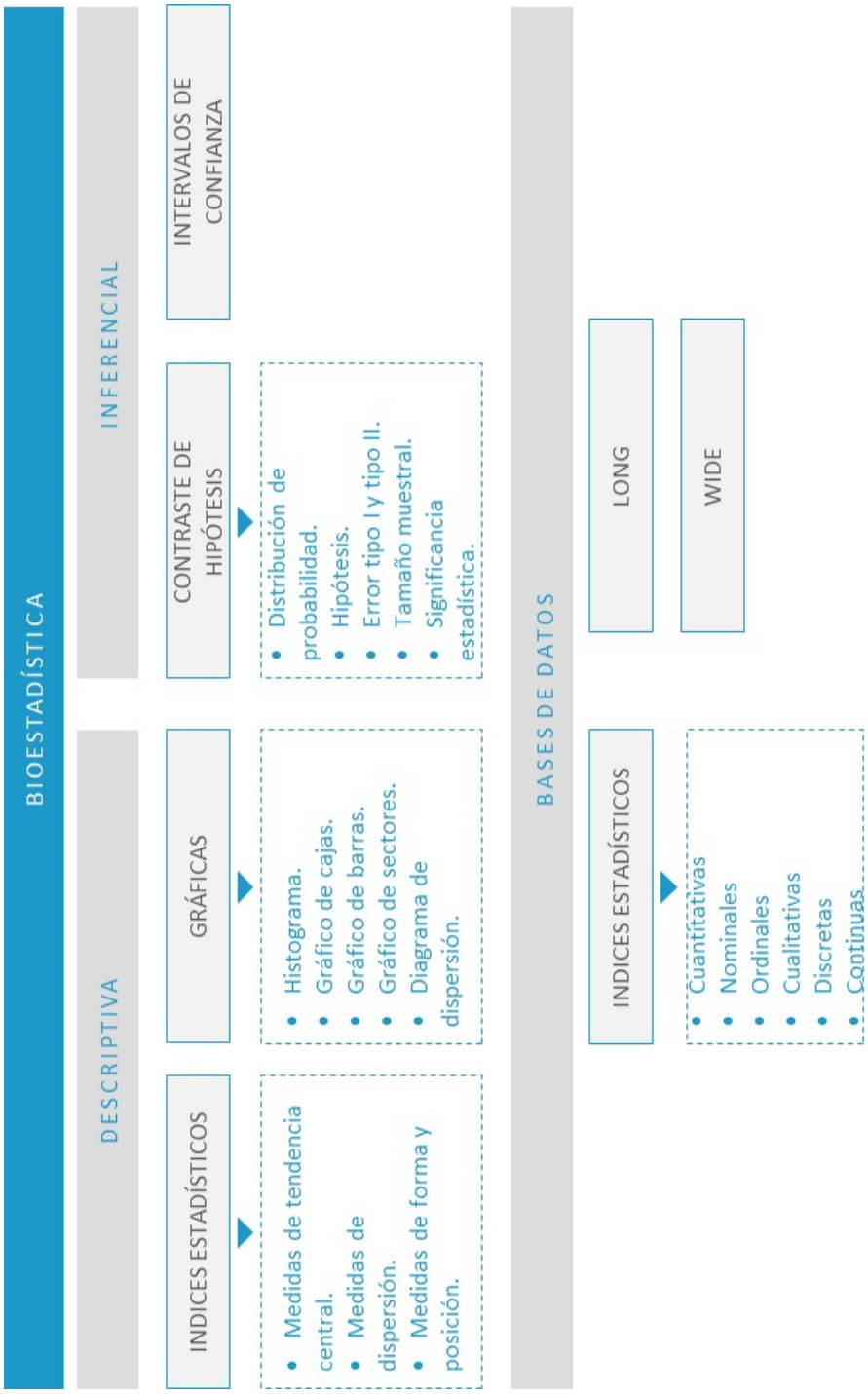
Estadística y R para Ciencias de la Salud

Tema 4. Fundamentos de la bioestadística

Índice

Esquema



© Universidad Internacional de La Rioja (UNIR)

4.1. Introducción y objetivos

La bioestadística es una rama de la estadística que se aplica en el campo de las ciencias, como la biología, química, física, nutrición, farmacia, bioquímica, medicina, enfermería, entre otras, para analizar y comprender los datos obtenidos en estudios y experimentos. La bioestadística es fundamental en la investigación biomédica, ya que permite tomar decisiones y hacer conclusiones basadas en la evidencia numérica [1]. Con una adecuada aplicación de la bioestadística descriptiva e inferencial, se pueden obtener resultados confiables y relevantes para la toma de decisiones clínicas y en la elaboración de políticas de salud pública [1].

Dentro de la bioestadística, diferenciamos la **bioestadística descriptiva** y la **bioestadística inferencial:**

La **bioestadística descriptiva** se enfoca en **describir** y **resumir** los datos mediante medidas de **tendencia central, dispersión** y **forma,** así como en la representación gráfica de los mismos [1]. Por su parte, la **bioestadística inferencial** utiliza técnicas estadísticas para hacer **inferencias** acerca de la población a partir de una **muestra** de datos.

En la bioestadística descriptiva se utilizan **herramientas,** como tablas de frecuencia, gráficos de barras, histogramas, diagramas de cajas y bigotes, entre otros, para **analizar** los datos y obtener información acerca de la **distribución, dispersión** y **forma**.

© Universidad Internacional de La Rioja (UNIR)

Ideas clave

En la bioestadística inferencial, sin embargo, se utilizan técnicas de estimación de **parámetros,** pruebas de **hipótesis** y **análisis de regresión** para hacer inferencias acerca de la población a partir de una muestra de datos [2]. Estas técnicas permiten determinar si las diferencias observadas entre grupos son **significativas** o no, lo que puede ser de gran importancia en la toma de **decisiones clínicas** y en la investigación biomédica en general.

El objetivo de este tema se centrará en:

- Introducir los conceptos básicos de la bioestadística descriptiva e inferencial y su importancia en la investigación biomédica.
- Adquirir habilidades para la selección y aplicación adecuada de herramientas estadísticas en la descripción y análisis de datos biomédicos.
- Identificar los conceptos clave de la bioestadística descriptiva, como saber cuáles son las medidas de tendencia central, dispersión y forma y cómo se aplican para describir y resumir los datos.
- Identificar los conceptos clave de la bioestadística inferencial, como la estimación de parámetros, pruebas de hipótesis y análisis de regresión y cómo se aplican para hacer inferencias acerca de la población a partir de una muestra de datos.
- Conocer los diferentes tipos de bases de datos, tipos de variables y categorización de datos utilizados en la investigación biomédica y cómo influyen en el análisis estadístico.

© Universidad Internacional de La Rioja (UNIR)

4.2. Conceptos clave de la bioestadística descriptiva

Por lo general, la bioestadística se enfoca, en un primera instancia, en el estudio de u n a **muestra** de individuos. Esta muestra o subconjunto comparte ciertas **características** que son de interés para el estudio en cuestión. La muestra se selecciona cuidadosamente para garantizar que sea **representativa** de la población total y que los resultados obtenidos de ella puedan, en la mayoría de los casos, generalizarse con precisión a la población completa [3]. La elección de una muestra adecuada es crucial en el análisis estadístico, ya que una muestra inadecuada puede llevar a conclusiones erróneas o **sesgadas.**

En la investigación científica, se asume que la población en su totalidad es **inaccesible** desde una perspectiva práctica, ya que incluiría a las personas que han vivido, viven y vivirán en el futuro. Por ello, solo se puede describir y estudiar una **muestra pequeña** (en comparación con la población total) y se busca extraer conclusiones válidas para toda la población a partir de ella.

