Análisis de Datos con Jamovi

Guía Práctica para la Investigación Científica

Prof. Dr. Miguel Lafuente Blasco

Departamento de Métodos Estadísticos

Instituto de Biocomputación y Física de Sistemas Complejos (BIFI)

Análisis de Datos con Jamovi

Guía Práctica para la Investigación Científica

Miguel Lafuente Blasco

1ª edición. Zaragoza, 2025.

Edita: Servicio de Publicaciones. Universidad de Zaragoza.

ISBN: 978-84-10169-43-2

Prefacio

Este texto surge de la necesidad de contar con un material didáctico adecuado para la asignatura *Análisis de Datos* del Máster Universitario en Evaluación y Entrenamiento Físico para la Salud de la Universidad de Zaragoza. Sin embargo, las técnicas descritas aquí no se limitan al ámbito de las ciencias del deporte o de la salud, sino que son aplicables en multitud de campos de la investigación científica.

Para ello, este libro ofrece una guía paso a paso para el análisis de datos con *jamovi*, un software sencillo e intuitivo que está ganando rápidamente popularidad, especialmente en ámbitos relacionados con la investigación en salud y sociología, desplazando a otros programas como *SPSS*.

Si bien este texto no tiene como objetivo principal la abordar la teoría estadística, se incluyen explicaciones sobre las razones y buenas prácticas a la hora de aplicar las técnicas de análisis de datos. Además, especialmente en las últimas secciones, se hace énfasis en la comunicación de los resultados en el contexto de la publicación científica.

Cada procedimiento se ilustra con capturas de pantalla detalladas, de modo que cualquier persona, incluso sin conocimientos previos en análisis de datos, pueda seguir el proceso con facilidad y replicar los resultados de manera autónoma.

El contenido de este libro está estructurado en cinco capítulos con ejercicios propuestos:

1. **Introducción y procesamiento de datos**: se presentan las herramientas básicas de *jamovi* y se explica cómo importar, organizar y manipular bases de datos.
2. **Estadística descriptiva univariante**: abarca los métodos para resumir y visualizar una única variable, sintetizando su distribución y características principales.
3. **Estadística descriptiva bivariante**: dedicada al análisis de relaciones entre dos variables, explorando tablas de contingencia, correlaciones y asociaciones.
4. **Inferencia estadística**: presenta las pruebas de hipótesis y estimaciones mediante intervalos de confianza.
5. **Modelos de regresión lineal**: técnicas adecuadas para describir relaciones más complejas entre variables y desarrollar modelos predictivos.

Por último, los datos empleados en este libro están disponibles en acceso abierto en el enlace de la referencia [1], desde donde pueden descargarse libremente. En ese mismo enlace también se encuentran materiales adaptados a otros software y contenidos teóricos sobre distintos aspectos de la Estadística.

Índice general

"Sin datos, sólo eres otra persona con una opinión."

W. Edwards Deming

1

Introducción a *jamovi*

1.1. Origen y usos de *jamovi*

jamovi, estilizado en minúsculas, es una herramienta estadística gratuita y de código abierto diseñada para el análisis de datos. Fue desarrollado por Jonathon Love, Damian Dropmann y Ravi Selker [5], quienes anteriormente colaboraron en el proyecto *JASP*. Este software combina la accesibilidad de una interfaz gráfica intuitiva con la potencia del lenguaje `R` [4].

- **Gratuito y accesible:** Disponible sin coste y de código abierto.
- **Interfaz intuitiva:** Facilita el análisis estadístico sin necesidad de conocimientos avanzados en programación.
- **Compatibilidad:** Permite importar y exportar datos en formatos como `.csv`, `.sav` (*SPSS*), `.sas7bdat` (*SAS*) y `.dta` (*Stata*).
- **Extensibilidad:** Ofrece la posibilidad de añadir módulos adicionales para análisis avanzados como meta-análisis o modelos bayesianos.
- **Actualización automática:** Recalcula los resultados dinámicamente al modificar los datos.
- **Multiplataforma:** Compatible con Windows, Linux y Mac.
- **Mejoras continuas:** Se actualiza periódicamente con nuevas funciones y optimizaciones.

Su diseño gráfico intuitivo lo convierte en una alternativa accesible a programas como *SPSS*, con la ventaja de ser completamente gratuito y de código abierto. Además, su arquitectura modular permite expandir sus funcionalidades a través de bibliotecas desarrolladas por la comunidad, lo que lo hace adaptable a distintos ámbitos de investigación y enseñanza.

Pensado principalmente para entornos académicos y de investigación, *jamovi* ofrece herramientas para estadística descriptiva, pruebas de hipótesis, regresión, análisis de tablas de contingencia y más, con la posibilidad de visualizar los resultados en tiempo real. También permite la integración con `R`, brindando a los usuarios avanzados la opción de personalizar sus análisis y desarrollar sus propios módulos.

1.1.1. Instalación

Como cualquier otro software, *jamovi* debe instalarse en el ordenador. Se puede descargar de forma gratuita desde su página oficial: `https://www.jamovi.org/`. En la sección de descargas, se encuentran enlaces específicos para los sistemas operativos Windows, Mac y Linux.

Para instalarlo, sigue estos pasos:

1. Ve a la página de descargas: `https://www.jamovi.org/download.html`.
2. Descarga el instalador correspondiente a tu sistema operativo.
3. Ejecuta el instalador y, si es necesario, otorga los permisos requeridos.

En el momento de escribir este texto, la versión disponible de *jamovi* es la 2.6, pero el programa se actualiza con frecuencia, por lo que es posible que haya una versión más reciente.

1.1.2. Interfaz

Una vez instalado, al abrir *jamovi* se muestra una interfaz dividida en dos secciones principales (ver Figura 1.1):

- **A la izquierda**: La vista de hoja de cálculo, donde se pueden introducir datos manualmente o importar archivos en formatos como `.csv`, `.xlsx` (*Excel*), `.sav` (*SPSS*), `.sas7bdat` (*SAS*), `.dta` (*Stata*) y `.jasp` (*JASP*). Desde esta sección es posible modificar valores, agregar o eliminar filas y columnas, así como definir las propiedades de las variables (tipo de medida, etiquetas, etc.).
- **A la derecha**: El área de resultados, donde se generan automáticamente las tablas y gráficos correspondientes a los análisis realizados.

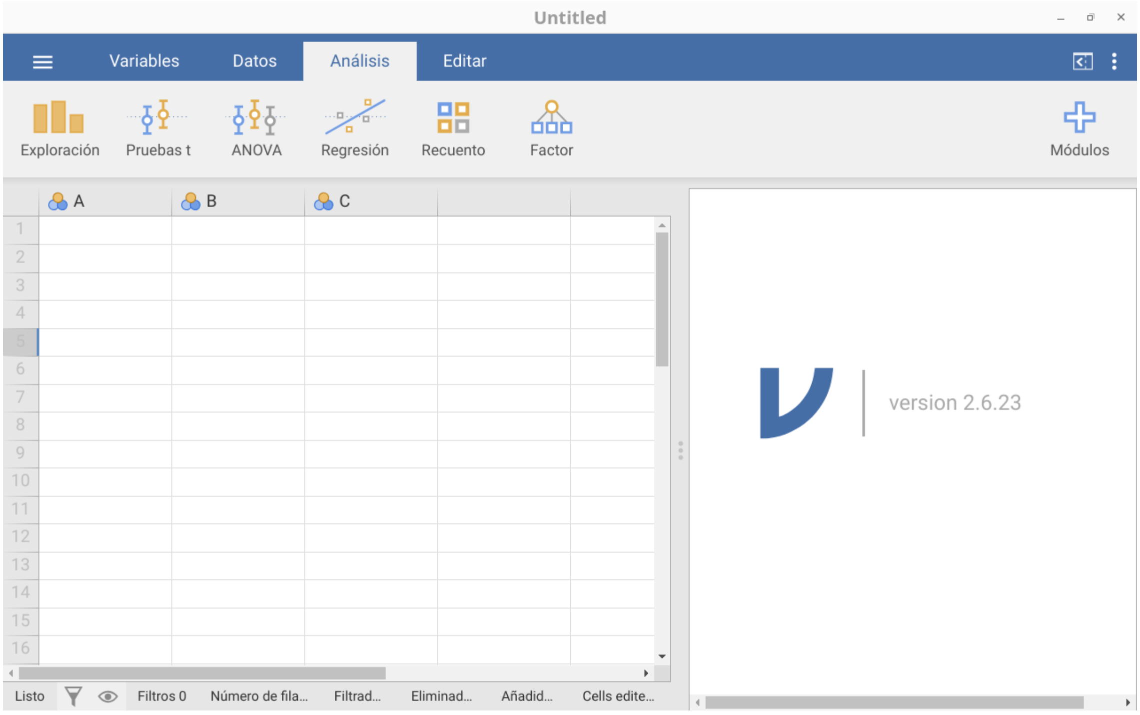


Figura 1.1: Interfaz de *jamovi*.

1.1.3. Funciones principales

El menú `Datos` permite realizar diversas operaciones sobre la base de datos, tales como:

- `Portapapeles`: Copiar, cortar y pegar observaciones.
- `Editar`: Deshacer o rehacer acciones previas.
- `Variables`: Configurar propiedades de las variables (`Configurar`), crear nuevas variables (`Calcular`), realizar transformaciones (`Transformar`) y añadir o eliminar variables (`Añadir/Eliminar`).
- `Ponderación`: Aplicar pesos a las observaciones para ajustar el análisis a la representatividad de los datos.
- `Filtros`: Aplicar filtros y gestionar observaciones agregándolas o eliminándolas.

El menú `Análisis` ofrece una amplia variedad de herramientas estadísticas:

- `Exploración`: Generación de tablas de resumen y gráficos descriptivos.
- `Pruebas t`: Pruebas paramétricas y no paramétricas para la comparación de medias o medianas en una muestra, dos muestras independientes o relacionadas.

- `ANOVA`: Análisis de varianza de uno o varios factores, ANCOVA, MANCOVA, modelos de medidas repetidas y pruebas no paramétricas asociadas.
- `Regresión`: Cálculo de correlaciones y ajuste de modelos de regresión lineal y logística (binaria, ordinal y multinomial).
- `Frecuencias`: Creación de tablas de contingencia y pruebas como Chi-cuadrado y McNemar.
- `Factor`: Análisis de fiabilidad, análisis de componentes principales y análisis factorial.

1.2. Manejo de un fichero de datos en *jamovi*

1.2.1. Creación de un fichero

La vista de datos de *jamovi* tiene un diseño similar al de una hoja de cálculo, como *Excel*. En esta sección se introducen los datos de la investigación. Cada columna representa una variable de estudio, y cada fila corresponde a un individuo (caso o unidad estadística) del que se ha recopilado información.

- La primera celda de cada fila muestra el número de caso.
- Los datos pueden introducirse por filas (usando → para avanzar entre celdas) o por columnas (usando la tecla `<return>`).
- En la parte superior de cada columna aparece el nombre de la variable. Al hacer doble clic en la cabecera de la columna, se accede a las propiedades de la variable, permitiendo editar su nombre y modificar sus atributos.

> **Sobre el tipo de medida** Dependiendo del tipo de variable, es importante indicar a *jamovi* la naturaleza de los datos para que el software pueda aplicar correctamente los análisis y operaciones estadísticos correspondientes (ver Figura 1.2).

Ejemplo de creación de un fichero: alturas de chicos y chicas

Para ilustrar el proceso, se estudiará la altura de chicos y chicas del último curso de bachillerato en un colegio. No es viable medir a todos los estudiantes (realizar un censo), por lo que se selecciona un subconjunto representativo (una muestra). Se eligen aleatoriamente 4 chicos y 4 chicas, y se registran sus alturas. Los datos obtenidos se muestran en la Tabla 1.1:

Chicos	Chicas
1.77	1.65
1.82	1.68
1.71	1.56
1.65	1.72

Tabla 1.1: Alturas de una muestra de chicos y chicas.

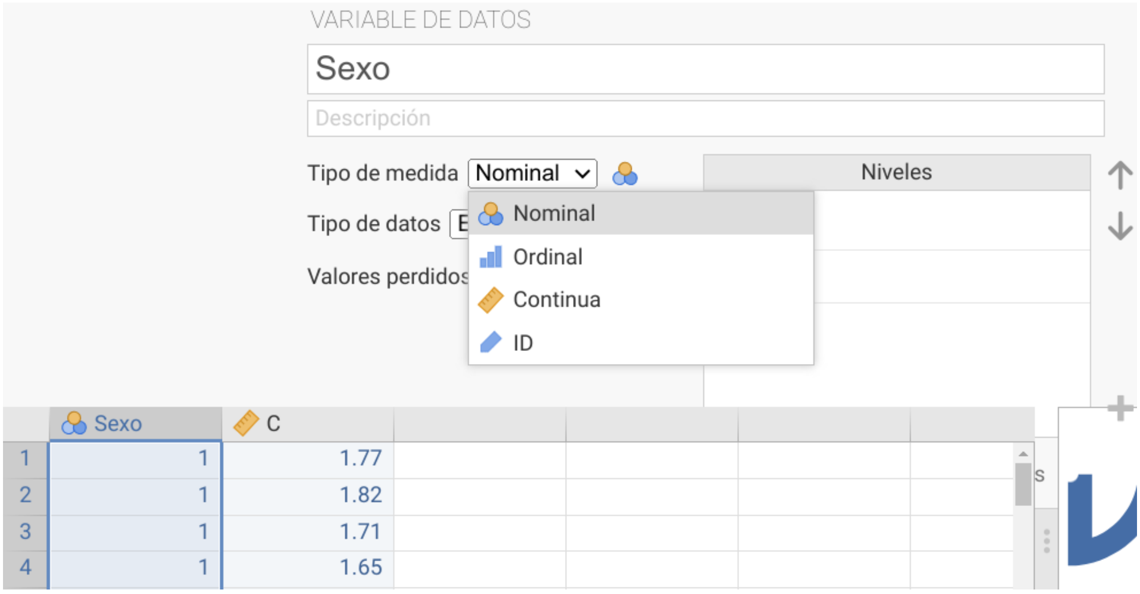


Figura 1.2: Modificación de los atributos de una variable.

