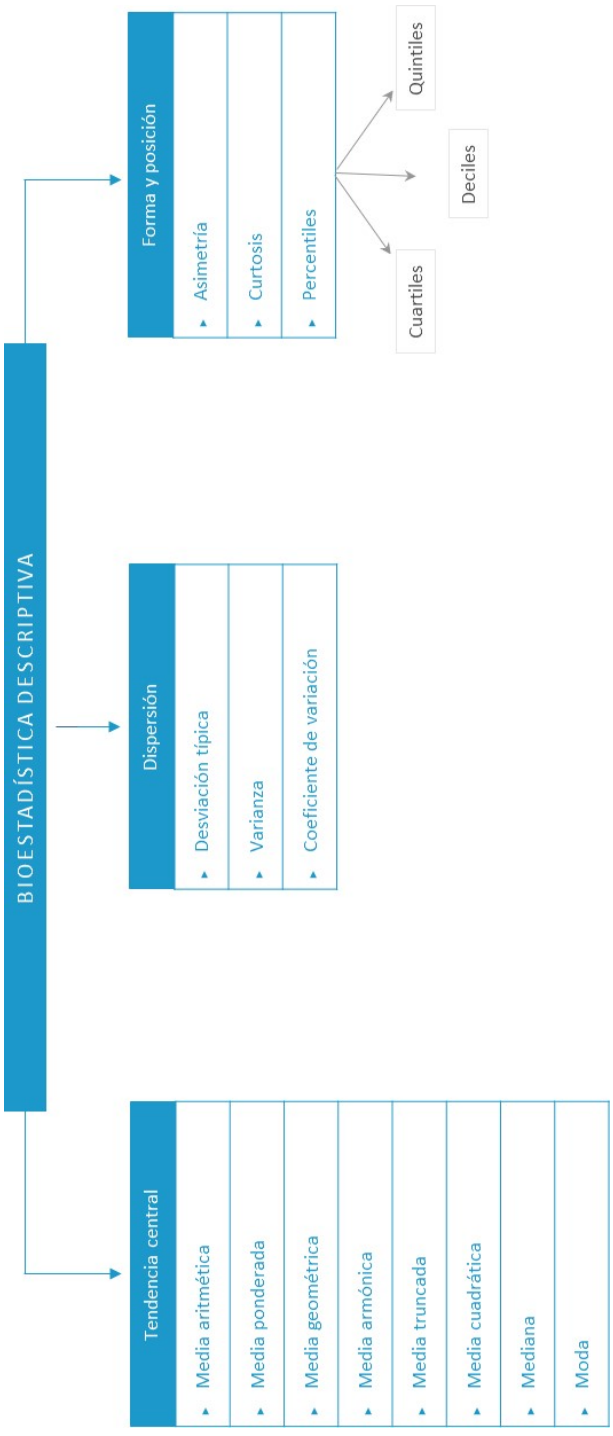
Estadística y R para Ciencias de la Salud

Tema 5. Bioestadística

Índice

Esquema



5.1. Introducción y objetivos

Si echamos la vista atrás, los primeros estudios descriptivos fueron los **censos** de población que se realizan sobre todos los habitantes y miembros. Esta práctica de elaborar censos se remonta a la Edad Antigua y continúa siendo una práctica en la actualidad.

A lo largo de la historia, existen numerosos ejemplos de actividad **estadística** en diferentes culturas, como, por ejemplo, en la antigua Babilonia, Egipto, China y Roma, donde se hacían estudios descriptivos de recuentos de población. Incluso se encuentran referencias a la estadística en **textos sagrados** de varias religiones, como en el libro de los Números de la Biblia.

Otros ejemplos de censos son el realizado en Egipto por Moisés y el empadronamiento llevado a cabo por los romanos en Judea. En Mesoamérica, durante la segunda migración de las tribus chichimecas en el año 1116, el rey Xólotl ordenó un censo de todos sus súbditos, el cual contó con la participación de cada persona que tiró una piedra en un montón llamado Nepohualco, dando como resultado un total de 3 200 000 personas.

A partir del siglo XIX, gracias a aportes de destacados estadísticos como **Adolphe Quetelet** (1796-1874), se desarrollaron distintos **métodos** para calcular probabilidades y analizar datos de diferentes fenómenos.

Por ello, la **estadística descriptiva** es una técnica que se utiliza para **describir** y **resumir cuantitativamente** las características de un conjunto de datos [1]. A diferencia de las estadística inferencial, que se utiliza para aprender sobre una población utilizando datos de muestra, la estadística descriptiva tiene como objetivo resumir y describir una muestra de datos sin basarse en la teoría de la probabilidad [1].

Ideas clave

Aunque la **estadística inferencial** es utilizada, principalmente, en el análisis de datos para **predecir modelos,** la **estadística descriptiva** es esencial para resumir y **presentar información** sobre una muestra, como, por ejemplo, el tamaño muestral, el tamaño de subgrupos importantes y las características demográficas o clínicas de los sujetos, entre otras [2]. Por ejemplo, las tablas que informan sobre sujetos humanos, células o incluso animales suelen incluir estadísticas descriptivas, como la edad, la proporción de sujetos estratificado por sexo, la proporción de sujetos con prevalencia de comorbilidades, entre otras.