Este proceso se conoce como **muestreo** y es una parte importante de la bioestadística [3]. La **interpretación** del tratamiento estadístico de los datos que se generaliza para toda la población se llama **inferencia.** Por lo tanto, se utilizan diversas **técnicas** de muestreo y análisis estadístico para garantizar la precisión y validez de los resultados obtenidos (Figura 1).

© Universidad Internacional de La Rioja (UNIR)

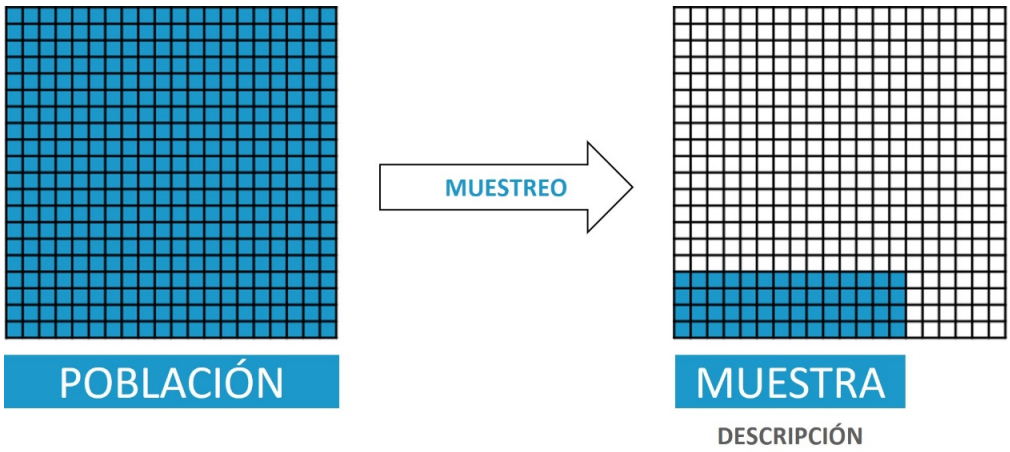


Figura 1. Proceso del muestreo. Fuente: elaboración propia

La primera etapa después del muestreo es la **bioestadística descriptiva.** La bioestadística descriptiva **sintetiza** y **resume** unos datos y los transforma en **información útil** y fácilmente interpretable [1]. Sirve para recoger, clasificar, representar y resumir datos de nuestra población de estudio [4]. La bioestadística descriptiva utiliza dos tipos de **procedimientos:** el **cálculo de índices estadísticos** y el **uso de representaciones gráficas** (Figura 2).

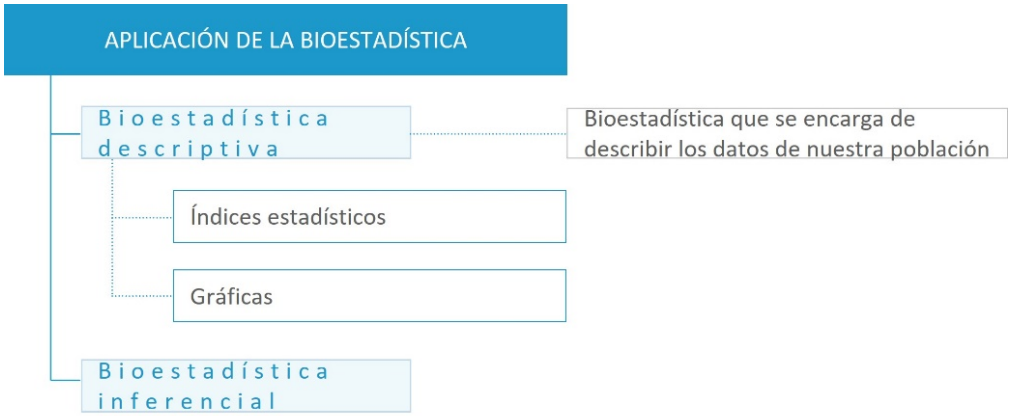


Figura 2. Aplicación de la bioestadística descriptiva. Fuente: elaboración propia

© Universidad Internacional de La Rioja (UNIR)

Cálculo de índices estadísticos

El cálculo de índices estadísticos son números que **resumen** de modo sencillo la **información** contenida en datos reales. Esto se refiere a la aplicación de **fórmulas matemáticas** y **estadísticas** para obtener medidas numéricas que resuman, de manera sencilla, la información contenida en un conjunto de datos reales. Estos índices permiten **describir** y **analizar** las características de una **población** o **muestra** de interés. Algunos ejemplos de índices estadísticos utilizados en bioestadística descriptiva son las medidas de **tendencia central** y las medidas de **dispersión** (Tabla 1).

Medidas de tendencia central (4)	Medidas de dispersión (5)	Medidas de forma y posición (6)
Media	Rango	Cuantiles
Mediana	Desviación estándar	Percentiles
Moda	Varianza	
	Valores mínimos y máximos	
	Curtosis	
	Asimetría	

Tabla 1. Índices estadísticos usados en la bioestadística descriptiva. Fuente: elaboración propia

Para entender la aplicación de la bioestadística descriptiva en la descripción de los datos de una muestra, se representa en la Tabla 2 un **ejemplo heurístico** de las medidas de tendencia central, dispersión, forma y posición de unos datos de **expresión génica** de unas proteínas (tamaño muestral [N] = 20).

© Universidad Internacional de La Rioja (UNIR)

Datos ΔCt de expresión génica para cada uno de los 20 genes

CCL2-Hs00234140 = -0,065; CCL5-Hs99999048 = -0,205; CCR5-Hs99999149 = 0,608; CD274-Hs00204257 = 1,006; CD36-Hs00354519 = -0,156; CHKA-Hs00957878 = 0,504; CPT1A-Hs00912671 = 0,688; CSF2-Hs00929873 = 0,33; CXCR1-Hs00174146 = 0,929; FASN-Hs01005622 = 0,714; FOXO3-Hs00818121 = -0,821; FOXP3-Hs01085834 = -0,023; G6PD-Hs00166169 = -0,096; GPX1-Hs00829989 = 0,439; IL10-Hs00961622 = -1,711; IL1B-Hs01555410 = -0,167; IL6-Hs00174131 = 2,948; IRS1-Hs00178563 = 0,929; JAK1-Hs01026983 = -0,113; STAT3-Hs00374280 = 0,577

Medidas	Resultado	Función en RStudio	Librería
Frecuencia	20	table()	base
Media (aritmética)	0,316	mean()	base
Mediana	0,384	median()	base
Moda	0,929	mode()	DescTools (7)
Rango	4,659	range()	base
Desviación estándar	0,900	sd()	base
Varianza	0,770	var()	base
Valores min y máx	-1,711 a 2,948	min() max()	base
Curtosis	3,873	kurtosis()	moments (8)
Asimetría	0,705	skewness()	moments (8)
Cuantiles	2: ≤-0,124 y > -0,125	quantile()	base
Percentiles	50: 0,384	quantile()	base

Tabla 2. Ejemplo de la aplicación de los índices estadísticos usados en la bioestadística descriptiva y aplicados a unos datos de expresión génica. Existen más librerías para sacar los siguientes datos mostrados. Fuente: elaboración propia

© Universidad Internacional de La Rioja (UNIR)

Ideas clave

Estos índices estadísticos permiten resumir de manera sencilla la información contenida en un conjunto de datos, lo que facilita su interpretación y análisis.

Representaciones gráficas

Las representaciones gráficas son herramientas importantes en la bioestadística descriptiva, ya que permiten **visualizar** y **resumir** grandes cantidades de datos de manera clara y concisa. Algunas de las representaciones gráficas más comunes en bioestadística descriptiva incluyen el histograma, gráfico de barras, gráfico de cajas, gráfico circular, diagrama de dispersión.