Los datos corresponden a una muestra de 8 individuos. Para representar esta información en *jamovi*, se definen dos variables:

- `sexo`: Variable categórica (cualitativa nominal) codificada como 1 =*hombre* y 2 =*mujer*. Alternativamente, se pueden registrar directamente los nombres de las categorías.
- `altura`: Variable numérica continua que representa la altura en metros.

Ingreso de los datos en *jamovi*

Para registrar la información en *jamovi*, primero introduciremos los valores en la hoja de datos y, posteriormente, configuraremos correctamente cada variable.

1. **Introducción de los valores en la hoja de datos** Los datos deben ser ingresados siguiendo la estructura de filas y columnas propia de *jamovi*:
   - En la primera columna, se registrarán los valores correspondientes a la variable `sexo`. Para ello, se escriben cuatro veces el número 1 seguido de cuatro veces el número 2, representando hombres y mujeres respectivamente.

- En la segunda columna, se anotan las alturas correspondientes a cada persona, manteniendo el orden de la variable anterior.

2. **Configuración de las variables en *jamovi*** Una vez introducidos los datos, es necesario indicar a *jamovi* el tipo de variable y su formato para que los análisis se realicen correctamente.

   a) Para la primera columna, donde se registró la variable `sexo`, se debe:
      - Escribir `sexo` en el campo de nombre de variable.
      - Incluir una descripción adecuada.
      - Seleccionar `Nominal` como tipo de medida.
      - Configurar los valores de la variable: 1 para *Hombre* y 2 para *Mujer*.

      En la Figura 1.3 se muestra cómo realizar esta configuración.

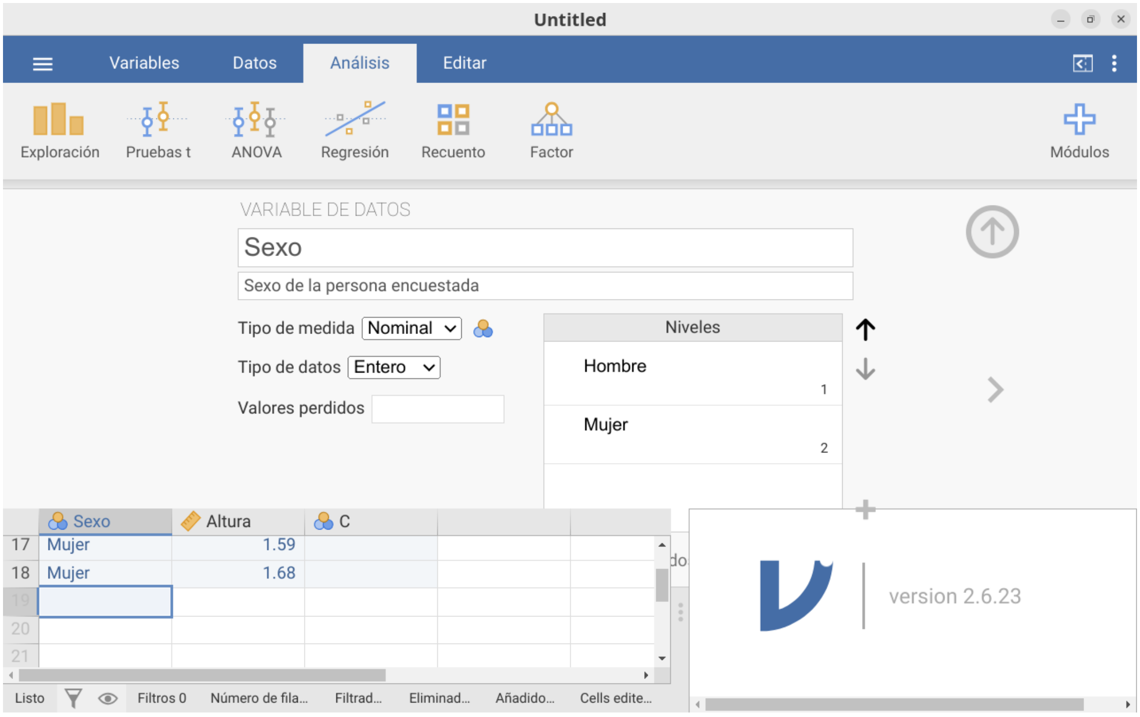


Figura 1.3: Configuración de la variable `sexo`.

   b) Para la segunda columna, que almacena la variable `altura`, se debe:
      - Escribir `altura` en el campo de nombre de variable.
      - Incluir una descripción adecuada.
      - Definir la medida como `Continua`.
      - Configurar el tipo de datos como `Decimal`.

      En la Figura 1.4 se muestra la configuración correspondiente.

1.2.2. Manejo de datos faltantes en *jamovi*

En *jamovi*, los datos faltantes pueden gestionarse de dos maneras:

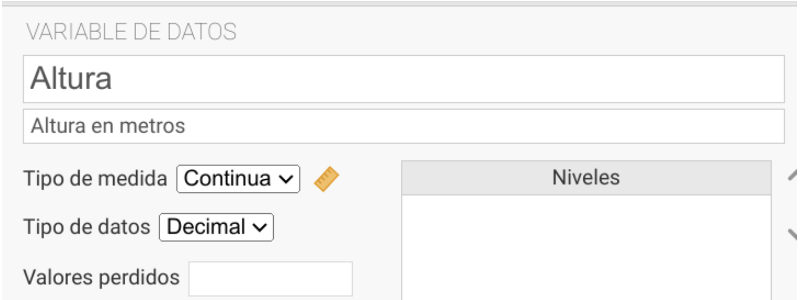


Figura 1.4: Configuración de la variable `altura`.

- **Celda vacía**: La opción más sencilla es dejar las celdas vacías en la hoja de datos. *Jamovi* identificará automáticamente estos valores como datos faltantes y los excluirá de los análisis.

- **Código de datos perdidos**: También es posible asignar un valor específico (por ejemplo, `-99`) para representar datos faltantes. Para que *jamovi* lo reconozca correctamente, es necesario acceder al menú `Variables` → `Configurar` o hacer doble clic en el nombre de la variable y definir el valor codificado como "Perdido". En el ejemplo de la Figura 1.5, se consideran como datos perdidos aquellos con valor igual a -99.

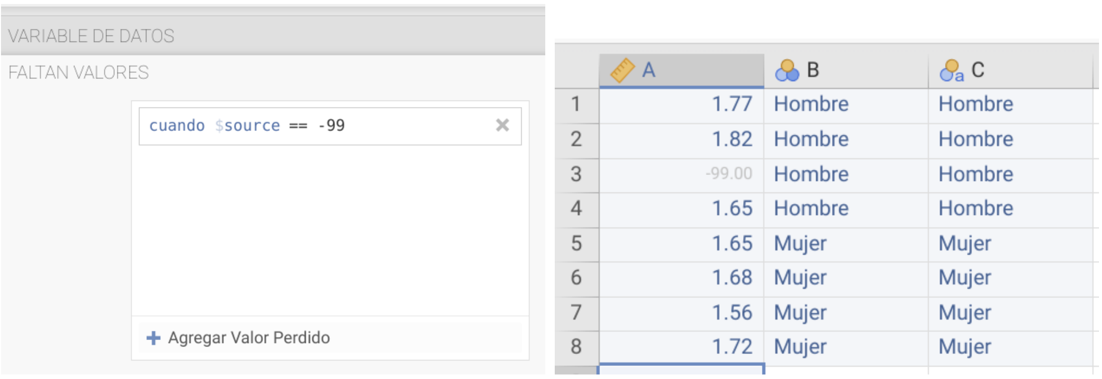


Figura 1.5: Configuración de códigos de valores perdidos y ejemplo de su apariencia en la hoja de datos.

1.2.3. Guardar el archivo

Para guardar el archivo sin salir de la vista de datos, selecciona `Archivo` (representado por el icono de los puntos y barras horizontales) y luego `Guardar como`. En el cuadro de diálogo que aparece, elige la ubicación deseada y asigna un nombre al archivo (por ejemplo, `altura.omv`). *Jamovi* guardará los datos en su formato nativo (`.omv`). También es posible `Exportar` los datos en formatos más comunes, como `.csv` o `.xlsx`.

Nota 1.1. El formato `.omv` es el formato nativo de *jamovi*, permitiendo almacenar tanto los datos como los análisis realizados dentro del programa, garantizando su conservación al volver a abrir el archivo. Sin embargo, es común trabajar con formatos más extendidos, como `.xlsx` (*Excel*) o `.csv` (valores separados por comas), que ofrecen mayor compatibilidad con otros programas.

1.3. Transformaciones de variables

A veces es necesario transformar los datos originales para generar nuevas variables que faciliten el análisis. Por ejemplo, a partir de los datos del archivo `altura.omv`, se puede crear una nueva variable llamada `centimetros`, que represente la altura en centímetros en lugar de metros.

Para realizar esta transformación:

1. Ir al menú `Datos` → `Calcular`.
2. Asignar el nombre `AlturaCM` a la nueva variable.
3. En la expresión, escribir `Altura * 100` (ver Figura 1.6).

   Nota: Es importante escribir exactamente el nombre de la variable, respetando mayúsculas, acentos y otros caracteres.

Nota 1.2. La nueva variable `AlturaCM` aparecerá automáticamente en la hoja de datos. En *jamovi*, las transformaciones son dinámicas: si se modifica un valor en la variable original (`altura`), el cambio se reflejará inmediatamente en `AlturaCM`.

Figura 1.6: Menú para calcular una variable a partir de otras.

1.4. Recodificación de variables

En algunas situaciones, es necesario modificar la codificación original de una variable, ya sea para agrupar valores dentro de nuevas categorías o para convertir una variable numérica en una variable categórica. Esto puede ser útil, por ejemplo, al clasificar la variable `Pulso1`, del fichero `pulso.omv` en diferentes niveles de frecuencia cardíaca o al agrupar rangos de edad en categorías más amplias. Estos datos corresponden a un experimento real [6] descrito en la Sección 1.8 de ejercicios.

Para recodificar una variable en *jamovi*:

1. Hacer clic derecho sobre la variable en la vista de datos.

2. Seleccionar la opción `Transformar`.

3. Asignar un nombre a la nueva variable.

4. Definir la transformación deseada. Véase el Ejemplo 1.3 y la Figura 1.7.

Ejemplo 1.3. Recodificación de la variable `Pulso1` en una nueva variable con tres categorías:

- *Baja*: Valores menores a 60.
- *Normal*: Valores entre 60 y menos de 80.
- *Alta*: Valores de 80 o superiores.

Para ello, se debe hacer clic en *Agregar condición de recodificación* tantas veces como sea necesario. Las condiciones se aplican sobre la variable `$source`, que en este caso es `pulso1`.

Figura 1.7: Recodificación de la variable `Pulso1`. Nótese que los valores ya asignados en un nivel anterior no se reasignan en las siguientes condiciones.

1.5. Modificar orden de las clases en variables categóricas

Las variables categóricas en *jamovi* pueden gestionarse desde el menú `Datos` → `Configurar`, o haciendo doble clic sobre el nombre de la variable. Aquí se pueden visualizar los niveles de la variable, reorganizarlos si es necesario o cambiar sus etiquetas.

Por defecto, los niveles de las variables categóricas se ordenan alfabéticamente, pero esto puede modificarse moviéndolos hacia arriba o hacia abajo. También es posible cambiar el nombre de los niveles simplemente sobrescribiendo sus valores.

Ejemplo 1.4. Cambiar las etiquetas de la variable `Correr` para que en lugar de `1` y `2` aparezcan “Hombre” y “Mujer”, como se había realizado en la Figura 1.3.

Ahora con las flechas se pueden reordenar los niveles, ver Figura 1.8. Esto servirá para ordenar correctamente las clases en una variable ordinal, o simplemente para mostrar tablas y gráficos en un orden que se considere más adecuado.

Figura 1.8: Cambio de orden de las clases de una variable.

1.6. Filtrado de datos

En muchas ocasiones, es útil seleccionar un subconjunto específico de registros para el análisis. *jamovi* permite aplicar filtros desde el menú `Datos` → `Filtros`, estableciendo condiciones que deben cumplir las observaciones seleccionadas.

Las condiciones pueden basarse en valores numéricos o categóricos. En el caso de variables categóricas, el valor de la categoría debe escribirse entre comillas, respetando mayúsculas y minúsculas para que el filtro se aplique correctamente.

Ejemplo 1.5. Filtrar únicamente las observaciones correspondientes a mujeres (`Sexo` = "Mujer").

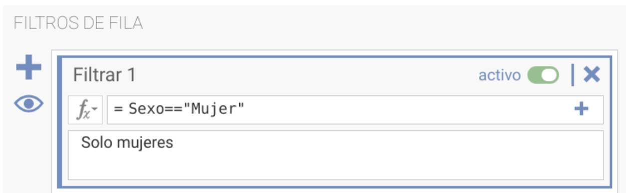


Figura 1.9: Creación de un filtro para seleccionar solo las observaciones de mujeres.

