediana de la muestra vale me_x =  6.5
```

```
AD = abs(X - me_x)
MAD = np.quantile(AD, 0.5, interpolation = "midpoint")

print("El MAD vale", MAD)
```

```
## El MAD vale 2.5
```

```
print("La desviación típica es", round(X.std(), 1))
```

```
## La desviación típica es 13.3
```

Se observa que el MAD es mucho más insensible a valores atípicos que la desviación típica de la muestra.