- **Histogramas:** un histograma es una representación gráfica de la distribución de frecuencia de un conjunto de datos numéricos continuos [9]. En un histograma, los datos se dividen en **intervalos** de igual ancho (llamados clases o bins) y se representa la frecuencia de datos que caen en cada intervalo con un **rectángulo vertical.** Los histogramas son útiles para identificar patrones de **distribución,** como simetría, sesgo o multimodalidad.

En el entorno RStudio, que usa el lenguaje de programación R, se utiliza el comando `hist()` que viene en la librería base. Sin embargo, caben muchas opciones, incluidas librerías y paquetes específicos, para **personalizar** un histograma usando R.

- **Gráficos de caja:** un gráfico de caja (también llamado diagrama de caja y bigotes) es una representación gráfica de la **distribución** de un **conjunto** de datos numéricos [10]. En un gráfico de caja, se dibuja un rectángulo que representa el **rango intercuartílico** (IQR) de los datos (es decir, el rango del 25 % al 75 % de los datos) y se dibujan líneas (los bigotes) que se extienden hasta los valores más extremos de los datos dentro de un cierto rango. Los gráficos de caja son útiles para **comparar** varias distribuciones de datos.

© Universidad Internacional de La Rioja (UNIR)

Ideas clave

En el entorno RStudio, que usa el lenguaje de programación R, se utiliza el comando `boxplot()` que viene en la librería base. Sin embargo, caben muchas opciones, incluidas librerías y paquetes específicos, para personalizar un diagrama de cajas usando R.

- **Gráfico de barras:** este tipo de gráfico se utiliza para representar **variables categóricas** [11]. El eje x representa las categorías y el eje y representa la frecuencia con la que aparecen esas categorías. Las barras se dibujan para cada categoría y su altura representa la **frecuencia.**

En el entorno RStudio, que usa el lenguaje de programación R, se utiliza la función `barplot()` que viene en la librería base. Sin embargo, caben muchas opciones, incluidas librerías y paquetes específicos, para personalizar un gráfico de barras usando R.

- **Gráfico de sectores:** Este tipo de gráfico se utiliza para representar variables categóricas. La variable se divide en **sectores,** cada uno de los cuales representa una categoría. El tamaño de cada sector es proporcional a la **frecuencia** de la categoría.

En el entorno RStudio, que usa el lenguaje de programación R, se utiliza el comando `pie()` que viene en la librería base. Sin embargo, caben muchas opciones, incluidas librerías y paquetes específicos, para personalizar un gráfico de sectores usando R.

- **Diagrama de dispersión:** un gráfico de dispersión se utiliza para representar la relación entre **dos variables continuas** [12]. Cada punto en el gráfico representa una observación. Si los puntos se agrupan cerca de una línea recta, hay una **correlación lineal** entre las variables.

© Universidad Internacional de La Rioja (UNIR)

Ideas clave

En el entorno RStudio, que usa el lenguaje de programación R, se utiliza el comando `plot()` que viene en la librería base. Sin embargo, caben muchas opciones, incluidas librerías y paquetes específicos, para personalizar un diagrama de dispersión usando R.

© Universidad Internacional de La Rioja (UNIR)

4.3. Conceptos clave de la bioestadística inferencial

La **bioestadística analítica** o inferencial va más allá de la estadística descriptiva ya que establece asociaciones o relaciones entre las características observadas [13]. Su misión es hacer **inferencias** o extraer **conclusiones científicas.** La presencia de estas asociaciones servirá de base para contrastar las hipótesis de una investigación frente a los datos recogidos empíricamente. La bioestadística analítica o inferencial usa, también, dos tipos de **procedimientos:** la **comprobación de hipótesis** (denominado el contraste de hipótesis [14]) y la estimación de **intervalos de confianza** [15] (Figura 3).

El contraste de hipótesis **confronta** los **resultados** encontrados en una muestra con una **hipótesis** inicial, teórica y universal, para la población de la que procede la muestra o subgrupo. Este proceso iterativo del avance del conocimiento biomédico se realiza gracias a la bioestadística descriptiva e inferencial.

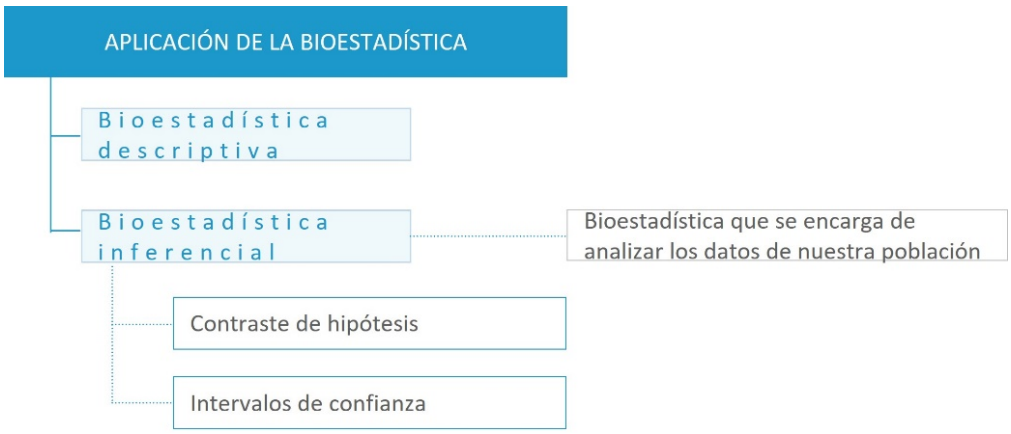


Figura 3. Aplicación de la bioestadística inferencial. Fuente: elaboración propia

© Universidad Internacional de La Rioja (UNIR)

Ideas clave

En la bioestadística inferencial, es fundamental tener en cuenta el concepto de **población** y **muestra.** La muestra es un **subconjunto** de la población que se utiliza para hacer inferencias sobre la población en su totalidad (Figura 4).

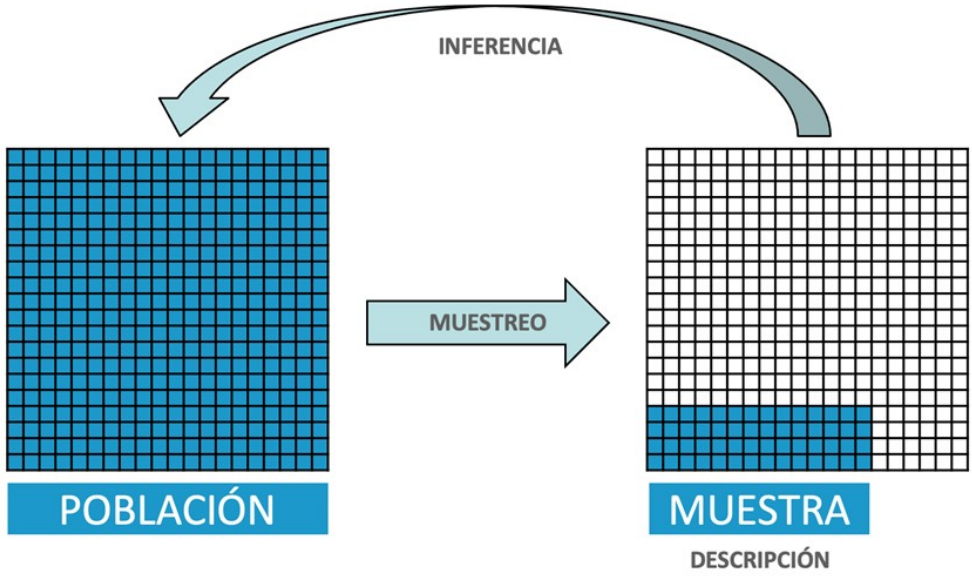


Figura 4. Proceso de inferencia. Fuente: elaboración propia

Uno de los conceptos clave en la bioestadística inferencial es la **distribución de probabilidad** [16]. La distribución de probabilidad es una función que describe la probabilidad de **ocurrencia** de cada valor posible de una **variable aleatoria** en una población. En la bioestadística inferencial se utilizan distribuciones de probabilidad, como la distribución normal o test de **contraste de hipótesis** para realizar pruebas de hipótesis y estimaciones de **parámetros poblacionales.**

Otro concepto clave en la bioestadística inferencial es la **hipótesis estadística** [14]. Una hipótesis estadística es una **afirmación** sobre una población que se quiere probar utilizando los datos de una muestra. En la bioestadística inferencial se utilizan **pruebas** de hipótesis para determinar si la evidencia de la muestra respalda o no una hipótesis estadística.