Nota 1.6.
- Si la variable tiene un espacio en el nombre, esta puede escribirse como `` `Nombre Variable` ``, con los acentos al revés encapsulando el nombre.
- En lugar de `==`, se puede usar símbolos de mayor o menor que, o incluso el signo *distinto de*: `!=`.

Una vez aplicado el filtro, en la primera columna de la vista de datos se indicará qué observaciones han sido excluidas del análisis y cuáles permanecen activas.

1.7. Otros consejos prácticos y módulos

- Puedes copiar tablas y gráficos directamente desde *jamovi* a documentos de *Word* o *Excel* utilizando `<Control> + C` y `<Control> + V`.

	Filtrar 1	Altura	Peso	Edad	Sexo
1	✓	173	57.0	18	Mujer
2	✓	179	58.0	19	Mujer
3	✓	167	62.0	18	Mujer
4	✗	195	84.0	18	Hombre
5	✓	173	64.0	18	Mujer
6	✗	184	74.0	22	Hombre
7	✓	162	57.0	20	Mujer
8	✓	169	55.0	18	Mujer
9	✓	164	56.0	19	Mujer
10	✗	168	60.0	23	Hombre
11	✗	170	75.0	20	Hombre

Figura 1.10: Vista de los datos con la indicación de qué observaciones han sido filtradas y cuáles siguen activas.

- **Módulos**: *Jamovi* permite ampliar sus funciones mediante módulos adicionales. Para instalarlos, accede al menú `Módulos` y luego a `Biblioteca jamovi`, donde podrás seleccionar e instalar el módulo deseado.

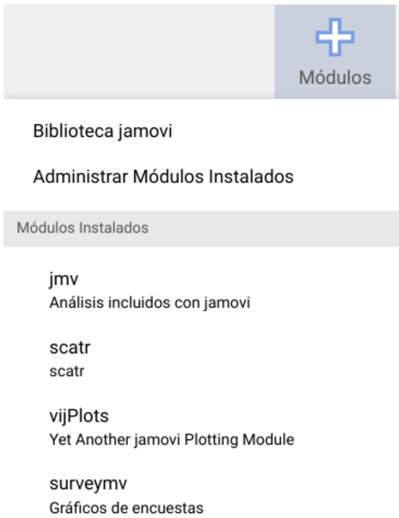


Figura 1.11: Menú de exploración e instalación de módulos en *jamovi*.

1.8. Ejercicios

1. En la referencia [6] se describen unos datos reales, llamados `pulso.omv` en nuestro caso, en el que a los estudiantes de un curso de estadística participan en el siguiente experimento:

    a) Primero, completan un cuestionario con información personal (`Altura`, `Peso`, etc.).

    b) A continuación, se toman el pulso en reposo (`Pulso1`).

    c) Luego, lanzan una moneda al aire. Si sale cara, deben correr durante un minuto sin moverse del lugar que ocupan; si sale cruz, deben permanecer quietos.

d) Transcurrido el minuto, todos los estudiantes se vuelven a tomar el pulso (`Pulso2`) y anotan si han corrido (`1`) o no (`2`) en la variable `Correr`.

Este experimento se ha repetido durante varios años y los resultados, junto con la información personal de los estudiantes, están almacenados en el archivo `pulso.sav`.

a) Abre el conjunto de datos en *jamovi*.

b) Revisa la ventana de datos y la ventana de variables. Cambia el nombre de la variable `Anno` por `Año`.

c) Asigna etiquetas adecuadas a las variables y configura su tipo de medida correctamente:

- `Altura`: Altura en centímetros.
- `Peso`: Peso en kilogramos.
- `Edad`: Edad de los estudiantes.
- `Sexo`: Sexo.
- `Fumador`: Hábito de fumar.
- `Alcohol`: Hábito de beber.
- `Ejercicio`: Regularidad en la realización de ejercicio físico.
- `Correr`: Indica si el estudiante ha corrido.
- `Pulso1`: Pulso antes de la actividad.
- `Pulso2`: Pulso después de la actividad.
- `Año`: Año en que se realizó la medición.

d) Define las correspondencias entre valores y etiquetas:

- `Sexo`: `1=hombre`, `2=mujer`.
- `Fumador`: `1=sí`, `2=no`.
- `Alcohol`: `1=sí`, `2=no`.
- `Ejercicio`: `1=alta`, `2=moderada`, `3=baja`.
- `Correr`: `1=sí`, `2=no`.

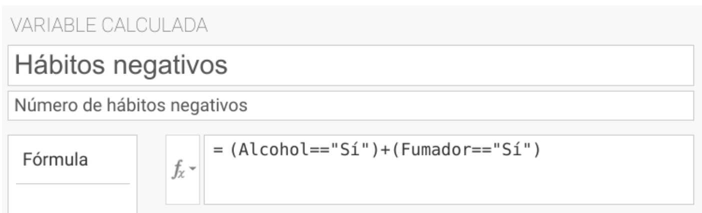
e) Examina el peso de los casos 77 y 78. ¿Notas alguna anomalía?

f) Revisa el caso 76. ¿Observas algo inusual?

g) Crea una nueva variable `Hábitos negativos`, que tome los valores:

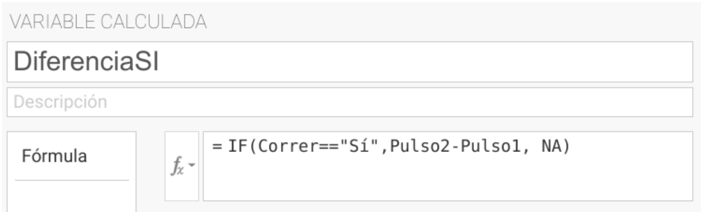
- `2` si la persona fuma y bebe.
- `1` si solo cumple una de estas condiciones.
- `0` si no fuma ni bebe.

Utiliza la función de cálculo en *jamovi*.



h) Crea la variable `Diferencia`, que almacene la diferencia entre `Pulso2` y `Pulso1`.

i) Genera la variable `DiferenciaSI`, que contenga la diferencia de pulso solo para aquellos individuos que hayan corrido (`Correr` = 1). Para ello, usa la función `IF`, que permite realizar cálculos condicionales, y el valor `NA` (*Not Available*) para representar datos faltantes.



j) Verifica que la variable `DiferenciaSI` tiene valores perdidos (*missing*) en los estudiantes que no han corrido.

k) Calcula el índice de masa corporal (IMC) como:

$$\text{IMC} = \frac{\text{Peso (kg)}}{\text{Altura (m)}^2}$$

1) Obtén el IMC para los individuos del fichero.
2) Según la *Organización Mundial de la Salud* (OMS), una persona mayor de 18 años se clasifica en:

Categoría	**IMC**
Infrapeso	< 18.5
Normal	$18.5 - 24.9$
Sobrepeso	≥ 25

3) Crea una nueva variable `IMCcod` con las siguientes categorías:
   - `1` = Infrapeso (`IMC` < 18.5).
   - `2` = Normal ($18.5 \leq$ `IMC` < 25).
   - `3` = Sobrepeso (`IMC` ≥ 25).

l) Guarda los datos modificados en formato nativo de *jamovi* con el nombre `pulsonuevo.omv` y en formato de hoja de cálculo *Excel* con el nombre `pulsonuevo.xlsx`.

2. El archivo `camisetas.csv` contiene 5 variables (`Identificacion`, `años`, `meses`, `altura` y `camiseta`). Ábrelo en *jamovi* y revisa su estructura.

a) Cambia el nombre de la variable `Identificacion` por `caso` y ajusta su tipo de variable.

b) Configura adecuadamente los tipos y nombres de las variables.

c) Crea una nueva variable `edad` en años, calculada como `años + meses/12`, utilizando la opción `Calcular` en el menú `Datos`.

d) Observa que la variable `camiseta` es de tipo texto. Recodifícala en una nueva variable numérica `talla` con los valores:

- 1 si `camiseta` es `s`.
- 2 si es `m`.
- 3 si es `l`.
- 4 si es `xl`.

3. El archivo `lorzas.sav` contiene información sobre la acumulación de grasa en diferentes partes del cuerpo en adolescentes [2]. A partir de estos datos:

    a) Abre `lorzas.sav` en *jamovi*.

    b) Selecciona únicamente las variables `ID`, `Edad`, `Sexo`, `Peso` y `Altura`.

    c) Guarda el nuevo archivo en formato `.csv` con el nombre `lorzas_filtrado.csv`.

“Si no puedo dibujarlo, es que no lo entiendo.”

Albert Einstein

2

Estadística descriptiva univariante

2.1. Motivación de la estadística descriptiva

Las técnicas de la estadística descriptiva permiten resumir y organizar la información contenida en una base de datos para facilitar su interpretación. Se utilizan para describir una muestra y analizar los datos, proporcionando información sobre la distribución de las variables, su tendencia central y su dispersión.

Este análisis también es un paso previo esencial para la estadística inferencial, ya que permite identificar patrones, detectar valores atípicos y verificar supuestos antes de aplicar modelos estadísticos más complejos. En particular, cuando no se dispone de un modelo predefinido para describir las relaciones entre variables, la estadística descriptiva ayuda a formular hipótesis iniciales y orientar el análisis posterior.

Las conclusiones obtenidas en este tipo de análisis son aplicables exclusivamente a la muestra estudiada. Para extrapolarlas a una población mayor, es necesario garantizar que la muestra sea representativa del colectivo de interés. Una muestra sesgada o no aleatoria puede llevar a conclusiones erróneas, por lo que es fundamental evaluar la calidad de los datos antes de realizar cualquier interpretación.

2.2. Validación de la base de datos

Antes de proceder con el análisis de los datos, es necesario validar la base de datos para garantizar su calidad y fiabilidad. Este proceso permite detectar posibles errores, asegurar la coherencia de las respuestas y garantizar que el análisis estadístico se base en datos correctos.

- **Revisión de las variables**: Identificar valores fuera de rango en las variables cuantitativas y comprobar la distribución de los datos mediante estadísticas descriptivas y gráficos. En el caso de variables categóricas, generar tablas de frecuencias ayuda a detectar inconsistencias o errores en la codificación.

- **Verificación de la coherencia interna**: En encuestas o estudios con múltiples variables relacionadas, es importante revisar la congruencia de las respuestas. Por ejemplo, si un individuo marca una opción en una pregunta, debe haber consistencia en otras respuestas asociadas.

Para llevar a cabo este proceso en *jamovi*, se recomienda utilizar el módulo `surveymv`, que permite realizar una exploración visual de los datos y detectar patrones inusuales. Para instalarlo, sigue estos pasos:

1. Ir a `Módulos` → `Biblioteca jamovi`.
2. Buscar `surveymv` y hacer clic en `Instalar`.
3. Una vez instalado, se añadirá un nuevo submenú dentro de `Análisis` → `Exploración`, desde donde se pueden seleccionar las variables de interés y generar gráficos para analizar su distribución por grupos. Ver Figura 2.1.

Ejemplo 2.1. Validación inicial de los datos en `pulsoProcesado.omv`.

1. Se genera un gráfico de exploración para todas las variables de la base de datos.

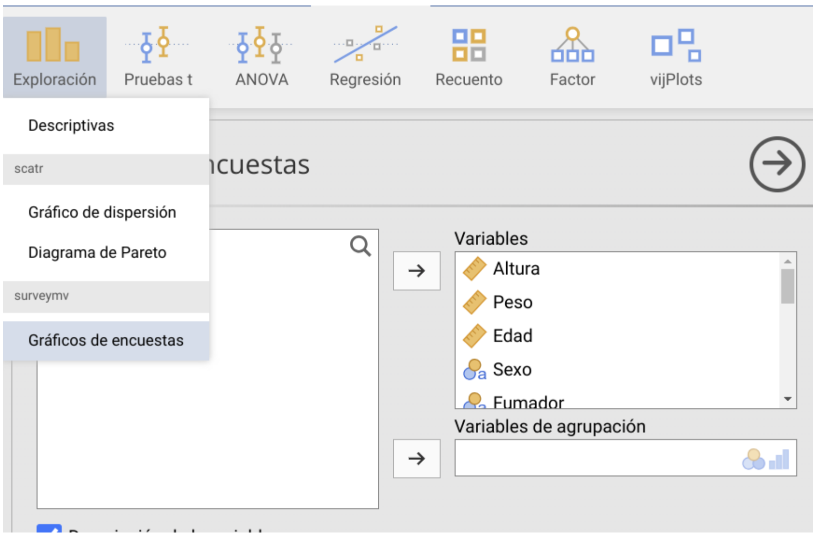


Figura 2.1: Exploración inicial de los datos en *jamovi* utilizando el módulo `surveymv`.

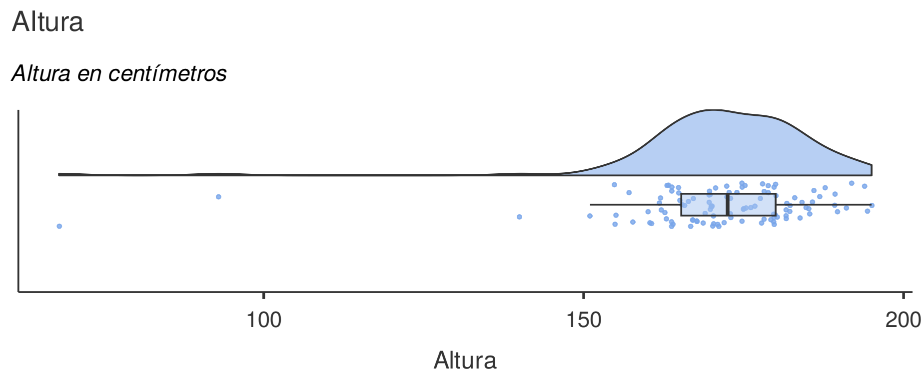


Figura 2.2: Detección de valores inusuales en la variable `Altura`.