Por lo tanto, el objetivo de este tema se centrará en:

- Definir qué son las medidas de tendencia central, cuáles son (media, mediana y moda) y cómo se calculan.
- Describir qué son las medidas de dispersión, cuáles son (rango, la varianza y la desviación estándar) y cómo se calculan e interpretan en el contexto de los datos biomédicos.
- Analizar qué son las medidas de forma y posición, cuáles son (curtosis y la asimetría) y cómo se calculan y pueden utilizarse para caracterizar la forma de la distribución de datos.
- Proporcionar ejemplos de aplicación de estas medidas en problemas biomédicos reales.
- Identificar las limitaciones de la bioestadística descriptiva y explicar por qué es necesario combinarla con otras técnicas estadísticas para hacer inferencias de la toma de decisiones en la investigación biomédica.

5.2. Medidas de tendencia central

Una medida de tendencia central es un valor que representa la **posición central** de una distribución de datos (3,4). La **media,** la **mediana** y la **moda** son las tres medidas de tendencia central más populares. Además, se utilizan medidas de posición, como los **cuantiles,** para describir la **ubicación** de los datos dentro de la distribución sin tener en cuenta su centralidad. El tipo de datos y la distribución de la muestra determinan la **medida** de tendencia central a utilizar. Las medidas de tendencia central se clasifican tal y como aparecen en la Figura 1.

Figura 1. Medidas de tendencia central. Fuente: elaboración propia

Media aritmética

La media aritmética, más comúnmente definida como promedio o media, es una **medida estadística** utilizada para obtener el **valor promedio** de un conjunto de datos [4]. Se obtiene al sumar todos los valores de la muestra y dividirlos por el número total de elementos (Figura 2). Es una medida comúnmente utilizada en diversas disciplinas **académicas,** como las ciencias biomédicas, economía,

antropología, historia, estadística, entre otras.

Aunque la media aritmética es un indicador de tendencia central de uso común, los **valores atípicos** (en inglés se conocen como *outliers* y son aquellos valores extremos o inusuales que se desvían significativamente del promedio de los datos) pueden tener un **impacto negativo** en la media aritmética.

$$media = \frac{x_1 + x_2 + \ldots x_n}{n}$$

Figura 2. Cálculo de la media aritmética. Fuente: elaboración propia

E n **distribuciones asimétricas** (es decir, que tienen más datos en posiciones extremas de la distribución), como, por ejemplo, en un estudio de la concentración de proteína en una muestra de sangre, un valor extremadamente alto o bajo de un paciente en particular puede afectar negativamente la media aritmética de la muestra. Por lo tanto, la media aritmética no es una estadística robusta para este caso. En biología, un ejemplo de una distribución asimétrica podría ser la distribución de tamaños de células en una población. Si una pequeña cantidad de células en la población son significativamente más grandes o pequeñas que el resto, la media aritmética de los tamaños de las células puede no reflejar adecuadamente la tendencia central. En este caso, se utilizarían **otras medidas** de tendencia central, como la mediana, que pueden proporcionar una mejor descripción de la tendencia central de los datos.

```
# Ejemplo 1: Calculamos la media aritmética de datos inventados

# Datos ficticios
datos <- c(10, 15, 20, 30, 35, 40, 50, 55, 60, 80, 85, 100)
mean(datos)

[1] 48.33333
```

Figura 3. Media aritmética de datos inventados. Fuente: elaboración propia.

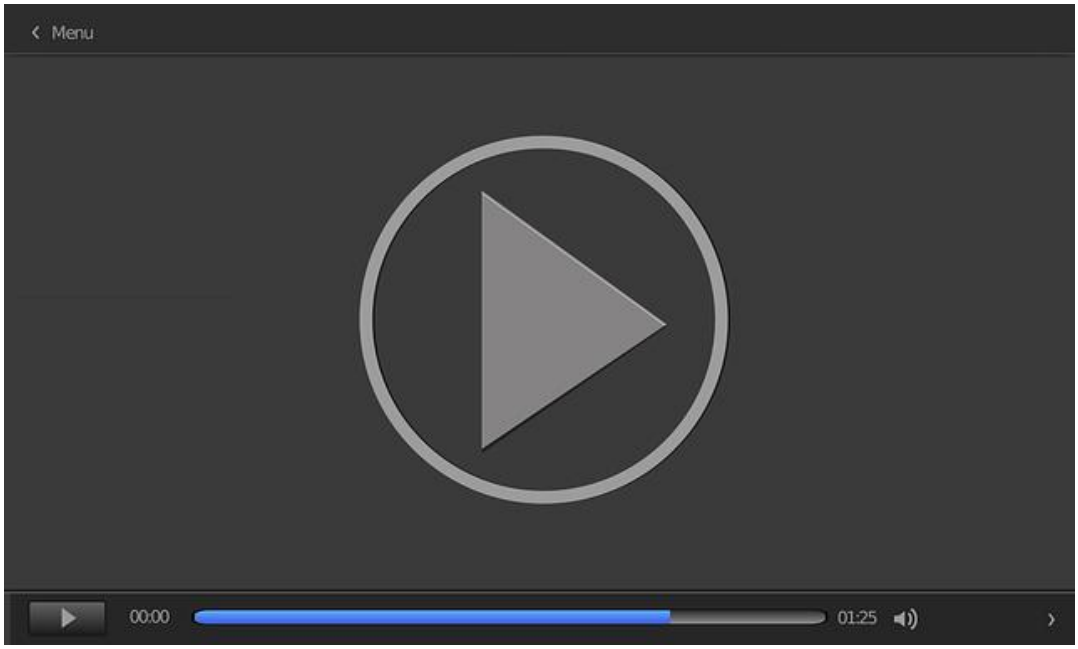
Media ponderada

La media ponderada es una medida útil de tendencia central cuando se quiere dar mayor **importancia** a ciertos datos dentro de un conjunto de datos (5). En esta medida, se asigna un peso a cada dato en función de su **importancia relativa** (Figura 4) y se calcula la media ponderada utilizando la **suma ponderada** de cada dato dividida entre la suma de los pesos.