© Universidad Internacional de La Rioja (UNIR)

Otro concepto clave que va enlazado a la hipótesis en la bioestadística inferencial es el **error tipo I** y el **error tipo II** [17]. El error tipo I se produce cuando se **rechaza** una **hipótesis nula** que es **verdadera.** El error tipo II, por otro lado, se produce cuando se **acepta** una **hipótesis nula** que es **falsa.** Estos errores son inevitables en la inferencia estadística y el objetivo es minimizar su probabilidad de ocurrencia.

Además, en la bioestadística inferencial es importante tener en cuenta el **tamaño muestral.** El tamaño muestral (depende, en parte, del valor del error tipo I y del error tipo II) y se refiere al número de **observaciones** en la muestra. Un tamaño de muestra más grande, generalmente, proporciona una mejor precisión en las estimaciones y en las pruebas de hipótesis. Por lo tanto, es importante considerar el tamaño de la muestra al diseñar un estudio y al interpretar los resultados.

La **significancia estadística** es otro concepto clave en la bioestadística inferencial. La significancia estadística se refiere al grado de **confianza** que se tiene en una **conclusión** estadística basada en los datos de la muestra [18]. Un resultado se considera significativo estadísticamente si la probabilidad de que el resultado es producto del azar es muy baja.

También, la estimación de **parámetros poblacionales** es otro concepto clave en la bioestadística inferencial. La estimación de parámetros poblacionales implica el uso de **estadísticas muestrales** para estimar valores desconocidos de la población. Las técnicas de estimación incluyen el **intervalo de confianza** (por ejemplo, confianza del 90 %, 95 %, o 99 %) y la **estimación puntual** (por ejemplo, calculado mediante el método de mínimo cuadrado u OLS) [19].

© Universidad Internacional de La Rioja (UNIR)

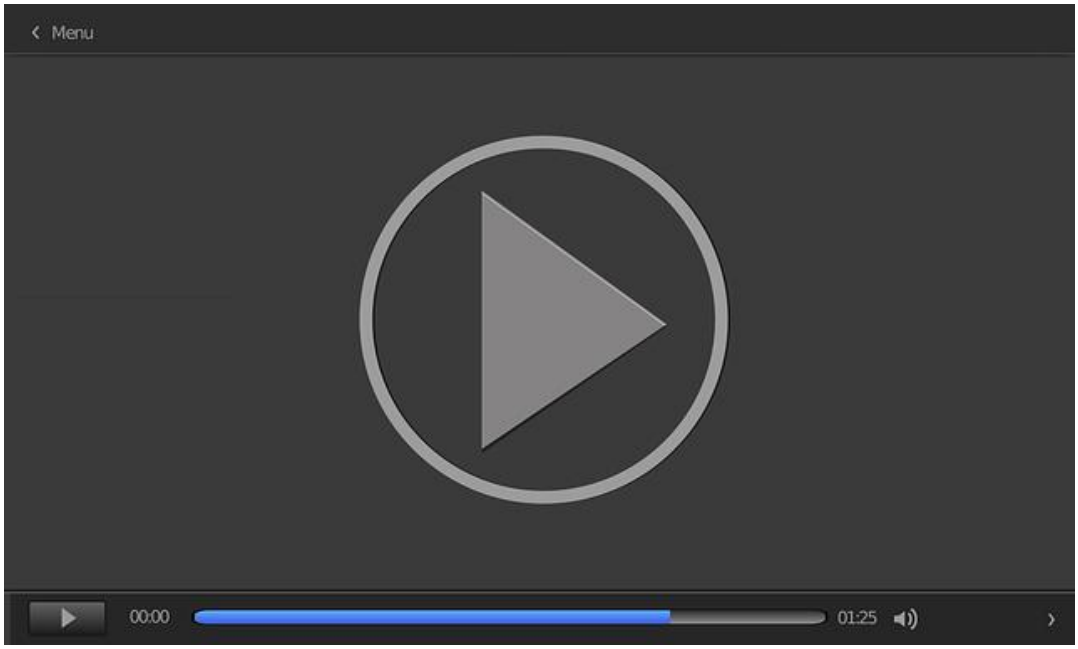
Ideas clave

Por último, en la bioestadística inferencial también se utilizan modelos estadísticos para analizar los datos. Los modelos estadísticos son una forma de describir la relación entre las **variables** y se utilizan para hacer **predicciones** y **estimaciones.** Los modelos pueden ser **paramétricos** o **no paramétricos,** dependiendo de si tienen parámetros desconocidos que se deben estimar a partir de los datos de la muestra.

© Universidad Internacional de La Rioja (UNIR)

4.4. Tipos de estructuras de datos, variables y categorización de datos

En el campo de la bioestadística y, sobre todo, en el ámbito de R, es fundamental comprender las diferentes **estructuras** de datos, variables y categorización de datos. Puedes acceder al vídeo *Estructuración de las bases de datos para análisis estadísticos.*



Estructuración de las bases de datos para análisis estadísticos.

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=23777f2f-1a75-42a9-85e9-b08900db998a

© Universidad Internacional de La Rioja (UNIR)

Estructuras de datos: *dataframes*

Los *dataframes* son estructuras que se encargan no solo de **almacenar** datos, sino también de **conectarlos** entre sí en una estructura con unidad lógica. En términos más generales, un *dataframe* es un conjunto de **datos estructurados** que pertenecen a un mismo **contexto** y, en cuanto a su función, se utiliza para **administrar** de forma electrónica grandes o pequeñas cantidades de **información** en función del objetivo del estudio. Existen dos formatos en la que se pueden encontrar los datos: en formato **longitudinales** o de tipo *long/narrow;* o en formato **transversales** o de tipo *wide.*

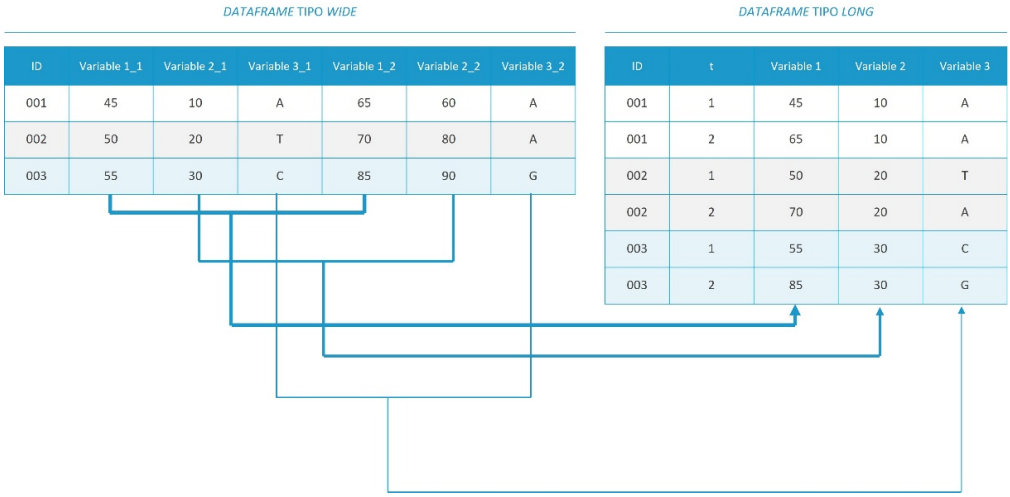
Dataframes* transversales o de tipo *wide

Los *dataframes* transversales son aquellas **bases de datos** que se recopilan en **un solo momento** en el tiempo (Figura 5). Por ejemplo, en un estudio transversal en el que se busca analizar la relación entre el consumo de tabaco y la salud pulmonar, se pueden recopilar datos de un grupo de personas en un momento determinado para estudiar esta relación.