2. Se revisa que los valores sean coherentes. En particular, se observa en la Figura 2.2 que algunos valores de altura parecen poco razonables, ya que hay al menos dos registros con alturas inverosímiles.

3. Se debe tomar una decisión respecto a estos valores atípicos. Una posible solución es aplicar un filtro para excluir estas observaciones, lo cual solo se recomienda si la variable fuera de rango es un elemento clave en el estudio. Otra opción es eliminar únicamente los valores erróneos sin excluir el resto de la observación.

   Para eliminar solo los valores incorrectos, se puede crear una nueva variable a partir de la original, conservando únicamente los valores superiores a 100 cm tal como se ilustra en la Figura 2.3.

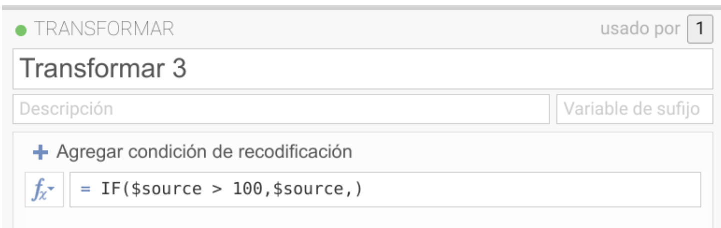


Figura 2.3: Creación de la variable `AlturaProcesada` para corregir valores atípicos.

4. Una vez calculada la nueva variable `AlturaProcesada`, se recomienda duplicarla, copiando y pegando, y eliminar tanto la variable original `Altura` como `AlturaProcesada` para evitar la duplicación de variables en la base de datos.

Nota 2.2. Revisa la variable `peso` y ajusta los valores para mantener únicamente los pesos creíbles.

2.3. Variables cualitativas nominales

Las variables cualitativas nominales representan categorías sin un orden inherente. Para resumir su información, se emplean tablas de frecuencias y gráficos como los diagramas de barras y los gráficos de sectores.

- **Tablas de frecuencias**: Permiten observar la distribución de cada categoría en términos absolutos y porcentuales. En el caso de variables nominales, el porcentaje acumulado no es informativo.
- **Gráficos de barras**: Representan visualmente la proporción de cada categoría. Los diagramas de barras son la opción más común, mientras que los gráficos de sectores circulares son útiles cuando la variable tiene pocas categorías.
- **Medidas de tendencia central**: La única medida que se puede calcular en variables nominales es la **moda**, que identifica la categoría más frecuente. Puede observarse directamente en la tabla de frecuencias o calcularse mediante *jamovi* si los datos se han codificado numéricamente.

Para analizar variables cualitativas en *jamovi*, se sigue el siguiente procedimiento:

`Análisis` → `Exploración` → `Descriptivas`

A continuación, se ilustrará su uso con el archivo `pulsoProcesado.omv`.

1. Accede al menú `Análisis` y seleccionar `Exploración` → `Descriptivas`.
2. Selecciona la variable cualitativa de interés y arrástrala al cuadro de `Variables`.
3. Marca la opción `Tablas de frecuencias` para obtener conteos y porcentajes. Ver Figura 2.4.

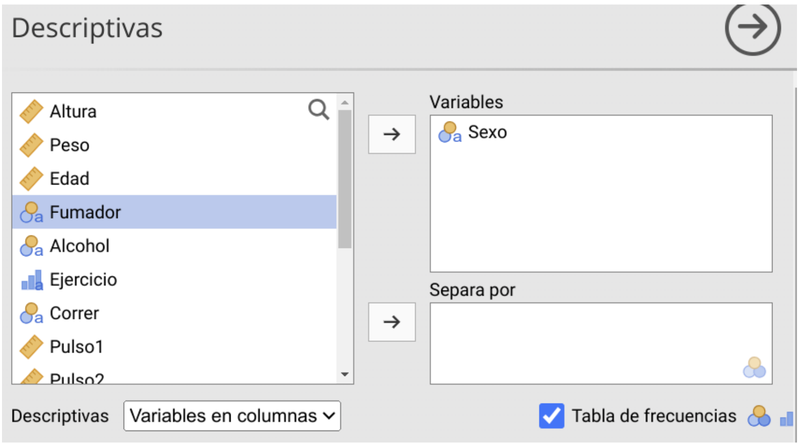


Figura 2.4: Opciones de análisis de tablas de frecuencias en *jamovi*.

4. Seleccionar las medidas de tendencia central adecuadas para la variable cualitativa nominal. Ver Figura 2.5.
5. Para visualizar la distribución de la variable, ir a la pestaña `Gráficos` y seleccionar `Diagrama de barras`.

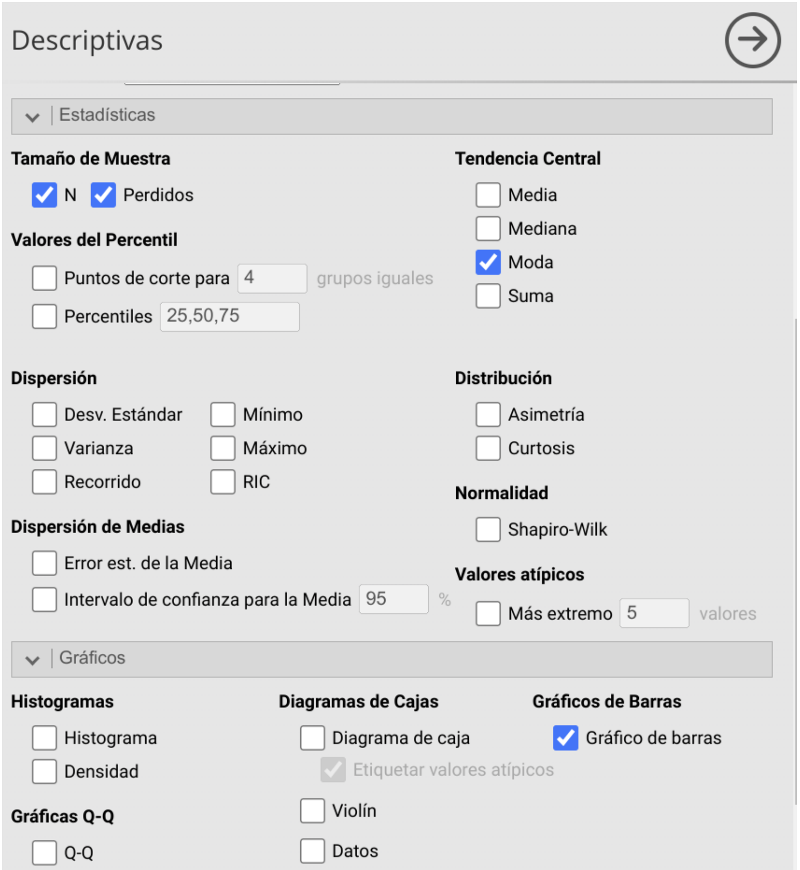


Figura 2.5: Selección de medidas de síntesis y opciones gráficas para variables cualitativas nominales.

Tras realizar estos pasos, *jamovi* generará automáticamente la tabla de frecuencias y los gráficos correspondientes, que se visualizarán en el área de resultados de la derecha y podrán copiarse o exportarse directamente.

Sexo	**N**
Total	110
Perdidos	0

Tabla 2.1: Resumen de los registros válidos y perdidos de la variable `Sexo`.

En este caso, el análisis de la variable `Sexo` revela que la muestra está compuesta por 110 individuos y no presenta valores perdidos (ver Tabla 2.1). Se observa una distribución equilibrada entre categorías, aunque con una ligera predominancia de los hombres (53.6 %) sobre las mujeres (46.4 %) (ver Tabla 2.2).

Como se muestra en la Figura 2.6, el gráfico de barras representa visualmente la proporción de cada categoría. La tabla de frecuencias proporciona un desglose detallado de la composición de la muestra, indicando los porcentajes absolutos y acumulados. Sin embargo, en variables nominales, el porcentaje acumulado no es relevante.

Tanto en la tabla de frecuencias como en el gráfico, se observa que la categoría más

Sexo	**Frecuencia**	**% del Total**	**% Acumulado**
Mujer	51	46.4 %	46.4 %
Hombre	59	53.6 %	100.0 %

Tabla 2.2: Distribución de la variable `Sexo`.

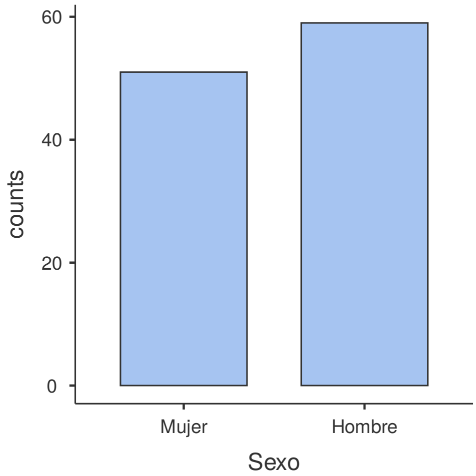


Figura 2.6: Gráfico de barras de la variable `Sexo`.

frecuente es `Hombre`, lo que indica que esta es la moda de la variable.

2.4. Variables cualitativas ordinales

En esta sección se describe el análisis de variables cualitativas ordinales, que son aquellas cuyas categorías poseen un orden inherente o natural. Un ejemplo es la variable `Regularidad en la realización de ejercicio físico`, que distingue entre los niveles *baja*, *moderada* y *alta*.

Para este tipo de variables, las herramientas más adecuadas incluyen, de manera similar a las variables nominales:

- **Tablas de frecuencias**: A diferencia de las variables nominales, en este caso los porcentajes acumulados son informativos, ya que permiten evaluar la distribución en los extremos de la variable con facilidad.
- **Gráficos de barras**: Es fundamental que el gráfico respete el orden de las categorías. El diagrama de barras es la representación más adecuada para este tipo de variables. En contraste, los gráficos de sectores no reflejan correctamente la jerarquía de las clases y, por lo tanto, no son recomendables.
- **Medidas de tendencia central**: Se recomienda utilizar la **mediana** como estadístico de posición y la **moda** para identificar la categoría más frecuente. Ambas se pueden obtener a partir de la tabla de frecuencias.

El procedimiento en *jamovi* es el mismo que para las variables cualitativas nominales:

`Análisis → Exploración → Descriptivas`

Una vez completados estos pasos, se seleccionan las mismas opciones que para las variables cualitativas nominales y *jamovi* generará automáticamente los resultados en la sección de salida.

Regularidad en la realización de ejercicio físico	N
Válido	110
Perdidos	0

Tabla 2.3: Resumen de los registros válidos y perdidos de la variable `Regularidad en la realización de ejercicio físico`.

Regularidad en la realización de ejercicio	Frecuencia	Porcentaje del Total	Porcentaje Acumulado
Baja	37	33,6 %	33,6 %
Moderada	59	53,6 %	87,3 %
Alta	14	12,7 %	100,0 %
Total	110	100,0 %	100,0 %

Tabla 2.4: Distribución de la variable `Regularidad en la realización de ejercicio físico`.

Como se observa en las Tablas 2.3 y 2.4, los resultados indican lo siguiente:

- No hay datos perdidos, se tiene el valor de la variable para los 110 individuos.
- La categoría *Moderada* es la más frecuente, representando un 53,6 % de la muestra.
- Un 33,6 % de los individuos realiza ejercicio de manera esporádica o con baja frecuencia.
- Solo el 12,7 % (100 % − 87,3 %) de la muestra reporta una práctica de ejercicio regular y alta.
- La moda de los datos coincide con la categoría *Moderada* (frecuencia o porcentaje más elevado), mientras que la mediana, al ser una variable ordinal, también se encuentra en esta categoría (primer acumulado que supera el 50 %).

Para visualizar gráficamente la distribución de la variable, se utiliza un diagrama de barras, mostrado en la Figura 2.7. Como se puede apreciar en el gráfico, la representación respeta el orden de las categorías, lo que garantiza una correcta interpretación de la información. A diferencia de los gráficos de sectores, este tipo de gráfico permite visualizar claramente la progresión en la regularidad de la práctica de ejercicio.

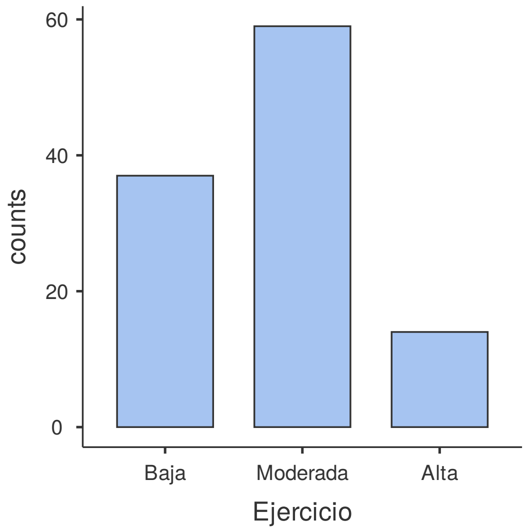


Figura 2.7: Gráfico de barras de la variable `Regularidad en la realización de ejercicio físico`.