$$media\ ponderada\ = \frac{x_1w_1 + x_2w_2 + \ldots x_nw_n}{w_1 + w_2 + w_n}$$

Figura 4. Cálculo de la media ponderada. La letra w representa el peso (*weight* en inglés). Fuente: elaboración propia

A continuación, puedes acceder al vídeo *Ponderación de los datos.*



Ponderación de los datos

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=6e5fb504-5b07-426a-8107-b08900db98d6

Un ejemplo de aplicación de la media ponderada en biología podría ser el **cálculo del promedio ponderado** de las concentraciones de diferentes proteínas en una muestra biológica, donde se quiere dar mayor importancia a las proteínas con una actividad biológica más relevante. En este caso, se asignaría un peso a cada proteína en función de su relevancia biológica y se calcularía la media ponderada de las concentraciones de proteínas en la muestra.

```
# Ejemplo 2: Calculamos la media ponderada de datos inventados

# Ejemplo de media ponderada: Datos ficticios y pesos
datos <- c(10, 20, 30, 40, 50)
pesos <- c(1, 2, 3, 2, 1)

# Cálculo de la media ponderada
media_ponderada <- weighted.mean(x = datos, w = pesos)
media_ponderada

[1] 30
```

Figura 5. Ejemplo de media ponderada. Fuente: elaboración propia.

Media geométrica

En matemáticas y estadística la media geométrica es una **métrica** de **tendencia central** (Figura 6). Sobre todo, es útil cuando se **comparan** diferentes aspectos cuyas actuaciones tienen unidades de medida en varios **rangos numéricos** [5]. La media geométrica se usa con frecuencia para analizar los datos de **expresión génica** en biología y biomedicina, es decir, en los estudios de genómica, donde uno tiene como objetivo identificar los genes que se sobreexpresan o subexpresan de manera más notable en diversas condiciones experimentales. Cuando se usa la media geométrica, los valores de varios rangos se **normalizan,** lo que significa que la media geométrica se ve afectada por **cambios** en cualquiera de las propiedades que son proporcionalmente **mayores que cero.** Debido a esto, la media geométrica permite realizar comparaciones precisas y justas entre varios aspectos con valores

que se encuentran dentro de diferentes rangos numéricos.

$$media\ geométrica = \sqrt[n]{x_1 * x_2 * \ldots x_n}$$

Figura 6. Cálculo de la media geométrica. Fuente: elaboración propia

El análisis de datos extensos de secuenciación de ARN (RNA-seq) es un ejemplo de cómo se usa la media geométrica en biología. Estos estudios emplean métodos de **secuenciación** para medir la **expresión génica** en diversas condiciones experimentales. La media geométrica de los valores de expresión se utiliza para **normalizar** los datos y **disminuir** el impacto de las variaciones en el tamaño de la biblioteca de secuenciación al comparar la expresión génica entre varias muestras. Esto hace posible identificar los genes que, en varias configuraciones experimentales, se sobreexpresan o infraexpresan notablemente.

```
# Ejemplo 3: Calculamos la media geométrica de datos inventados

# Crear un conjunto de datos de expresión génica para 3 muestras y 5 genes

expresion_genes <- matrix(c(10, 20, 30, 5, 7, 9, 40, 60, 80, 2, 4, 6, 15, 30, 45), ncol=3)
colnames(expresion_genes) <- c("Muestra1", "Muestra2", "Muestra3")
rownames(expresion_genes) <- paste("Gen", 1:5, sep="")
expresion_genes

      Muestra1 Muestra2 Muestra3
Gen1       10        5       40
Gen2       20        7       60
Gen3       30        9       80
Gen4        5        2       15
Gen5        7        4       30

# Calcular la media geométrica de los valores de expresión
media_geom <- exp(rowMeans(log(expresion_genes)))
media_geom

[1] 30
```

Figura 7. Ejemplo de media geométrica. Fuente: elaboración propia.

Media armónica

La media armónica es una medida de tendencia central utilizada en estadística y matemáticas que se obtiene mediante el **cálculo del inverso** de la **media aritmética** (Figura 8). Es recomendada para **promediar valores inversamente proporcionales** [5]. Se emplea para promediar múltiplos o cocientes, así como para promediar trayectorias de igual longitud o variables, como el tiempo, y que suelen darse en diferentes tiempos o en diferentes medidas.

$$media\ armónica\ = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots \frac{1}{x_n}}$$

Figura 8. Cálculo de la media armónica. Fuente: elaboración propia

Un ejemplo de aplicación de la media armónica en biología podría ser en el análisis de la tasa de crecimiento de las **poblaciones bacterianas.** En este caso, se podría calcular la media armónica de las tasas de crecimiento de varias poblaciones de bacterias para obtener una medida representativa de su velocidad de crecimiento.

```
# Ejemplo 4: Calculamos la media armónica de datos inventados

# Crear un vector con valores numéricos
datos <- c(10, 20, 30, 40, 50)

# Calcular la media armónica
media_arm <- length(datos) / sum(1/datos)
media_arm

[1] 21.89781
```

Figura 9. Media armónica. Fuente: elaboración propia.