Dataframes* longitudinales o de tipo *long/narrow

Los *dataframes* longitudinales recopilan datos en **diferentes momentos** a lo largo del tiempo (Figura 5). Por ejemplo, en un estudio en el que se busca analizar la evolución de una enfermedad a lo largo del tiempo, se pueden recopilar datos de un mismo grupo de pacientes en diferentes momentos para estudiar la **evolución** de la enfermedad. Este tipo de estructura de datos hace que las observaciones (filas o *rows*) estén duplicados (estructuras jerárquicas) por lo que se requiere de **modelos estadísticos precisos** a la hora de generar inferencias de los datos.

© Universidad Internacional de La Rioja (UNIR)



ID	Variable 1_1	Variable 2_1	Variable 3_1	Variable 1_2	Variable 2_2	Variable 3_2
001	45	10	A	65	60	A
002	50	20	T	70	80	A
003	55	30	C	85	90	G

ID	t	Variable 1	Variable 2	Variable 3
001	1	45	10	A
001	2	65	10	A
002	1	50	20	T
002	2	70	20	A
003	1	55	30	C
003	2	85	30	G

Tabla 3. Remodelación de *dataframes.* Fuente: elaboración propia

Existen diferentes formas de realizar la **remodelación** o *reshape* de los *dataframes* con diferentes librerías de RStudio. A continuación, se muestra un ejemplo y el código para realizar el *reshape* de los datos de un *dataframe* de datos tipo *wide* a un *dataframe* tipo *narrow.* En color negro se muestra la línea de **comandos** y en color morado se muestra la **salida** de RStudio.

© Universidad Internacional de La Rioja (UNIR)

```
# Ejemplo 1: Reshape del DataFrame de formato wide a narrow

#Creamos un DataFrame en formato wide
df <- data.frame(player=c('A', 'B', 'C', 'D'),
                 year1=c(12, 15, 19, 19),
                 year2=c(22, 29, 18, 12))

#Desplegamos el DataFrame
df

  player year1 year2
1      A    12    22
2      B    15    29
3      C    19    18
4      D    19    12

#Importamos librería
install.packages("tidyr") ## si no está instalado
library(tidyr)

#Reshape de los datos a formato narrow
df %>% pivot_longer(cols=c('year1', 'year2'),
                    names_to='year',
                    values_to='points')

  player year  points
1 A      year1     12
2 A      year2     22
3 B      year1     15
4 B      year2     29
5 C      year1     19
6 C      year2     18
7 D      year1     19
8 D      year2     12
```

Figura 5. *Reshape* de un *dataframe* de formato *wide* a *narrow.* Fuente: elaboración propia.

El operador %>% en R usando paquetes del entorno Tidyverse se conoce como pipe. Este operador se utiliza para encadenar **múltiples operaciones** de manera clara y legible. La función de este operador es tomar el resultado de una expresión y pasarlo como el **primer argumento** a la siguiente función en la cadena, lo que permite una escritura más compacta y legible del código

© Universidad Internacional de La Rioja (UNIR)

```
# Ejemplo 2: Reshape de la DataFrame de formato narrow a wide

#Creamos un DataFrame en formato narrow
df <- data.frame(player=rep(c('A', 'B'), each=4),
                 year=rep(c(1, 1, 2, 2), times=2),
                 stat=rep(c('points', 'assists'), times=4),
                 amount=c(14, 6, 18, 7, 22, 9, 38, 4))

#Desplegamos el DataFrame
df

  player year    stat amount
1      A    1  points     14
2      A    1 assists      6
3      A    2  points     18
4      A    2 assists      7
5      B    1  points     22
6      B    1 assists      9
7      B    2  points     38
8      B    2 assists      4

#Importamos librería
install.packages("tidyr") ## si no está instalado
library(tidyr)

#Reshape de los datos a formato wide
df %>% pivot_wider(names_from = stat,
                   values_from = amount)

  player  year points assists
1 A          1     14       6
2 A          2     18       7
3 B          1     22       9
4 B          2     38       4
```

Figura 6. *Reshape* de un *dataframe* de formato *narrow* a *wide*. Fuente: elaboración propia.

© Universidad Internacional de La Rioja (UNIR)

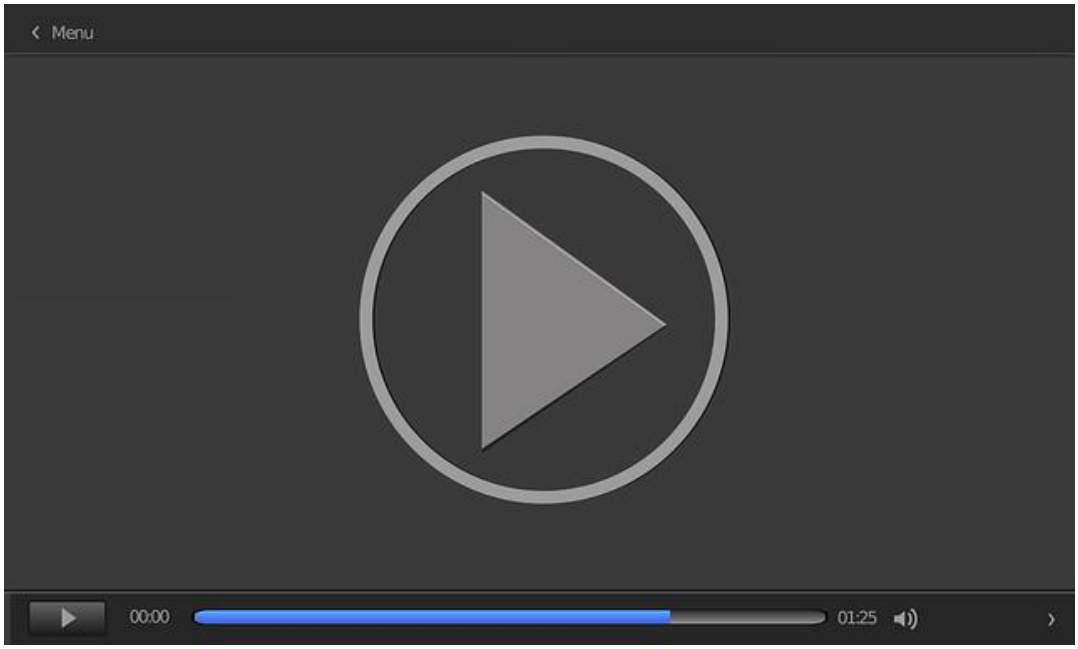
```
# Ejemplo 1: Reshape del DataFrame de formato wide a narrow

#Creamos un DataFrame en formato wide
df <- data.frame(player=c('A', 'B', 'C', 'D'),
                 year1=c(12, 15, 19, 19),
                 year2=c(22, 29, 18, 12))

#Desplegamos el DataFrame
df

  player year1 year2
1      A    12    22
2      B    15    29
3      C    19    18
4      D    19    12

#Importamos librería
install.packages("tidyr") ## si no está instalado
library(tidyr)

#Reshape de los datos a formato narrow
df %>% pivot_longer(cols=c('year1', 'year2'),
                    names_to='year',
                    values_to='points')

  player year  points
1 A      year1     12
2 A      year2     22
3 B      year1     15
4 B      year2     29
5 C      year1     19
6 C      year2     18
7 D      year1     19
8 D      year2     12
```

Los valores de la columna stat ahora se usan como nombres de columna, y los valores de la columna amount se usan como valores de celda en estas nuevas columnas.