2.5. Más allá de los gráficos por defecto

2.5.1. Gráficos con porcentajes

Cuando se trabaja con una muestra y se desea extrapolar los resultados o compararlos con otros estudios, es recomendable expresar los gráficos en términos de porcentajes en lugar de valores absolutos. Esto permite una interpretación más clara y facilita la comparación entre estudios con diferentes tamaños muestrales. Sin embargo, *jamovi* no incluye esta opción de forma predeterminada en sus gráficos de barras.

Para solucionar esta limitación, se pueden utilizar módulos adicionales que permiten personalizar la visualización de los datos.

En particular, el módulo `surveymv` permite generar gráficos de barras horizontales, lo que puede resultar útil cuando se trabaja con variables cualitativas ordinales con etiquetas largas, ya que mejora la legibilidad. Un ejemplo de este tipo de visualización se muestra en la Figura 2.8.

Para generar este gráfico utilizando el módulo `surveymv`, sigue estos pasos:

1. Ve al menú `Exploración` → `Gráficos de encuestas`.
2. Arrastra la variable de interés, en este caso `Regularidad en el ejercicio`, al cuadro de variables.
3. En el submenú `Tipo de frecuencia`, selecciona *Porcentajes* para visualizar los valores en términos relativos en lugar de absolutos.

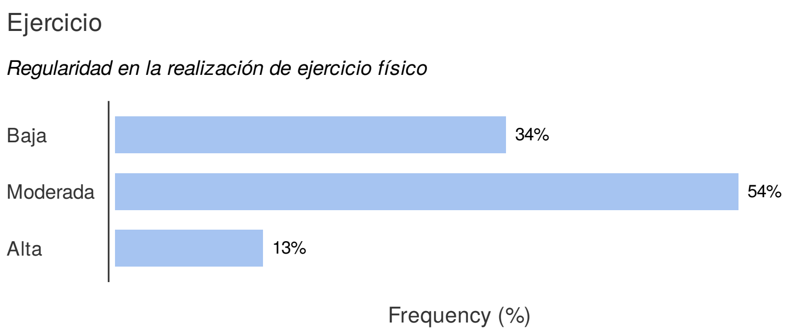


Figura 2.8: Gráfico de barras horizontales con el módulo `surveymv`, representando los valores en porcentaje.

Otra opción para generar gráficos con porcentajes es el módulo `vijPlots`, que permite visualizar los datos en términos porcentuales y personalizar su presentación. Un ejemplo generado con este módulo se muestra en la Figura 2.11.

Para construir este gráfico, sigue los pasos siguientes:

1. Haz clic en la imagen del módulo `vijPlots` en la biblioteca de módulos.

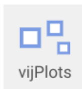


Figura 2.9: Módulo `vijPlots` en la biblioteca de *jamovi*.

2. Selecciona la opción `Bar Plot` (diagrama de barras en inglés) y completa las opciones del recuadro de configuración de la siguiente manera:

   - En *Category Axis*, selecciona la variable que se desea representar.
   - En *Y-axis (without group)*, selecciona *Percent* para que los valores se expresen en porcentaje.
   - Se pueden modificar otras opciones, como el orden de los niveles, aunque generalmente no es necesario si estos han sido organizados previamente en el preprocesamiento de los datos. También se puede optar por una representación horizontal de los datos.

Una vez configuradas las opciones, el gráfico generado se muestra en la Figura 2.11.

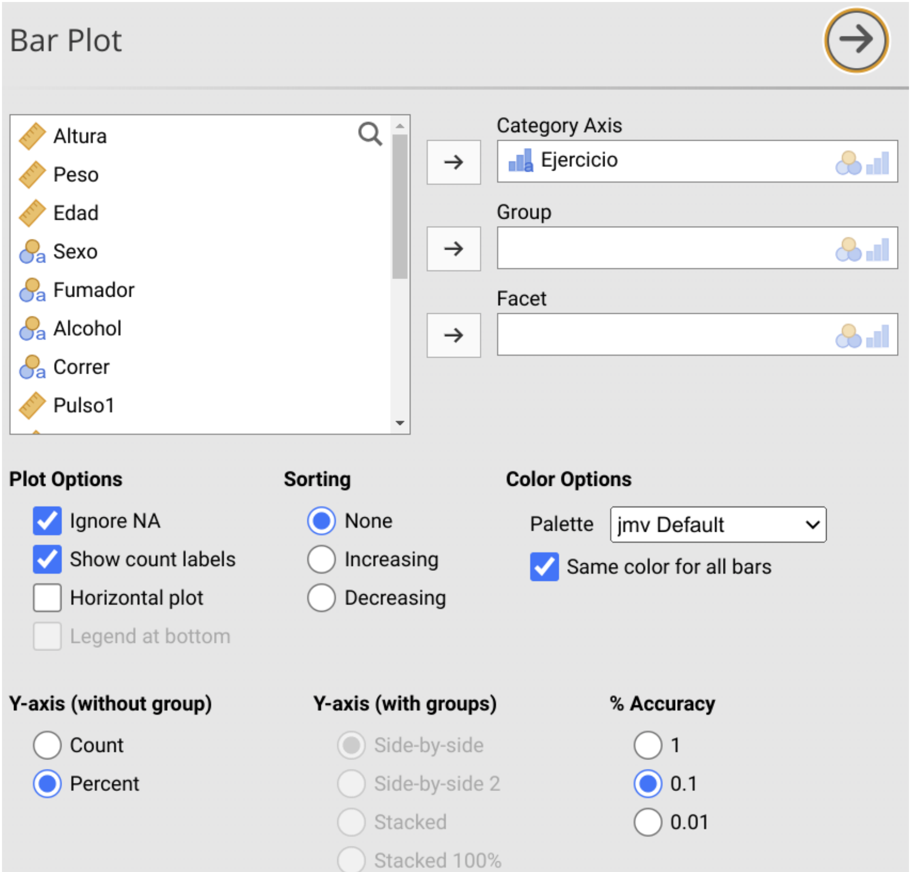


Figura 2.10: Configuración de `vijPlots` para generar gráficos de barras con porcentajes.

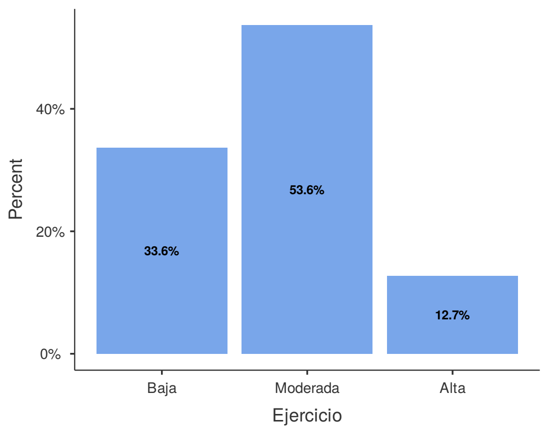


Figura 2.11: Gráfico alternativo de barras horizontales con el módulo `vijPlots`.

2.5.2. Diagramas de sectores

Los diagramas de sectores circulares (*pie charts*) son una alternativa a los diagramas de barras para representar variables cualitativas nominales. Este tipo de gráfico permite visualizar la proporción de cada categoría de manera clara, especialmente cuando el número de categorías es reducido. Son útiles para comparar rápidamente la contribución relativa de cada grupo dentro del total de la muestra. Sin embargo, cuando se tienen muchas categorías, los diagramas de barras suelen ser más adecuados, ya que facilitan la comparación de diferencias pequeñas. Véase [3] para una explicación sencilla pero detallada sobre la comparación de la efectividad de estos gráficos.

En la Sección 2.5.1, se explicó cómo representar variables categóricas mediante diagramas de barras utilizando el módulo `vijPlots`. Ahora, siguiendo un procedimiento similar, se puede generar un gráfico de sectores como el que se muestra en la Figura 2.12.

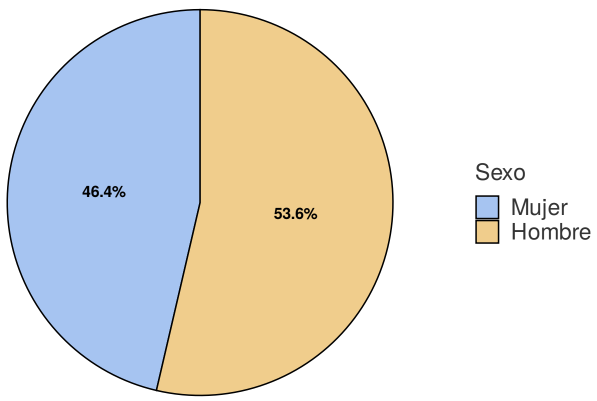


Figura 2.12: Gráfico de sectores generado con el módulo `vijPlots`.

1. Acceder al módulo `vijPlots` haciendo clic en su icono en la barra de herramientas.

2. Seleccionar la opción `Pie Chart` en el menú de tipos de gráficos.

3. Configurar las opciones del gráfico. En la Figura 2.13, se muestran los ajustes recomendados:

   - En el campo `Variable`, selecciona la variable categórica que se desea representar. En este caso, se utiliza `Regularidad en el ejercicio`.
   - En la sección `Labels`, selecciona la opción `Percent` para que los datos se visualicen en términos de porcentaje en lugar de conteos absolutos.
   - Para cambiar la disposición de la leyenda y colocarla en la parte inferior del gráfico, activa la opción `Legend at Bottom`.
   - Si se desea personalizar la apariencia del gráfico, se pueden ajustar opciones estéticas como la paleta de colores o agregar bordes a las secciones del gráfico.
   - Para obtener un gráfico con estilo de *donut*, se debe activar la opción `I prefer Donuts!`.

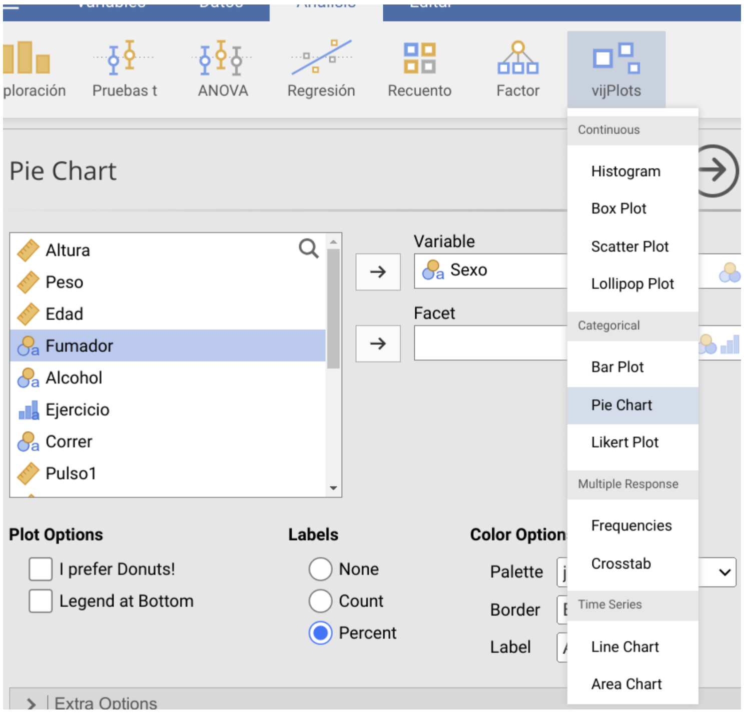


Figura 2.13: Opciones en el módulo `vijPlots` para generar un gráfico de sectores con porcentajes.

2.6. Variables cuantitativas

Las variables cuantitativas representan magnitudes numéricas y pueden tomar múltiples valores distintos. En particular, se debe, de algún modo, abordar las siguientes cuestiones:

1. ¿Cuáles son los valores más frecuentes, típicos o centrales?
2. ¿Cómo se distribuyen los valores respecto a este centro? ¿Hay mucha variabilidad? ¿Están muy concentrados o dispersos?

Para obtener un resumen de estas variables en *jamovi*, se debe seguir el mismo procedimiento que con las variables cualitativas, pero seleccionando variables numéricas:

`Análisis` → `Exploración` → `Descriptivas`

2.6.1. Medidas de localización

Las medidas de localización permiten describir el centro de la distribución de los datos. Las más utilizadas son:

- **Media** ($\bar{X}$): Representa el centro de masas de la distribución.
- **Mediana**: Es el valor central de la distribución, dividiendo los datos en dos partes iguales.
- **Moda**: Indica el valor más frecuente en los datos.

En distribuciones simétricas, la media y la mediana coinciden. Sin embargo, en distribuciones asimétricas, la mediana suele ser más representativa, ya que la media es más sensible a valores extremos.

Otras medidas de posición incluyen los **percentiles** o **cuartiles**, que dividen la distribución en partes iguales:

- Q_1 (percentil 25): Primer cuartil, indica el valor por debajo del cual se encuentra el 25 % de los datos.
- Q_2 (percentil 50): Coincide con la mediana.
- Q_3 (percentil 75): Indica el valor por debajo del cual se encuentra el 75 % de los datos.

Para obtener estas medidas en *jamovi*, se debe acceder a la pestaña `Tendencia Central` y marcar las siguientes opciones, como se ilustra en la Figura 2.14:

- `Media`, `Mediana` y `Moda`.
- En `Valores del Percentil`, seleccionar `Percentiles` e ingresar los valores 25, 50, 75, que aparecen por defecto si no se modifican.