Media truncada

La media truncada es una medida de tendencia central utilizada en estadística que se asemeja a la media aritmética y la mediana [5]. En este caso, se **excluyen** los **valores extremos** inferiores y superiores antes de calcular el promedio. Por lo general, se eliminan cantidades iguales de valores en ambos extremos. Esta medida se utiliza cuando se desea r**educir** e l **impacto** de los valores extremos en el promedio y obtener una medida más robusta de la tendencia central de los datos.

Un ejemplo de aplicación de la media truncada en biología puede ser en el análisis de datos de expresión génica en estudios de **transcriptómica.** En estos estudios es común que algunos genes presenten una expresión muy baja o alta, lo que puede afectar la medida de la tendencia central. Al utilizar la media truncada, se pueden excluir estos valores extremos y obtener una medida más r**epresentativa** de la expresión de los genes en una muestra determinada.

```
# Ejemplo 5: Calculamos la media geométrica de datos inventados
# Nos inventamos 20 datos
set.seed(1234)
datos <- rnorm(20)
datos

-1.20706575  0.27742924  1.08444118 -2.34569770  0.42912469  0.50605589 -0.57473996 -
0.54663186 -0.56445200 -0.89003783 -0.47719270 -0.99838644 -0.77625389  0.06445882
0.95949406
-0.11028549 -0.51100951 -0.91119542 -0.83717168  2.41583518

# Calcular la media truncada a 0.1
mean(datos, trim = 0.1)
[1] 30
```

Figura 10. Media truncada. Fuente: elaboración propia.

Media cuadrática

Una medida estadística utilizada para cuantificar la **magnitud** de una variable, que puede tomar valores tanto positivos y negativos como errores, es la raíz cuadrática media, también conocida como valor cuadrático medio [5]. Se utiliza como medida en las ciencias experimentales. La media cuadrática implica **elevar al cuadrado** todas las observaciones, encontrar sus medias aritméticas y sacar la **raíz cuadrada** de esas medias para volver a la unidad de medida original en lugar de tener en cuenta la media aritmética de los valores absolutos de la variable en cuestión.

$$media\ cuadratica\ = \sqrt{\frac{x_1^2 + x_2^2 + \dots x_n^2}{n}}$$

Figura 11. Cálculo de la media armónica. Fuente: elaboración propia

Por ejemplo, en biología la media cuadrática puede ser utilizada para analizar la variabilidad de la intensidad de fluorescencia en células, en estudios de expresión génica, lo que permite una **comparación** más **adecuada** y **precisa** entre diferentes muestras. La desviación estándar es un ejemplo de medida cuadrática. En resumen, la media cuadrática es una herramienta útil para describir la **variabilidad** de una variable que puede tomar valores tanto positivos como negativos.

```
# Ejemplo 6: Calculamos la media cuadrática de datos inventados

# Nos inventamos 20 datos
valores <- c(2, -3, 1, 0, -5, 4, -2, 0, 3)

# Calcular la media cuadrática
sqrt(mean(valores^2))
[1] 2.841874
```

Figura 12. Media cuadrática. Fuente: elaboración propia.

Mediana

La mediana o percentil 50 representa el **valor central** en un conjunto de **datos regulados** [5]. Esta medida es denotada como la variable de posición central. Si el conjunto de datos es **par,** la mediana es calculada como la media aritmética de las dos puntuaciones centrales. La mediana es una medida útil cuando hay **valores extremos** o **datos atípicos** en el conjunto de datos, ya que no es tan sensible a ellos como la media aritmética.

Un ejemplo de aplicación de la mediana en biología es en el análisis de datos de expresión génica de RNA-seq. En estos estudios los valores de expresión génica pueden **variar** ampliamente entre diferentes muestras debido a la presencia de datos atípicos o genes altamente expresados. Para reducir el impacto de estos valores atípicos en la medida de la expresión génica central se utiliza la mediana en lugar de la media aritmética. De esta manera, se puede identificar la expresión génica central que mejor representa el conjunto de datos.

```
# Ejemplo 7: Calculamos la mediana de datos inventados

# Nos inventamos 20 datos
set.seed(1234)
datos <- rnorm(20)

# Calcular la mediana
median(datos)

[1] -0.5288207
```

Figura 13. Mediana. Fuente: elaboración propia.

Moda

La moda es un valor estadístico que representa el valor más **frecuente** o con un mayor número de **apariciones** en un conjunto de datos [5]. En el caso de una variable en la que no se puede tomar ningún valor entre dos **números consecutivos** (variable discreta), la moda se define como el **valor x** donde la probabilidad y frecuencia de sus datos alcanza su máximo. La moda es una medida importante de tendencia central que proporciona información sobre la frecuencia con la que aparece un determinado valor en una muestra de datos.

```
# Ejemplo 8: Calculamos la moda de datos inventados
# Vector de datos
datos <- c(1, 2, 3, 3, 4, 4, 4, 5, 5, 5, 5)

moda <- function(x) {
   return(as.numeric(names(which.max(table(x)))))
}

# Calcular la moda
moda(datos)
[1] 5
```

Figura 14. Moda. Fuente: elaboración propia.