© Universidad Internacional de La Rioja (UNIR)

Ideas clave

Puedes ampliar en el siguiente vídeo titulado *Depuración de base de datos.*



Depuración de base de datos

Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=289259e1-04b4-4555-a6d4-b08900db9969

© Universidad Internacional de La Rioja (UNIR)

Estructuras de datos: vectores, matrices y listas

Otras estructuras de datos que se pueden encontrar en el programa R son los vectores, matrices y las listas, entre otros. Los **vectores** son una estructura de datos que pueden **almacenar elementos** del mismo tipo. Se puede pensar en un vector como una **secuencia ordenada** de valores. Los elementos en un vector pueden ser números, caracteres, lógicos u otros objetos de R. En RStudio puedes crear un vector utilizando la función `c()` (que significa concatenar). Dentro de los paréntesis de la función `c()` puedes proporcionar los elementos separados por **comas** para construir el vector.

Por otro lado, una **matriz** es una **estructura** de datos **bidimensional** que contiene elementos del mismo tipo organizados en filas y columnas. De otra forma, una matriz es una tabla rectangular donde cada elemento tiene una **posición única** definida por su fila y columna. En RStudio puedes crear una matriz utilizando la función `matrix()`. Esta función toma varios **argumentos** incluyendo los valores de los elementos, el número de filas y el número de columnas.

Por último, una **lista** es una estructura de **datos** que puede contener diferentes tipos de elementos, como **vectores, matrices, *dataframes*** u otras listas. A diferencia de los vectores y matrices, que contienen elementos del mismo tipo, una lista puede contener elementos de **diferentes tipos.** En RStudio puedes crear una lista utilizando la función `list()`. Dentro de los paréntesis de la función `list()` puedes proporcionar los elementos separados por **comas** para construir la lista.

Tipos de variables

Resulta básico distinguir los diferentes tipos de variables según las **escalas** que se usen para medirlas. Diferenciar con claridad los tipos de variables previene muchos errores en la aplicación e interpretación de cualquier procedimiento estadístico. Para entender los tipos de variables que podemos encontrar pasemos a explicarlo con un ejemplo:

© Universidad Internacional de La Rioja (UNIR)

_n	sexo	edad	grupo	peso	altura	educ	smoke	n_int_fum	n_hijos	dep_smoke	ecivil
1	0	20	1	80,3	156,7	0	0	.	2	.	0
2	1	82	2	56,8	178,4	1	1	1	4	1	2
3	0	40	1	79,4	167,3	1	2	4	5	3	3
4	1	56	2	80,3	192,5	3	0	.	1	.	2

Tabla 4. Ejemplo de base de datos para explicar el tipo de variables. En la parte superior se encuentra el nombre de las variables. Fuente: elaboración propia

Es evidente que encontrar el número 1 en la variable [sexo] no tiene el mismo significado que hallar ese mismo número en la variable [edad]. En el primer caso (variable [sexo]), el número es solo un **indicador** o código que se ha querido asignar en este ejemplo a todos los individuos de sexo femenino (Tabla 4). En cambio, en la variable [edad], el número 1 sería una **cantidad real,** ya que correspondería exactamente a la edad del participante.

La variable [sexo] es una variable **cualitativa** o **categórica;** en cambio, la variable [edad] es una variable **cuantitativa.** Las variables cualitativas o categóricas están medidas en una **escala nominal,** aunque a sus valores se les asignen códigos numéricos, los números son, en realidad, una traducción de sus nombres. Por su parte, la escala de las variables **cuantitativas** corresponde a los **valores reales** de los números que toman. Una diferencia fundamental entre ambas escalas, como se ve en la Tabla 4, es que, por ejemplo, el número 20 en la columna de la variable [edad] corresponde a un valor que es exactamente la mitad del de otra casilla que tuviese el número 40, y también vale exactamente el doble que cuando la casilla contiene un 10. En cambio, cuando el número 2 figura en la variable [grupo] no supone que quienes pertenecen al grupo control valgan el doble que los del grupo de intervención, pues a efectos prácticos hubiese dado lo mismo (incluso hubiera sido preferible) codificar el control con 0 y la intervención con 1.

Además de [edad], otras variables como [peso], [altura] y [presión arterial sistólica] son **cuantitativas** y, por lo tanto, los datos que aparecen en ellas corresponden

© Universidad Internacional de La Rioja (UNIR)

realmente a **números.** En cambio, [sexo], [smoke] son variables cualitativas o categóricas. A su vez, dentro de las variables cuantitativas o realmente numéricas hay dos **posibilidades:**

- Aquellas que admiten **cualquier valor** dentro de un **intervalo** (continuas) sin más restricción que el límite del aparato de medida.
- Aquellas que solo pueden tomar **números enteros** (discretas).

El [peso] y la [talla] son variables **cuantitativas continuas,** ya que, teóricamente, un individuo puede tener un [peso] que tome **cualquier valor.** Por ejemplo, entre 80 y 81 kg podría pesar realmente 80,3333693 kg, y lo mismo se aplica para la [talla] (Tabla 4). En cambio, otras variables, por ejemplo, [n_int_fum], solo pueden tomar números enteros. Otro ejemplo claro es la variable [n_hijos], ya que ninguna familia puede haber tenido realmente 4,33 hijos, o tiene cuatro o tiene cinco (Tabla 4). Estas variables que solo pueden tomar valores de **números enteros** se conocen como **variables cuantitativas discretas.**

Queda por definir otro tipo de variables, las que están en una situación **intermedia.** Vamos a seguir el mismo ejemplo del hábito tabáquico en función del interés en dejar de fumar (variable [dep_smoke]). En este tipo de variables se puede decir que un grado 2 de interés es más **intenso** que un grado 1, pero nunca puede interpretarse como que tener un código 2 implique exactamente el doble de interés que el 1 (Tabla 4). Este tipo de variables se llaman **ordinales** y su uso es muy frecuente en medicina y en todas las demás **ciencias de la salud.** Así, el dolor se puede clasificar en ausente/leve/moderado/intenso y se asignarán, respectivamente, los códigos 0/1/2/3 a cada categoría.

Otro ejemplo de **variable ordinal** es el máximo nivel de estudios alcanzado (variable [educ]) (Tabla 4). No lo es, sin embargo, el estado civil (variable [ecivil]), pues no sería estadística ni políticamente correcto ordenar o **jerarquizar** los diferentes estados civiles. En el campo de la farmacología la respuesta a un tratamiento podría

© Universidad Internacional de La Rioja (UNIR)

valorarse mediante una escala ordinal, asignando, por ejemplo, el código -1 a los que empeoran, el 0 a los que quedan igual, el +1 a los que mejoran algo y el +2 a los que mejoran mucho.

De este modo, podemos dar nombre a las variables teniendo en cuenta la naturaleza de los datos (Figura 7):

- **Variables cualitativas** o categóricas: son variables que corresponden a **características** que están codificadas en **valores numéricos.** De este modo, se distinguen dos tipos de variables cualitativas o categóricas:

  - **Cualitativas nominales:** son aquellas que se utilizan para **clasificar** o categorizar objetos, personas o eventos en grupos distintos. No se establece un orden o jerarquía entre las categorías. Ejemplos comunes de variables nominales son: el género (masculino/femenino), la nacionalidad (español/alemán/francés), la religión (católico/musulmán/judío), entre otras.