2.6.2. Medidas de dispersión

Las medidas de dispersión cuantifican cómo se distribuyen los valores respecto a la media. Las principales son:

- **Rango**: Diferencia entre el valor máximo y mínimo. Indica cuánto pueden variar como mucho los datos.
- **Rango intercuartílico (RIC o IQR por sus siglas en inglés)**: Diferencia entre el tercer y el primer cuartil ($Q_3 - Q_1$). Representa la dispersión, cuánto pueden variar como mucho, el 50 % central de los datos.
- **Varianza** (S^2): Mide la dispersión de los datos respecto a la media. Se reporta pero no es fácilmente interpretable, ya que sus unidades son el cuadrado de las unidades de la variable original.

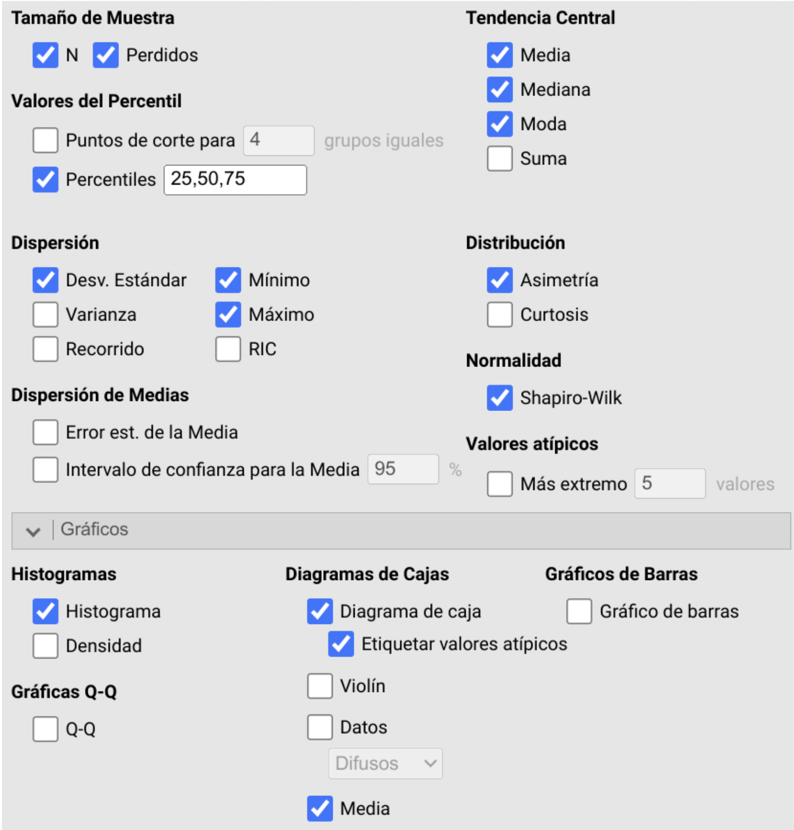


Figura 2.14: Opciones para obtener las medidas de síntesis y gráficas adecuadas para variables cuantitativas.

- **Desviación estándar** (s): Raíz cuadrada de la varianza. Se expresa en las mismas unidades que los datos originales y es la medida de dispersión más utilizada.
- **Error estándar de la media**: Se calcula como $\frac{s}{\sqrt{n}}$ y mide la precisión de la media muestral.
- **Coeficiente de Variación** (CV): Se calcula como $s/\bar{X}$, y no lo reporta *jamovi* directamente. Es una medida de variabilidad en relación a la media.

Para obtener estas medidas en *jamovi*, se deben seleccionar las opciones correspondientes en la pestaña `Dispersión`, como se muestra en la Figura 2.14.

2.6.3. Forma de la distribución

Adicionalmente, se pueden calcular medidas que describen la forma de la distribución:

- **Asimetría**: Calcula el Coeficiente de Asimetría de Fisher. Indica si la distribución está sesgada hacia la derecha (asimetría positiva) o hacia la izquierda (asimetría negativa).

- **Curtosis**: Mide el grado de apuntamiento de la distribución. Una curtosis alta indica colas más gruesas que la distribución normal, mientras que una curtosis baja indica colas más delgadas.

Para comprobar si la distribución se ajusta a un modelo normal, se puede realizar la prueba de Shapiro-Wilk, que devuelve un valor p. Si $p < 0{,}05$, se rechaza la hipótesis de normalidad, lo que sugiere que la distribución no sigue una normalidad perfecta. Este concepto se estudiará con más detalle en el Capítulo 4.

Las opciones para calcular estos estadísticos se encuentran en la pestaña `Forma de la distribución` y se ilustran de nuevo en la Figura 2.14.

Ejemplo: análisis de la variable `Pulso1`

A continuación, se presentan los estadísticos descriptivos obtenidos para la variable `Pulso1`. En la Tabla 2.5 se muestran las medidas de localización, mientras que la Tabla 2.6 recoge las medidas de dispersión y forma de la distribución.

Pulso1	**Valor**
N	109
Perdidos	1
Media	75.69
Mediana	76.00
Moda	76.00[a]

Tabla 2.5: Medidas de localización para la variable `Pulso1`.

Pulso1	**Valor**
Desviación estándar	13.30
Mínimo	47.00
Máximo	145.00
Rango intercuartílico (IQR)	14.00
Asimetría	1.512
Error estándar de asimetría	0.231

Tabla 2.6: Medidas de dispersión y forma de la distribución para `Pulso1`.

Prueba de normalidad	**Valor**
W de Shapiro-Wilk	0.904
Valor p	$< ,001$

Tabla 2.7: Prueba de normalidad de Shapiro-Wilk para `Pulso1`.

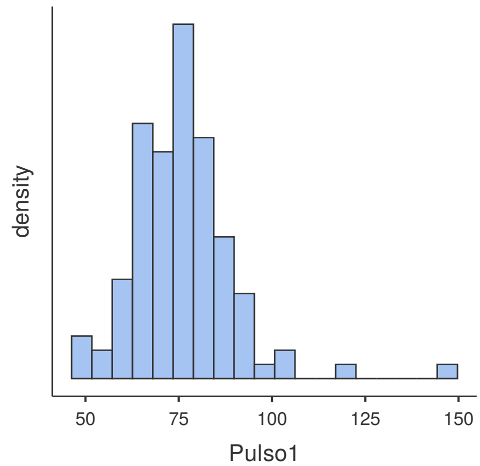


Figura 2.15: Histograma de la variable `Pulso1`.

Interpretación

- El pulso medio es de 75,7 pulsaciones por minuto, con una desviación típica de 13,3 pulsaciones por minuto. El rango de variación del pulso es de 98 pulsaciones por minuto, desde un mínimo de 47 hasta un máximo de 145 (ver Tabla 2.6).
- Al menos el 50 % de los individuos tiene un pulso menor o igual (o mayor o igual) a 76, que es la mediana. Asimismo, al menos el 25 % presenta pulsaciones menores o iguales que 68 y al menos otro 25 % presenta pulsaciones mayores o iguales que 82 (ver Tabla 2.5).
- El coeficiente de asimetría es de 1,5, lo que indica una distribución sesgada hacia la derecha. Su magnitud es grande en comparación con el error estándar de asimetría (0,23) (ver Tabla 2.6).
- Como se observa en la Figura 2.15, el histograma permite intuir la forma de la distribución y proporciona una representación visual complementaria a los estadísticos de forma. En este caso, parece relativamente normal, con valores alrededor de la media, salvo por algunos casos atípicos con pulsaciones elevadas.
- La prueba de Shapiro-Wilk (Tabla 2.7) devuelve un valor $p < 0{,}001$, lo que confirma que la distribución no sigue una normalidad estadística.

Nota 2.3. El rango intercuartílico es una medida de dispersión más robusta ante valores atípicos que la desviación estándar.

Los resultados indican que la distribución de `Pulso1` es moderadamente asimétrica, con la mayoría de los individuos presentando pulsaciones entre 68 y 82 ppm. Se observa la presencia de valores atípicos en el extremo superior. Para complementar este análisis, es recomendable utilizar diagramas de cajas, los cuales permiten visualizar mejor la dispersión de los datos y detectar valores extremos.

2.6.4. Valores extremos y detección de errores

En la Figura 2.14, se observa que, dentro de la opción `Valores atípicos`, *jamovi* permite mostrar los puntos `Más extremos`. Estos valores pueden ser útiles para detectar posibles errores en la base de datos y se presentan en la Tabla 2.8.

En este caso, los valores de 145 (caso 73) y 119 (caso 106) pulsaciones por minuto pueden indicar errores en la medición o manipulación de los datos. Sin embargo, antes de descartar estas observaciones, es importante comprender bien la naturaleza del experimento y su contexto. Como parte del proceso de validación de la base de datos, descrito en la Sección 2.2, se debe analizar si estos registros deben eliminarse o corregirse.

Nota 2.4. Como se ha indicado a la vista de la Figura 2.15 en la Sección 2.6.3, la asimetría hacia la derecha no es tan evidente si se eliminan estos valores (posiblemente erróneos), lo

Valores extremos de `Pulso1`		
Categoría	**Número de fila**	**Valor**
Máximos	73	145
	106	119
	40	104
	46	104
	3	96
Mínimos	61	47
	71	49
	62	50
	82	52
	55	56

Tabla 2.8: Valores extremos de la variable `Pulso1`.

que sugiere que su presencia influye en la distribución observada.

2.6.5. Modificación del histograma

El módulo por defecto de *jamovi* determina automáticamente la posición y la anchura de los intervalos de clase al generar un histograma. No obstante, en ciertos casos, puede ser necesario definir intervalos más amplios o más estrechos para mejorar la representación de los datos.

Para modificar esta configuración, se utiliza el módulo `vijPlots`, siguiendo estos pasos:

1. Acceder al módulo `vijPlots` haciendo clic en su icono en la barra de herramientas.
2. Seleccionar la opción `Histogram`.
3. Arrastrar la variable `Pulso1` al recuadro de variables.
4. Configurar el punto inicial del histograma en 40 y definir una anchura de intervalos de 10 (Figura 2.16).

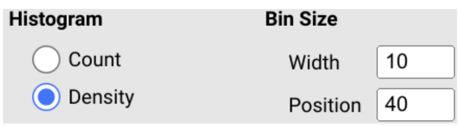


Figura 2.16: Configuración del histograma de `Pulso1`.

El resultado es un histograma con clases más naturales (Figura 2.17). Dependiendo del nivel de detalle deseado, pueden ajustarse otros parámetros para modificar la visualización de los datos.

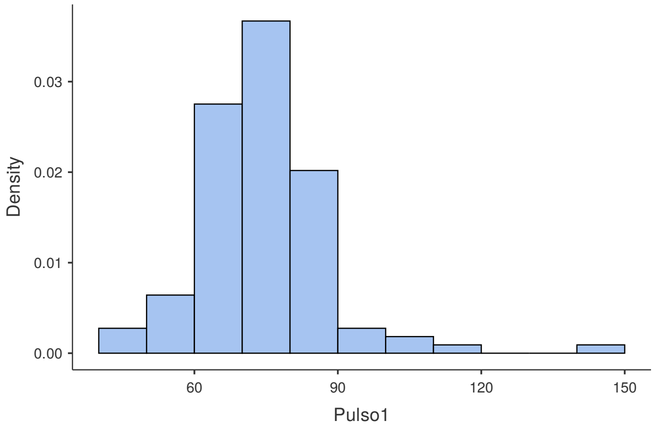


Figura 2.17: Histograma de `Pulso1` con clases personalizadas.

2.6.6. Diagrama de caja

El diagrama de caja permite visualizar la mediana, los percentiles 25 y 75, el mínimo, el máximo y los valores atípicos. Los valores extremos se marcan con un signo • si son considerados *atípicos*.

Interpretación del diagrama de caja
Los valores de la variable se representan en el eje Y:

- La tapa superior de la caja corresponde al percentil 75.
- La tapa inferior representa el percentil 25.
- La línea central más gruesa indica el percentil 50 (mediana).

En la Figura 2.18, se observa el diagrama de caja generado para la variable `Pulso1`. Se identifican como valores atípicos los siguientes casos:

- Caso 73, con 145 pulsaciones por minuto.
- Caso 106, con 119 pulsaciones por minuto.
- Casos 40 y 46, ambos con 104 pulsaciones por minuto.
- No se detectan valores atípicos en la parte baja de la distribución.

El rango de valores no atípicos está representado por los *bigotes* del diagrama, que se extienden desde 47 (mínimo) hasta 96 (el quinto valor más alto, quitando los atípicos, según la Tabla 2.8).

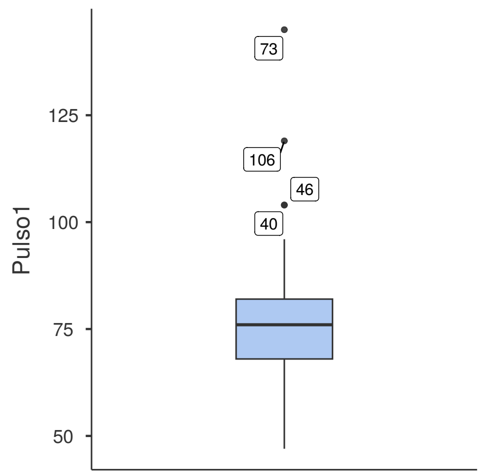


Figura 2.18: Diagrama de cajas de `Pulso1`.

Consideraciones sobre la presentación de gráficos

Tanto en los histogramas como en los diagramas de caja, para una presentación formal de los gráficos, es recomendable incorporar las unidades en el eje correspondiente. En este caso, la unidad adecuada sería *ppm* (*pulsaciones por minuto*).