5.3. Medidas de dispersión

En estadística las medidas de dispersión son un **conjunto de estadísticos** que nos permiten conocer la **variabilidad** de los datos de una variable. Mientras que las medidas de tendencia central (como la media, mediana y moda) nos dan una idea de la ubicación de los datos, las medidas de dispersión nos dan una idea de cuánto se **desvían** los datos de esa ubicación [6].

Las medidas de dispersión se clasificarían en:

- Desviación típica o estándar
- Varianza
- Coeficiente de variación

Varianza

La desviación de los valores de una variable aleatoria de su media está determinada por su varianza, que es una medida de dispersión. Se describe como **varían** una serie de **datos** con respecto a la **media** [7]. En la Figura 15 la varianza se muestra como el **valor al cuadrado** de la variable medida. Por ejemplo, la varianza se expresará en gramos por litro al cuadrado si la variable mide la concentración de una proteína en sangre en gramos por litro.

$$varianza = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots (x_n - \mu)^2}{n}$$

Figura 15. Cálculo de la varianza. Fuente: elaboración propia

```
# Ejemplo 9: Calculamos la varianza de datos inventados
# Crear un vector con las alturas de las plantas
altura_plantas <- c(10, 12, 11, 13, 10, 9, 14, 13, 11)

# Calcular la varianza de las alturas
var(altura_plantas)
[1] 2.777778
```

Figura 16. Varianza. Fuente: elaboración propia.

A continuación, veremos un ejemplo ficticio (Figura 17) en el que puede haber dos poblaciones, muestras o variables que pueden tener la misma media, pero diferentes varianzas.

```
# Ejemplo 10: Calculamos la varianza de datos inventados

# Generar dos muestras de poblaciones

set.seed(123)

muestra1 <- rnorm(n = 10000, mean = 10, sd = 2)

muestra2 <- rnorm(n = 10000, mean = 10, sd = 4)

# Calcular las medias y varianzas de ambas muestras

media1 <- mean(muestra1)

media2 <- mean(muestra2)

varianza1 <- var(muestra1)

varianza2 <- var(muestra2)

# Crear un histograma de ambas muestras para comparar visualmente
```

```
library(ggplot2)

datos <- data.frame(

Poblacion = rep(c("Muestra 1", "Muestra 2"), each = length(muestra1)),

Valor = c(muestra1, muestra2))

ggplot(datos, aes(x = Valor, fill = Poblacion)) +

geom_histogram(aes(y = ..density..),

binwidth = 0.5,

colour = "black",

alpha = 0.5,

position = "identity") +

geom_density(alpha = .2) +

 ggtitle("Distribución de datos con diferentes asimetrías") +

xlab("Datos") +

ylab("Densidad") +

geom_vline(aes(xintercept = media1, color = "Media"),

linetype = "dashed",

size = 1) +

geom_vline(aes(xintercept = media2, color = "Media"),

linetype = "dashed",
```

```
size = 1) +

scale_color_manual(values = c("red", "black")) +

scale_fill_manual(values = c("blue", "green")) +

theme()
```

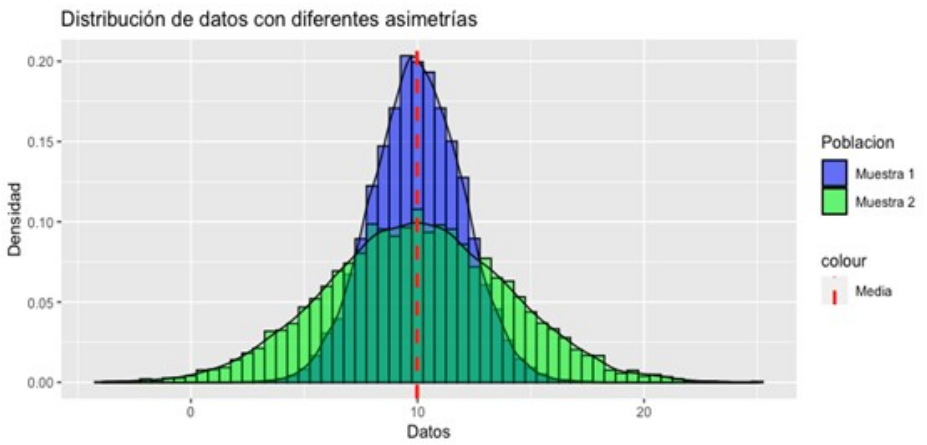


Figura 17. Distribución de las 2 poblaciones con media de 10 (línea discontinua) pero con diferentes varianzas. Fuente: elaboración propia con el programa R Studio.

Desviación típica o estándar

La desviación estándar es una medida de dispersión que se utiliza comúnmente [8]. Se interpreta como la **raíz cuadrada positiva** de la varianza y se expresa en las mismas unidades que los datos de la variable (Figura 18) [8]. Por lo tanto, si la variable mide la concentración de una proteína en sangre en gramos por litro, la desviación estándar se expresará en gramos por litro.

$$desviación\ estándar = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ \ldots (x_n - \mu)^2}{n}}$$

Figura 18. Cálculo de la desviación típica y estándar. Fuente: elaboración propia

```
# Ejemplo 11: Calculamos la desviación estándar de datos inventados
# Generar una muestra aleatoria de 50 datos con media 10 y DE= 2
set.seed(123)
muestra <- rnorm(n = 50, mean = 10, sd = 2)

# Calcular la media, varianza y desviación estándar de la muestra
mean(muestra)
[1] 10.06881

var(muestra)
[1] 3.428941

sd(muestra)
[1] 1.85174
```

Figura 19. Desviación estándar. Fuente: elaboración propia.