  - **Cualitativas ordinales:** son aquellas que permiten **ordenar** o clasificar los objetos, personas o eventos en **categorías jerarquizadas,** pero no se establece una distancia numérica entre ellas. Ejemplos de variables ordinales son: el nivel educativo (primaria/secundaria/universidad), el grado de satisfacción (muy insatisfecho/insatisfecho/satisfecho/muy satisfecho), el estado civil (soltero/casado/viudo), entre otras.

- **Variables cuantitativas:** son variables cuyos números equivalen realmente y con exactitud a los **verdaderos datos.** De este modo se distinguen dos tipos de variables cuantitativas:

  - **Cuantitativas discretas:** son variables numéricas que solo pueden tomar **valores enteros** y no pueden ser fraccionales. Ejemplos de variables discretas son: el número de hijos en una familia, el número de estudiantes en una clase, la cantidad de veces que un evento ocurre en un período de tiempo, etc.

© Universidad Internacional de La Rioja (UNIR)

- **Cuantitativas continuas:** son variables numéricas que pueden tomar **cualquier valor** en un **rango** de valores. Ejemplos de variables continuas son: la altura de una persona, el peso de una persona, la temperatura, el ingreso, etc. Estas variables se miden típicamente en unidades de medida, como metros, kilogramos, grados Celsius o Fahrenheit, dólares, etc.

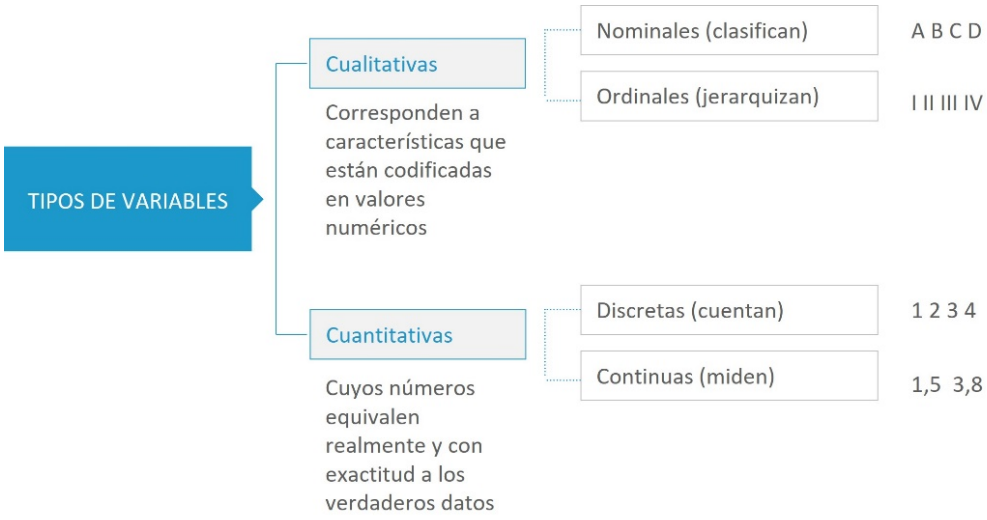


Figura 7. Tipos de variables. Fuente: elaboración propia.

Tipos de variables según el lenguaje de programación R y su aplicación en RStudio.

En R los tipos de variables se denominan de la siguiente manera:

- **Variables nominales:** se denominan factores en R. Los factores son variables **categóricas** que representan una serie de categorías o niveles que **no tienen** un **orden** o jerarquía natural.

- **Variables ordinales:** en R las variables ordinales también se representan como **factores.** La única diferencia es que los factores tienen un **orden lógico,** que se especifica utilizando la función ordered .

© Universidad Internacional de La Rioja (UNIR)

```
# Supongamos que tienes un conjunto de datos que contiene
# información sobre la abundancia relativa de diferentes tipos de
# ARN en células de diferentes tejidos. Para analizar estos datos
# en R, es útil convertir la variable de tipo de tejido en un
# factor ordenado, de manera que se pueda visualizar la abundancia
# de ARN de cada tipo de tejido en orden creciente o decreciente

# Crear un vector con los tipos de tejido
educacion <- c("hígado",
               "pulmón",
               "riñón",
               "cerebro",
               "corazón")

# Crear un factor ordenado para representar los tejidos
educacion_ordenada <- ordered(educacion,
                              levels = c("cerebro",
                                         "corazón",
                                         "hígado",
                                         "riñón",
                                         "pulmón"))
```

Figura 8. Variables. Fuente: elaboración propia.

- **Variables discretas:** en R las variables discretas se pueden representar como **variables numéricas enteras** o como factores con un número finito de niveles. En algunos casos también se pueden representar como variables booleanas (TRUE o FALSE).

- **Variables continuas:** en R las variables continuas se representan, típicamente, como **variables numéricas** con valores **decimales.**

Categorización de datos: transformación de variables

La categorización es una de las transformaciones más comunes en el análisis de datos. Se utiliza para **convertir variables cuantitativas** en **variables cualitativas ordinales,** agrupando los valores individuales en un número limitado de categorías. Por ejemplo, si se quisieran categorizar los valores de expresión génica se podrían crear seis categorías diferentes. El resultado de la categorización es una **nueva variable** denominada [expresion_cat], que contiene solo seis valores posibles

© Universidad Internacional de La Rioja (UNIR)

(0=Muy bajo, 1=Bajo, 2=Medio, 3=Alto, 4=Muy alto). Este proceso convierte una escala de razón en una escala ordinal, pero también implica una **pérdida de información.** Por lo tanto, se recomienda no recopilar información en una escala ordinal si se puede hacer en una escala cuantitativa más precisa. En general, es mejor recopilar variables cuantitativas con el máximo nivel de detalle posible y solo después categorizarlas, si es necesario, para el tipo de análisis estadístico que se quiere realizar.

```
# Generamos los datos inventados
set.seed(123)
datos_expresion <- data.frame(expresion = rnorm(10,
                                                mean = 3,
                                                sd =2))

# Creamos la variable de categorías
datos_expresion$expresion_cat <- cut(datos_expresion$expresion,
                                     breaks=c(0,1,2,3,4,Inf),
                                     labels=c("Muy bajo",
                                              "Bajo",
                                              "Medio",
                                              "Alto",
                                              "Muy alto"))

# Imprimimos los primeros registros de los datos
head(datos_expresion)

  expresion expresion_cat
1  1.879049          Bajo
2  2.539645         Medio
3  6.117417      Muy alto
4  3.141017          Alto
5  3.258575          Alto
6  6.430130      Muy alto
```

Figura 9. Categorización de datos. Fuente: elaboración propia.

Variables *dummies*

Las variables *dummies* son variables que se utilizan para representar **información categórica** en **modelos estadísticos.** Son variables binarias que toman el valor 1 cuando la categoría es `TRUE` y 0 cuando es `FALSE`. Su función principal es representar información categórica en modelos de **regresión lineal,** ya que estos

© Universidad Internacional de La Rioja (UNIR)

modelos requieren que las variables explicativas sean numéricas. En el caso de las variables categóricas es necesario transformarlas en variables numéricas que puedan ser utilizadas en el modelo. En R se pueden crear variables *dummies* utilizando la función dummyVars de la librería caret. A continuación, se pone un ejemplo de cómo generar variables *dummies* en R:

```
# Supongamos que tenemos un conjunto de datos que contiene
# información sobre diferentes especies de plantas y queremos
# crear variables dummies para la variable "tipo de planta", que
# puede ser "arbusto", "herbácea" o "árbol"

# Cargar el paquete caret
install.packages("caret") ## si no está instalado
library(caret)

# Generar los datos inventados
set.seed(123)
plantas <- data.frame(tipo = sample(c("arbusto",
                                      "herbácea",
                                      "árbol"),
                                      50,
                                      replace = TRUE))

# Crear variables dummies
vars_dummies <- dummyVars(~tipo,
                          data = plantas)

# Aplicar la transformación a los datos
datos_dummies <- predict(vars_dummies,
                         newdata = plantas)

# Imprimir los primeros registros de los datos transformados
head(datos_dummies)

  tipoárbol tipoarbusto tipoherbácea
1         1           0            0
2         0           1            0
3         0           0            1
```

Figura 10. Variables *dummies*. Fuente: elaboración propia.