Sin embargo, *jamovi* no permite añadir unidades en los ejes de forma nativa. Para incluirlas, sería necesario utilizar un módulo adicional o editar el gráfico manualmente en un software externo.

2.7. Gestión de datos ya tabulados: ponderación de casos

En algunas ocasiones, no se dispone de un fichero con los datos individuales de la muestra, sino únicamente de una tabla de frecuencias que resume esta información.

En los archivos `hijos.sav` y `hijos.xlsx` se encuentra la misma información sobre el número de hijos por pareja. Al abrir ambos ficheros en *jamovi*, se puede observar que `hijos.sav` contiene los microdatos, es decir, una fila para cada pareja, mientras que `hijos.xlsx` presenta los datos en forma de tabla de frecuencias, como se muestra en la Tabla 2.9.

Para analizar estos datos en *jamovi*, una opción sería reconstruir el fichero con los microdatos, escribiendo 37 veces el valor 0, 76 veces el valor 1, etc. Sin embargo, un método más eficiente es utilizar la opción de *ponderación de casos*.

1. Verificar que la variable que representa la ponderación, en este caso `frecuencias`, esté configurada como una variable continua. Si no lo está, modificar su tipo de

Hijos	**Frecuencia**
0	37
1	76
2	68
3	53
4	42
5	15
6	6

Tabla 2.9: Tabla de frecuencias del archivo `hijos.xlsx`.

medida.

2. Acceder al menú `Datos` → `Ponderaciones` (Figura 2.19).

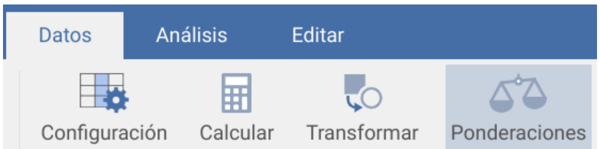


Figura 2.19: Menú de ponderación de casos en *jamovi*.

3. En la opción `Variable del peso`, seleccionar la variable que contiene el número de repeticiones de cada caso, en este caso `frecuencias` (Figura 2.20).

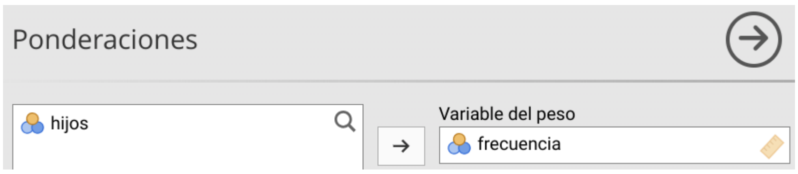


Figura 2.20: Selección de la variable de ponderación en *jamovi*.

4. Una vez aplicada la ponderación, en la ventana de resultados aparecerá un mensaje confirmando que los datos han sido ponderados correctamente (Figura 2.21).

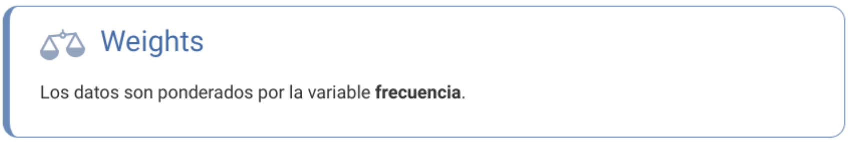


Figura 2.21: Confirmación de la aplicación de la ponderación en *jamovi*.

Al calcular las medidas descriptivas de la variable `hijos`, se observa en la Tabla 2.10 que se está trabajando con un total de 297 parejas.

Para visualizar la tabla de frecuencias, es necesario asegurarse de que la variable `hijos` no esté configurada como continua. Una vez verificado esto, se podrá generar una salida similar a la Tabla 2.11 siguiendo las instrucciones de la Sección 2.3.

Estadísticos	**Número de hijos por pareja**
Válido	297
Perdidos	0
Media	2.189
Mediana	2.00
Moda	1
Desviación típica	1.499
Asimetría	0.431
Error estándar de asimetría	0.141
Mínimo	0
Máximo	6
Percentiles	
25	1.00
50	2.00
75	3.00

Tabla 2.10: Estadísticos descriptivos para la variable `hijos`.

Hijos	**Frecuencia**	**% del Total**	**% Acumulado**
0	37	12.5	12.5
1	76	25.6	38.0
2	68	22.9	60.9
3	53	17.8	78.8
4	42	14.1	92.9
5	15	5.1	98.0
6	6	2.0	100.0
Total	297	100.0	100.0

Tabla 2.11: Tabla de frecuencias para la variable `hijos`.

Para generar un gráfico, se puede utilizar el módulo `vijPlots`, como se describió en la Sección 2.5.1. El resultado obtenido se muestra en la Figura 2.22.

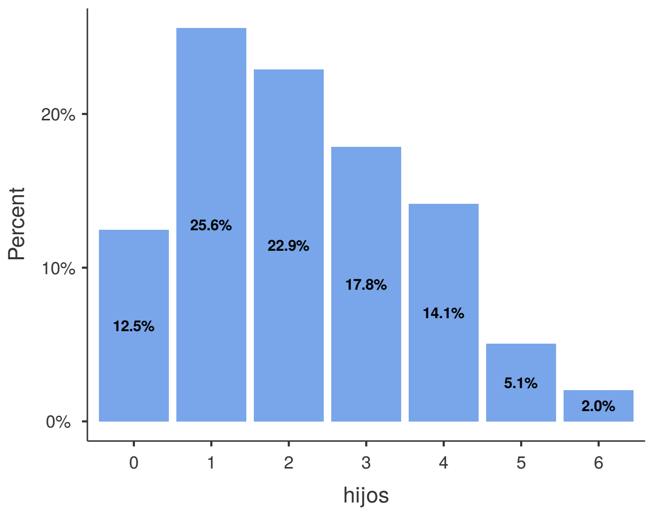


Figura 2.22: Diagrama de barras del número de hijos.

Comentarios sobre la distribución de la variable *Número de hijos*

- La mayoría de las 297 parejas encuestadas tiene un solo hijo, lo que representa la moda con un 25,6 % del total.
- La mediana (percentil 50) es 2, lo que indica que al menos la mitad de la muestra tiene, como máximo, 2 hijos, mientras que la otra mitad tiene, como mínimo, la misma cantidad.
- El percentil 25 es 1, lo que significa que al menos una cuarta parte de las parejas tiene un hijo o menos. El percentil 75 es 3, lo que implica que al menos el 75 % de las parejas tiene como mucho tres hijos.
- El porcentaje de familias numerosas (tres o más hijos) asciende al 39,1 %. La media es 2,19 hijos, un valor muy próximo a la mediana, lo que sugiere una distribución equilibrada.
- La desviación estándar es de 1,50 hijos. Para medir la variabilidad relativa con respecto a la media, se calcula el coeficiente de variación (CV) como:

$$CV = \frac{\text{Desviación estándar}}{\text{Media}} = \frac{1{,}50}{2{,}19} = 68\,\%.$$

  Este valor, al ser moderadamente grande, refleja una variabilidad elevada en la cantidad de hijos por pareja. Es decir, las parejas son muy desiguales en cuanto a la tenencia de hijos.
- El coeficiente de asimetría es positivo (0,43), indicando una distribución sesgada hacia la derecha, en concordancia con la visualización en la Figura 2.22.

2.8. Ejercicios

1. En el fichero `pulsoProcesado.omv`, realiza un análisis univariante de las siguientes variables:
   a) `Edad`.
   b) `Peso`.
   c) `Altura`.
   d) `IMC`.
   e) `IMC en niveles`.
   f) `Hábitos negativos` (modifica su tipo para poder generar un gráfico de barras).
   g) `Fumador`.
2. Analiza la variable `Pulso1` (pulso en reposo) en el fichero `pulsoProcesado.omv`. Incluye en el estudio solo los valores de `Pulso1` inferiores a 110. Examina su distribución mediante estadísticos, un histograma y un diagrama de caja. Evalúa si la variable es simétrica o presenta asimetría y analiza su apuntamiento (curtosis).

3. En un servicio de Traumatología, se busca conocer la distribución de las patologías (*lesión de rodilla, de cadera, de tobillo, de cráneo, otras*) de los pacientes atendidos en Urgencias en los últimos seis meses para facilitar la planificación del servicio. Los datos se encuentran en el archivo `patologia.sav`. Representa la distribución de las patologías mediante una tabla de frecuencias, identifica la moda y elabora un diagrama de sectores circulares.

   Nota 2.5. Observa cómo el gráfico de barras generado por el módulo predeterminado de *jamovi* para esta variable no es adecuado para representar correctamente la información. Explora la alternativa con los módulos `surveymv` y `vijPlots`.

4. En un servicio de Medicina Interna, se ha realizado un muestreo aleatorio del número de urgencias atendidas por día, obteniendo una muestra de 60 días al azar. Los datos se encuentran en el archivo `urgencias.sav`. Realiza un estudio descriptivo de la variable `Número de urgencias atendidas`, generando un histograma con clases adecuadas, un gráfico de caja y los estadísticos de posición, dispersión y forma.

   Además, en el apartado `Tendencia Central`, selecciona la opción `Suma`. Interpreta el significado de esta medida.

5. Analiza la distribución de la variable `Número de DVD` en el archivo `dvd.sav`, que contiene información sobre la cantidad de DVD que poseen 56 estudiantes. Examina su distribución mediante un histograma y un diagrama de caja. Comenta las estadísticas descriptivas de posición, dispersión y forma.

   Recomendación sobre medidas estadísticas
   En general, la media es la medida de tendencia central más utilizada, y la desviación típica, la medida de dispersión más adecuada para variables numéricas. Sin embargo, cuando existen valores atípicos o datos erróneos, se recomienda el uso de la mediana como medida de tendencia central y del rango intercuartílico como medida de dispersión.
   ¿Qué medidas considerarías más adecuadas en el caso de los DVD si se garantiza que los datos no contienen errores?

6. Un grupo de 100 profesores ha sido evaluado según un índice de rendimiento laboral, donde valores más altos indican mayor rendimiento. Realiza un análisis descriptivo de esta variable utilizando los estadísticos y gráficos que consideres adecuados. En la Tabla 2.12 se muestra la distribución de la variable.

Índice	**Frecuencia**
3	4
4	7
5	12
6	18
7	21
8	15
9	11
10	8
11	4

Tabla 2.12: Distribución del índice de rendimiento laboral de los profesores.

7. Describe la distribución de la variable `Altura` en el archivo `camisetas.csv`, utilizando los estadísticos de posición, dispersión y forma, así como los gráficos que consideres adecuados. Evalúa si la distribución de alturas se ajusta a un modelo normal.

"Felix qui potuit rerum cognoscere causas." Dichoso aquel que puede conocer las causas de las cosas.

Virgilio, Geórgicas II, 490. Año 29 A.C.

3

Estadística descriptiva bivariante

3.1. Introducción a la Estadística Descriptiva Bivariante

Antes de aplicar métodos de inferencia estadística, es fundamental comprender y describir adecuadamente las relaciones entre pares de variables en nuestra muestra. La **estadística descriptiva bivariante** permite explorar la asociación entre variables, detectar patrones y proporcionar una visión preliminar de los datos antes de realizar contrastes de hipótesis (Capítulo 4) o modelos predictivos (Capítulo 5).

Las técnicas a aplicar en el análisis bivariante difieren según el tipo de variables involucradas:

- Si una variable es **cualitativa** y la otra **cuantitativa**, se pueden comparar distribuciones a través de medidas de tendencia central y dispersión en cada categoría, complementándolo con diagramas de cajas (Sección 3.2).
- Si ambas variables son **cualitativas**, se utilizan tablas de contingencia y gráficos de barras para las frecuencias condicionadas (Sección 3.3).
- Si ambas variables son **cuantitativas**, se analizan medidas de correlación y gráficos de dispersión para identificar relaciones lineales o no lineales (Sección 3.4).

A partir de esta sección, se hará uso del fichero `pulsoProcesado.omv`, correspondiente al fichero de `pulso` original, habiendo eliminado los datos de las alturas inferiores a 120 cm, pesos menores a 40 kg, e individuos con más de 130 ppm en la variable `pulso1` o datos faltantes en esta variable.

3.2. Análisis de una variable cualitativa y otra cuantitativa

En muchas situaciones interesa estudiar el comportamiento de una variable numérica para diferentes subgrupos de población. Por ejemplo, puede ser de interés analizar las posibles diferencias del pulso en reposo entre hombres y mujeres.

El objetivo es resumir numéricamente la variable cuantitativa en función de los grupos que define la variable cualitativa. Para ello, la opción `Exploración` permite realizar este análisis:

1. Ir a `Análisis` → `Exploración` → `Descriptivas`.
2. Arrastrar las variables `Pulso1` a `variables`, y el `Sexo` a la ventana de `Separa por`.

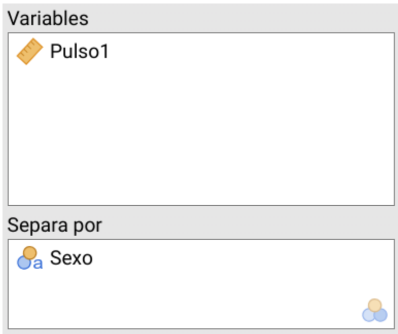


Figura 3.1: Análisis de una variable cuantitativa por grupos.