Coeficiente de variación

El coeficiente de variación se define como una **medida normalizada** de la dispersión de una distribución de datos expresada como un porcentaje [9]. Para obtener esta medida se calcula la **relación** entre la **desviación estándar** y la **media** de los datos (Figura 20). El coeficiente de variación resulta útil al comparar la **variabilidad** entre dos o más conjuntos de datos que pueden tener distintas unidades de medida o escalas de magnitud. En el ámbito biológico se aplica el coeficiente de variación al investigar la variabilidad de la frecuencia cardíaca según el sexo y la edad de los pacientes. Esto permite obtener información sobre la **dispersión** de los **datos** y posibles **diferencias** en la variabilidad en función de esas variables.

$$coeficiente\ de\ variación = \frac{desviación\ estándar}{|media|}$$

Figura 20. Cálculo del coeficiente de variación. Fuente: elaboración propia

```
# Ejemplo 12: Calculamos el coeficiente de variación de datos inventados
# Generar una muestra aleatoria de 100 datos con media 50 y desviación estándar 10
set.seed(123)
datos <- rnorm(n = 100, mean = 50, sd = 10)

# Calcular el coeficiente de variación de la muestra
cv <- sd(datos) / mean(datos) * 100  # como un porcentaje
cv

[1] 17.93208
```

Figura 21. Coeficiente de variación. Fuente: elaboración propia.

5.4. Medidas de forma y posición

Dentro del ámbito estadístico existen medidas de forma que permiten describir la **configuración** de una **distribución de probabilidad** sin necesidad de recurrir a representaciones gráficas [10]. Estas medidas se clasifican en dos categorías principales: **asimetría** y **curtosis.**

La asimetría se emplea para evaluar el grado de simetría presente en una distribución, mientras que la curtosis cuantifica el nivel de concentración que exhibe la distribución alrededor de su valor central.

Un ejemplo de aplicación de estas medidas en biomedicina podría ser el estudio de la distribución de la talla en un grupo de pacientes. La asimetría podría indicar si la mayoría de los pacientes tienen una talla similar o si hay una mayor cantidad de pacientes con una talla muy por encima o por debajo de la media, mientras que la curtosis podría indicar si la mayoría de los pacientes tienen una talla cercana a la media o si hay una dispersión más amplia de las tallas.

Asimetría

Como se ha mencionado anteriormente, la simetría se traduce como el grado de distribución perfecto de una distribución de datos [6] Principalmente, se distinguen **tres tipos** de simetría (Figura 23): **asimetría positiva, simetría** y **simetría negativa.** Para ello, primero calculamos la simetría de nuestros datos usando la librería Moments [11].

```
# Ejemplo 13: Calculamos la simetría de datos inventados
# Generar datos biomédicos inventados
set.seed(111)
datos_asimetria <- rnorm(n = 10000, mean = 10, sd = 5)
datos_asimetria_pos <- rchisq(n = 10000, df = 5)
datos_asimetria_neg <- -rchisq(n = 10000, df = 5)

# Calcular la asimetría y la curtosis de los datos
install.packages("moments")
library(moments)
skewness(datos_asimetria)
[1] -0.01843534
skewness(datos_asimetria_pos)
[1] 1.2509
skewness(datos_asimetria_neg)
[1] -1.26402
```

Figura 22. Simetría. Fuente: elaboración propia.

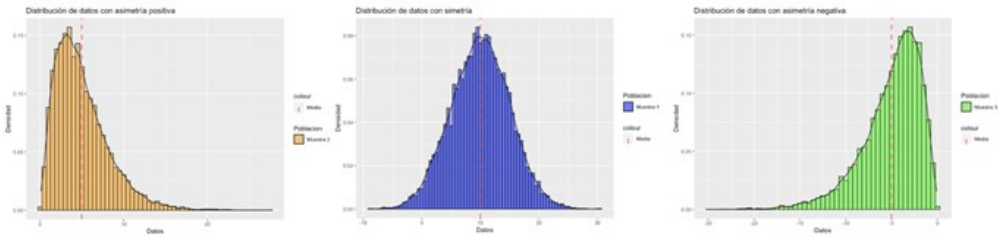


Figura 23. Distribución de datos según su simetría. Fuente: elaboración propia

- En el caso de que el **coeficiente** de asimetría sea **positivo** se considera que la distribución presenta **asimetría positiva.**
- Si el coeficiente de asimetría es **igual a cero** se interpreta que la distribución es **simétrica.**
- Por último, si el coeficiente de asimetría resulta **negativo** se concluye que la distribución exhibe **asimetría negativa.**

Curtosis

La curtosis, conocida también como apuntamiento, es un parámetro estadístico que permite evaluar el grado de **concentración** de una distribución de datos en torno a su media (Figura 24). La curtosis determina si una distribución es más puntiaguda o aplanada. Una distribución con una **curtosis elevada** es más **puntiaguda** y **concentrada** (leptocúrtica), mientras que una distribución con una **curtosis baja** es más **aplanada** y **dispersa** [5]. Por ejemplo, en el campo de la biología o la medicina la curtosis se puede utilizar para evaluar la distribución de ciertos valores en una muestra de pacientes, como el índice de masa corporal o la concentración de ciertos componentes en la sangre.

```
# Ejemplo 14: Calculamos la curtosis de datos inventados
# Generar datos biomédicos inventados
set.seed(123)
datos_mesocurticos <- rnorm(n = 1000, mean = 10, sd = 2)
datos_leptocurticos <- rchisq(n = 1000, df = 2)
datos_platicurticos <- runif(n = 1000, min = 0, max = 20)

# Calcular la curtosis de los datos
library(moments)
kurtosis(datos_mesocurticos)
[1] 2.925747
kurtosis(datos_leptocurticos)
[1] 8.890374
kurtosis(datos_platicurticos)
[1] 1.893801
```