© Universidad Internacional de La Rioja (UNIR)

4.5. Referencias bibliográficas

1.Mishra P, Pandey C, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth. 2019 en.;22(1):67.

2.Guetterman TC. Basics of statistics for primary care research. Fam Med Community Health. 2019 mzo. 28;7(2).

3. de Leon J. Evidence-Based Medicine Versus Personalized Medicine. J Clin Psychopharmacol. 2012 abr.;32(2):153–64.

4. Cristopher AN. Interpreting and Using Statistics in Psychological Research. Cap. 1; Drawing Conclusions From Data: Descriptive Statistics, Inferential Statistics, and Hypothesis Testing; p. 145-183.

5. National Institute of Standards and Technology. NIST/SEMATECH [Internet]. 2012 abr. Engineering Statistics [citado 2023 sept. 11]. Disponible en: https://www.itl.nist.gov/div898/handbook/

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

7. Signorell A. Cran [Internet]. 2023 sept. 6. DescTools: Tools for Descriptive Statistics [citado 2023 sept. 11]. Disponible en: https://cran.r-project.org/web/packages/DescTools/index.html

8. Komsta L, Novomestky F. Cran [Internet]. 2022 may. 2. moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests [citado 2023 sept. 11]. Disponible en: https://cran.r-project.org/web/packages/moments/index.html

9. Pearson K. Contributions to the mathematical theory of evolution. Philos Trans R Soc Lond A. 1895 nov 16; 1185(1894): 71-110.

© Universidad Internacional de La Rioja (UNIR)

10. Toit SHC, Steyn AGW y Stumpf RH. Graphical Exploratory Data Analysis. Springer; 1986.

11.Beniger JR, Robyn DL. Quantitative Graphics in Statistics: A Brief History. Am Stat. 1978 febr.; 32(1): 1–11.

12.Jarrell, Stephen B. Basic Statistics. Edición especial prepublicación. William C Brown; 1994.

13. Audi R. Inference. The Cambridge Dictionary of Philosophy. Cambride University Press.

14.Lehmann E.L. y Romano JP. (2005). Testing Statistical Hypotheses. Nueva York: Springer; 2022.

15. Dekking FM, Kraaikamp C, Lopuhaa HP y Meester LE. A Modern Introduction to Probability and Statistics: Understanding Why and How. Londres: Springer; 2005.

16. Ash RB. (2008). Basic probability theory. Mineola: Dover Publications; 2008.

17. Dekking FM, Kraaikamp C, Lopuhaa HP y Meester LE. A Modern Introduction to Probability and Statistics: Understanding Why and How. Londres: Springer; 2005.

18. Borror CM. The Certified Quality Engineer Handbook. Statistical decision making; 418-472.

19.Pradip KS, Pal SR y Das AK. Estimation and Inferential Statistics. India: Springer; 2015.

© Universidad Internacional de La Rioja (UNIR)

Extensión en tipos de variables para su aplicación en R

Tipos de variables [Internet]. RPubs; [Citado 29 de marzo de 2023]. Recuperado a partir de: https://rpubs.com/gustavomtzv/1062185

Conocimientos sobre qué son los vectores, vectores numéricos, vectores de caracteres, vectores de fechas, vectores de categorías numéricas y etiquetas.

© Universidad Internacional de La Rioja (UNIR)

Extensión de etiquetar y recodificar variables en R

García-García F. Biocosas [Internet]. 2014 dic. 16. Gestión de datos con R [citado 2023 sept. 11]. Disponible en: https://biocosas.github.io/R/020_gestion_datos.html

Conocimiento sobre cómo recodificar diferentes variables dependiendo de si recodificamos los valores de una variable por otros valores en esa misma variable o si realizamos una recodificación de valores, pero en otra variable distinta de la inicial. También se explica cómo categorizar una variable cuantitativa en una cualitativa.

© Universidad Internacional de La Rioja (UNIR)

Tutorial de extensión de data reshaping en R

Smith O. Datacamp [Internet]. 2020 may. Data Reshaping in R tutorial [citado 2023 sept. 11]. Disponible en: https://www.datacamp.com/tutorial/data-reshaping-in-r

Conocimiento sobre cómo hacer el *reshape* de los datos utilizando otras funciones y librerías diferentes (rbind() , cbind() , along with Melt() , Dcast()). Además, se muestra la función transponer t() , muy útil en el manejo de matrices.

© Universidad Internacional de La Rioja (UNIR)

1. Señale la respuesta incorrecta: ¿por qué es importante seleccionar cuidadosamente una muestra adecuada en el análisis estadístico?

A. Porque una muestra inadecuada puede llevar a conclusiones erróneas o sesgadas.

B. Porque una muestra grande garantiza la precisión de los resultados obtenidos.

C. Porque una muestra aleatoria siempre representa con precisión la población completa.

D. Porque la elección de una muestra no tiene impacto en la validez de los resultados obtenidos.

2. Señale la respuesta incorrecta: ¿cuál de las siguientes representaciones gráficas es la más adecuada para mostrar la distribución de frecuencia de una variable continua?

A. Gráfico de barras

B. Histograma

C. Gráfico de sectores

D. Gráfico de líneas

© Universidad Internacional de La Rioja (UNIR)

3. ¿Cuál de las siguientes afirmaciones es falsa acerca de los diagramas de caja y bigotes?

A. Los diagramas de caja y bigotes muestran la mediana, el rango intercuartil, el valor mínimo y el valor máximo de una distribución de datos.

B. Los bigotes se extienden hasta el valor mínimo y máximo de los datos respectivamente.

C. Los puntos que están fuera de los bigotes representan valores atípicos o extremos.

D. Los diagramas de caja y bigotes son útiles solo para distribuciones simétricas.

4. ¿Cuál de las siguientes opciones no es un tipo de variable?

A. Nominal

B. Ordinal

C. Fraccional

D. Continua

5. Señala la opción incorrecta. ¿Cuál es la principal diferencia entre una variable ordinal y una variable nominal?

A. La variable nominal tiene un orden establecido, mientras que la variable ordinal no lo tiene.

B. La variable ordinal tiene un orden establecido, mientras que la variable nominal no lo tiene.

C. Ambas variables tienen el mismo orden establecido.

D. No hay diferencia entre las dos variables.

© Universidad Internacional de La Rioja (UNIR)

6. ¿Qué tipo de variable es la temperatura?

A. Nominal

B. Ordinal

C. Discreta

D. Continua

7. ¿Qué tipo de variable es el nivel de educación (primaria, secundaria, universidad, etc.)?

A. Nominal

B. Ordinal

C. Discreta

D. Continua

8. ¿Qué tipo de variable es el número de hermanos que tiene una persona?

A. Nominal

B. Ordinal

C. Discreta

D. Continua

9. ¿Qué tipo de variable es el ingreso anual de una persona?

A. Nominal

B. Ordinal

C. Discreta

D. Continua

© Universidad Internacional de La Rioja (UNIR)

10. ¿Cuál de las siguientes opciones representa una variable que tiene valores que no se pueden medir?

A. Variable continua

B. Variable nominal

C. Variable ordinal

D. Variable discreta

© Universidad Internacional de La Rioja (UNIR)