Figura 24. Curtosis. Fuente: elaboración propia

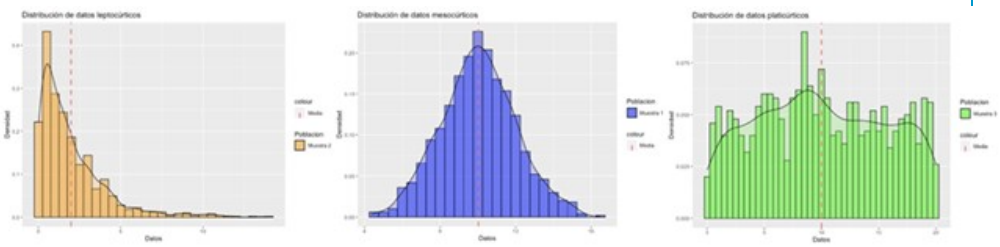


Figura 25. Distribución de datos según su curtosis. Fuente: elaboración propia

Ideas clave

Existen dos tipos de medidas de posición en estadística: las **centrales** y las **no centrales.** Las medidas de posición centrales, como la media y la mediana, se concentran en el centro de la **distribución.** Por otro lado, las medidas de posición no centrales, como los cuantiles, dividen la distribución en **secciones iguales** y reflejan los valores superiores, medios e inferiores de los datos. Por ejemplo, se puede utilizar los cuartiles para dividir una muestra en cuatro partes iguales, lo que puede ser útil para analizar datos biomédicos, como la distribución del tiempo que tarda una muestra de pacientes en recuperarse de una enfermedad.

Las medidas de **posición** más **habituales** son:

- Cuartiles.
- Quintiles.
- Deciles.
- Percentiles.

Cuartiles

Los cuartiles se utilizan para dividir los datos en **cuatro partes iguales,** representando cada parte un 25 % del total. El **primer cuartil** (Q1) corresponde al valor que separa el conjunto de datos en dos partes iguales, donde el 25 % de los valores se sitúan por debajo de Q1 y el 75 % restante se encuentran por encima de este valor. El **segundo cuartil** (Q2), que también se conoce como la mediana, divide los datos en dos partes iguales, con un 50 % de los valores por debajo de Q2 y el otro 50 % por encima. Por último, el **tercer cuartil** (Q3) separa los datos en dos partes iguales, donde el 75 % de los valores se encuentran por debajo de Q3 y el 25 % restante se sitúa por encima de este valor.

Quintiles, deciles y percentiles

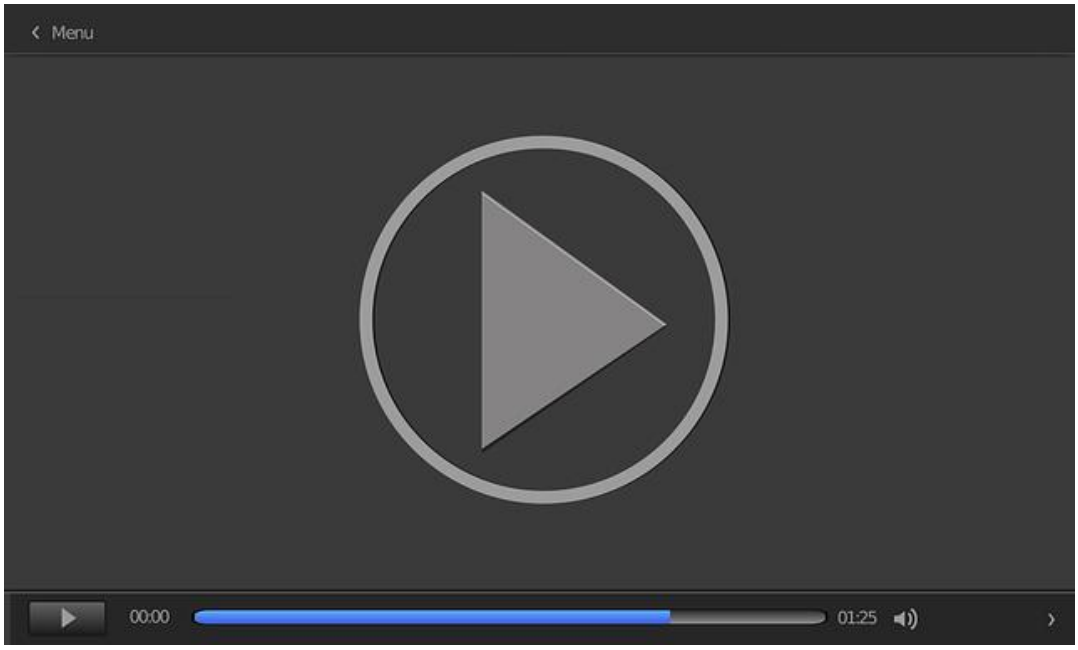
Los quintiles, deciles y percentiles son medidas utilizadas para dividir los datos en partes iguales, correspondiendo a 5%, 10% y 1% respectivamente. Por ejemplo, el percentil 20 representa el valor por debajo del cual se sitúa el 20% de los datos, mientras que por encima de ese valor se encuentran el 80% restante de los datos. De manera similar, el decil 3 indica el valor por debajo del cual se sitúa el 30% de los datos, mientras que por encima de él se encuentran el 70% de los datos.

```
# Ejemplo 15: Calculamos los percentiles de datos inventados
# Generamos datos aleatorios de niveles de glucosa en sangre de 100 pacientes
glucosa <- runif(100, 70, 200)

# Calculamos el percentil 10
quantile(glucosa, 0.1)

# Calculamos el percentil 90
quantile(glucosa, 0.9)
```

Figura 26. Percentiles. Fuente: elaboración propia.

Ideas clave

A continuación, puedes acceder al vídeo *Transformación de bases de datos (narrow a wide).*



ransformación de bases de datos (narrow a wide)

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=d3475942-c994-4671-aa52-b08900db98bb

5.5. Referencias bibliográficas

1. Gravetter FJ, Wallnau LB. Fundamentos de la estadística para las ciencias del comportamiento. Wadsworth; 2013.

2. Larson MG. Descriptive Statistics and Graphical Displays. Circ. 2006 jul. 4; 114(1):76–81.

3. Khorana A, Pareek A, Ollivier M, Madjarova SJ, Kunze KN, Nwachukwu BU, et al. Choosing the appropriate measure of central tendency: mean, median, or mode? Knee Surg Sports Traumatol Arthrosc. 2023 en.; 31(1):12–5.

4.Hazra A, Gogtay N. Biostatistics series module 1: Basics of biostatistics. Indian J Dermatol. 2016; 61(1):10.

5. Torres MF. Estadística. Décima edición. Pearson; 2017.

6. McQuarrie DA. Statistical Mechanics. Nueva York: Harper & Row; 1976.

7.Fisher RA. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Earth Environ. Sci. Trans. R. Soc. Edinb. 2012 jul. 6; 52(2): 399-433.

8. Bland JM, Altman DG. Statistics notes: Measurement error. BMJ. 1996 jun. 29; 312(7047):1654–1654.

9.Koopmans LH, Owen DB y Rosenblatt JI. Confidence intervals for the coefficient of variation for the normal and log normal distributions. Biometrika. 1964 jun.; 51(1–2): 25–32.

10.Hyndman RJ, Fan Y. Sample Quantiles in Statistical Packages. Am Stat. 1996 nov.; 50(4): 361.

11. Komsta L, Novomestky F. Cran [Internet]. moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests (citado 2023 sept. 12]. Disponible en: https://cran.r-project.org/web/packages/moments/index.html

Ejemplos de ejercicios

Hernández-Barajas F. Manual de R. Cap. 10, Medidas de tendencia central. Disponible en: https://fhernanb.github.io/Manual-de-R/central.html

Web en la que podrás profundizar no solo en temas de programación de R, sino también en ejercicios de medidas de tendencia central, variabilidad y de posición. Los ejercicios están resueltos para que puedas obtener la solución de forma más sencilla y fácil.

Medias de dispersión

López JF. Economipedia [Internet]. 2019 sept. 27. Medidas de dispersión [citado 2023 sept. 12]. Disponible en: https://economipedia.com/definiciones/medidas-de-dispersion.html

Web en la que podrás profundizar en el conocimiento de las medidas de dispersión más utilizadas.

Medias de posición

Rus-Arias E. Economipedia [Internet]. 2021 febr. 1. Medidas de posición [citado 2023 sept. 12]. Disponible en: https://economipedia.com/definiciones/medidas-de-posicion.html

Web en la que podrás profundizar en el conocimiento de las medidas de posición más utilizadas

Test

1. ¿Cuál es la medida de tendencia central que se ve más afectada por valores extremos?

A. Media aritmética.

B. Mediana.

C. Moda.

D. Media geométrica.

2. ¿Qué medida de tendencia central se utiliza para datos categóricos?

A. Media aritmética.

B. Mediana.

C. Moda.

D. Media geométrica.

3. ¿Qué medida de dispersión mide la diferencia entre el valor más alto y el valor más bajo de un conjunto de datos?

A. Rango.

B. Desviación estándar.

C. Varianza.

D. Coeficiente de variación.

4. ¿Cuál es la medida de dispersión que se utiliza para conocer la dispersión relativa de los datos?

A. Rango.

B. Desviación estándar.

C. Varianza.

D. Coeficiente de variación.

5. ¿Qué medida de forma y posición indica la cantidad de datos que se encuentran por encima o por debajo de un determinado valor?

A. Percentil.

B. Cuartil.

C. Decil.

D. Mediana.

6. ¿Cuál es la medida de forma y posición que divide el conjunto de datos en cuatro partes iguales?

A. Percentil.

B. Cuartil.

C. Decil.

D. Mediana.

7. ¿Qué función se utiliza en R para calcular la media aritmética?

A. median()

B. sd()

C. var()

D. mean()

8. ¿Qué función se utiliza en RStudio para calcular la mediana?

A. median()

B. sd()

C. var()

D. mean()

9. ¿Qué función se utiliza en R para calcular la desviación estándar?

A. median()

B. sd()

C. var()

D. mean()

10. ¿Qué función se utiliza en R para calcular el rango intercuartílico?

A. IQR()

B. sd()

C. var()

D. mean()