3. En la sección `Estadísticas`, seleccionar las medidas que se deseen como: N, Perdidos, Media, Mediana, Desviación estándar o Asimetría entre otras. Ver Tabla 3.1.

   Nota 3.1. En este caso también hemos marcado la opción de *Normalidad – Shapiro-Wilk* para hacer un test de normalidad por grupo. Esto será especialmente útil en el Capítulo 4.

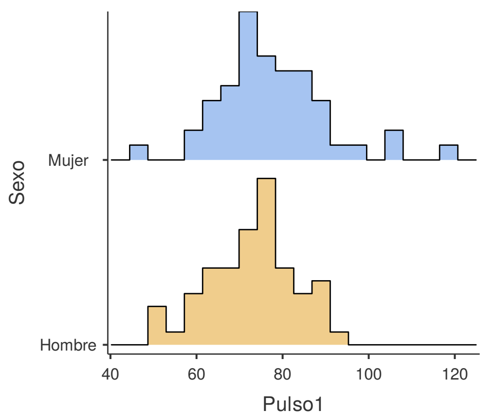


Figura 3.2: Histogramas del pulso en hombres y mujeres.

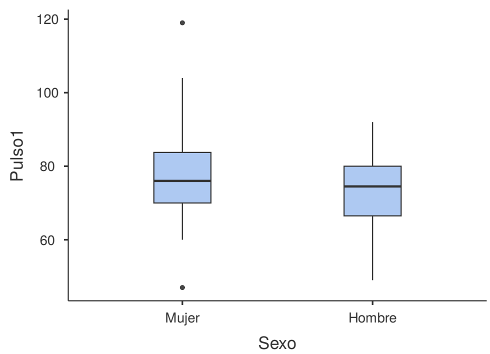


Figura 3.3: Diagrama de cajas del pulso según el sexo.

4. Para mostrar un gráfico, se puede elegir entre un gráfico de cajas o dos histogramas en el mismo menú que en el análisis univariante.

Estadístico	**Sexo**	**Pulso1**
N	Mujer	50
	Hombre	58
Perdidos	Mujer	0
	Hombre	0
Media	Mujer	77.500
	Hombre	72.931
Mediana	Mujer	76.000
	Hombre	74.500
Desviación estándar	Mujer	12.629
	Hombre	10.151
Mínimo	Mujer	47.000
	Hombre	49.000
Máximo	Mujer	119.000
	Hombre	92.000
Asimetría	Mujer	0.755
	Hombre	-0.357
Error est. asimetría	Mujer	0.337
	Hombre	0.314
W de Shapiro-Wilk	Mujer	0.958
	Hombre	0.978
Valor p de Shapiro-Wilk	Mujer	0.073
	Hombre	0.357
Percentil 25	Mujer	70.000
	Hombre	66.500
Percentil 50 (Mediana)	Mujer	76.000
	Hombre	74.500
Percentil 75	Mujer	83.750
	Hombre	80.000

Tabla 3.1: Estadísticos descriptivos del pulso según el sexo

Nota 3.2. El histograma por grupos (Figura 3.2) representa una información más completa que el diagrama de cajas (Figura 3.3). Sin embargo, el diagrama de cajas, por su simplicidad y estructura, transmite de manera más eficaz las posibles diferencias entre grupos.

Comentario sobre los resultados

En primer lugar, la Tabla 3.1 muestra los estadísticos descriptivos para cada grupo, indicando que la media del pulso en mujeres (77.500 ppm) es algo mayor que la de los hombres (72.931 ppm), aunque la diferencia es menor que la variabilidad interna de cada grupo, reflejada en las desviaciones estándar de 12.629 ppm y 10.151 ppm, respectivamente.

Los coeficientes de asimetría sugieren una ligera diferencia en la distribución de los datos: en los hombres, la distribución presenta una cola más larga hacia la izquierda (asimetría negativa), mientras que en las mujeres se observa un sesgo positivo, con valores más elevados en la cola derecha. Sin embargo, ambos valores de asimetría son pequeños en comparación con sus errores estándar, lo que indica que la distribución no se aleja significativamente de la simetría.

El análisis gráfico refuerza estas observaciones. En la Figura 3.2, los histogramas muestran la forma general de las distribuciones para cada grupo, permitiendo evaluar visualmente la dispersión y la simetría. Por otro lado, el diagrama de cajas de la Figura 3.3 proporciona una representación clara de la mediana y los cuartiles, facilitando la comparación de la variabilidad entre los grupos. En este gráfico, los diagramas de caja de ambos grupos se presentan en la misma escala, lo que permite comparaciones directas. Se observa que las distribuciones son muy similares y bastante simétricas, aunque los valores en el grupo de mujeres tienden a ser ligeramente mayores la mediana y los percentiles 25 y 75. Asimismo, se aprecia una dispersión algo mayor en el grupo de mujeres, aunque esta diferencia parece estar influenciada por unos pocos valores extremos.

Además, en el gráfico de cajas se puede notar la ligera asimetría hacia la izquierda en el grupo de hombres y la ligera asimetría hacia la derecha en el grupo de mujeres, en concordancia con los coeficientes de asimetría observados en la Tabla 3.1. Esta coherencia entre los valores numéricos y la representación gráfica sugiere que la distribución del pulso en ambos grupos es relativamente simétrica, sin indicios de asimetrías pronunciadas.

Si los datos presentan valores atípicos o errores de medición, el uso de medidas robustas como la mediana y el rango intercuartílico resulta más adecuado para describir la tendencia central y la dispersión. En este caso, las conclusiones no varían significativamente al comparar la media y la mediana, lo que sugiere que los valores extremos no tienen un impacto considerable en la distribución general del pulso en cada grupo.

Personalizar la anchura de las clases de los histogramas
Con el módulo `vijPlots` se puede hacer un gráfico con varios histogramas muy similar a la Figura 3.2 pero ajustando nosotros las características del histograma como la amplitud de las clases. Para ello, en el módulo mencionado, dentro del menú `Histogram`, se ha de llevar la variable `Pulso1` a la ventana de `variable` y la de `Sexo` a la ventana `Facet`.

3.3. Análisis de dos variables cualitativas.

3.3.1. Cálculo de las frecuencias absolutas conjuntas y representación gráfica

La opción `Explorar` permite realizar este análisis para una muestra concreta. Vamos a ilustrarlo con la variable `Fumador` en función de `Sexo` del fichero `pulsoProcesado.omv`:

1. Ir a `Análisis` → `Recuento` → `Muestras Independientes`.
2. Arrastrar la variable `Sexo` a la casilla `Filas`.

3. Arrastrar la variable `Fumador` a la casilla `Columnas`.

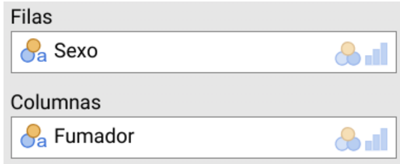


Figura 3.4: Análisis de dos variables cualitativas.

4. En la pestaña de `Celdas`, seleccionar si se quieren representar los porcentajes totales, por fila o por columna, en lugar de las frecuencias absolutas. Ver Figura 3.5.

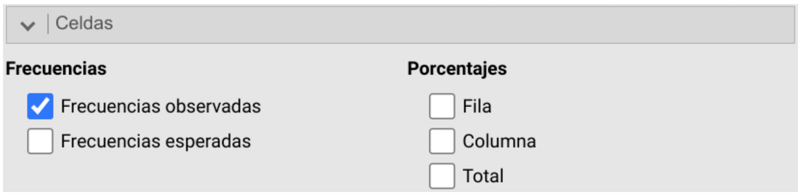


Figura 3.5: Menú de selección del tipo de frecuencias representadas.

5. Para representar los datos gráficamente, ve a `Gráficos` (ver Figura 3.6) y selecciona allí:

   - El tipo de gráfico (*Al lado* o *Alineados*).
   - Si se representan frecuencias absolutas o porcentuales (*Eje Y*), y dentro de la misma, si se representan porcentajes totales o condicionados a filas o columnas. Se recomienda que el gráfico muestre los mismos porcentajes que se ha seleccionado en `Frecuencias`.
   - Variable de las dos que se coloca en el eje X. (*Eje X*).

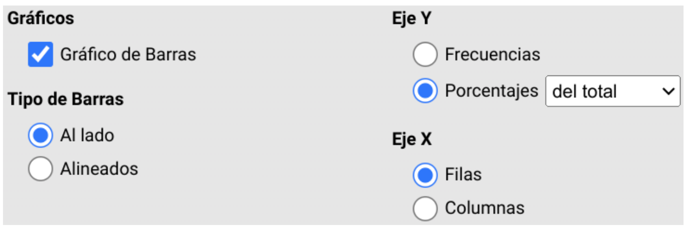


Figura 3.6: Menú de las opciones gráficas.

Nota 3.3. Las filas y las columnas en una tabla de contingencia tienen un papel intercambiable. En este caso, se ha colocado `Sexo` en las filas y `Fumador` en las columnas para facilitar la lectura de las frecuencias dentro de cada grupo de sexo.

Nota 3.4. En la ventana de `Capas`, se puede arrastrar una tercera variable para segmentar el análisis según la misma.

La tabla de contingencia obtenida se presenta en la Tabla 3.2. En ella se observa que en la muestra hay un total de 108 individuos, de los cuales 50 son mujeres y 58 son hombres.

En términos de consumo de tabaco, se registran 11 personas fumadoras y 97 no fumadoras. De las personas fumadoras, 3 son mujeres y 8 hombres.

Sexo	**Sí**	**No**	**Total**
Mujer	3	47	50
Hombre	8	50	58
Total	11	97	108

Tabla 3.2: Tabla de contingencia de frecuencias absolutas entre `Fumador` y `Sexo`.

3.3.2. Interpretación del gráfico de barras

En la Figura 3.7, se observa la distribución del hábito de fumar y sexo en términos de frecuencias absolutas. Este gráfico muestra cuántas personas pertenecen a cada categoría, separando a hombres y mujeres para facilitar la comparación.

El gráfico de la Figura 3.8 muestra la misma información, pero en porcentajes. Nótese que la forma del gráfico es la misma.

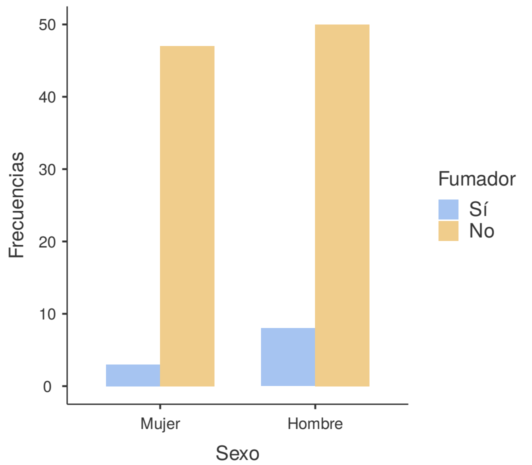


Figura 3.7: Gráfico de barras de frecuencias absolutas del hábito de fumar por sexo.

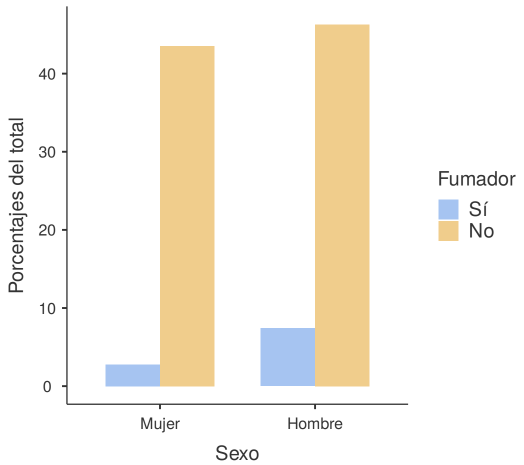


Figura 3.8: Gráfico de barras de frecuencias porcentuales del hábito de fumar por sexo.

3.3.3. Distribuciones condicionadas

Sin embargo, a la vista de la Figura 3.7, surge la pregunta: ¿Podemos afirmar que existe una relación entre el sexo y el hábito tabáquico?

Dado que en la muestra hay más hombres que mujeres, es natural que también haya más fumadores entre los hombres. Para responder adecuadamente, es necesario normalizar los datos utilizando distribuciones condicionadas.

El gráfico de barras de la Figura 3.7 permite describir la muestra mostrando cuántas personas pertenecen a cada categoría.

No obstante, para analizar la relación entre las variables, es más apropiado utilizar distribuciones condicionadas, como se muestra en la Figura 3.9.
La lógica es estudiar a las mujeres por un lado para conocerlas, y por otro, estudiar a los hombres. Sabiendo cómo son ellos y ellas de manera independiente, comparar dicha información.

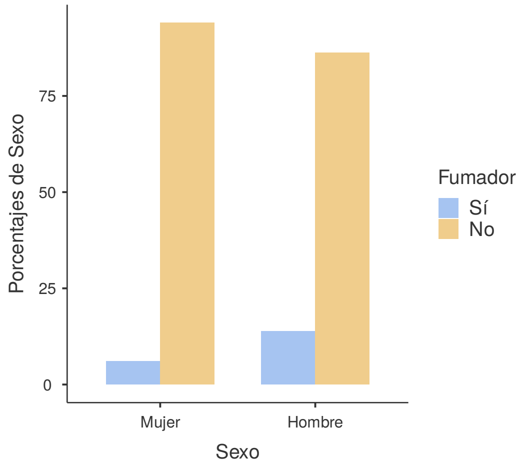


Figura 3.9: Gráfico de barras de frecuencias porcentuales del hábito de fumar por sexo.

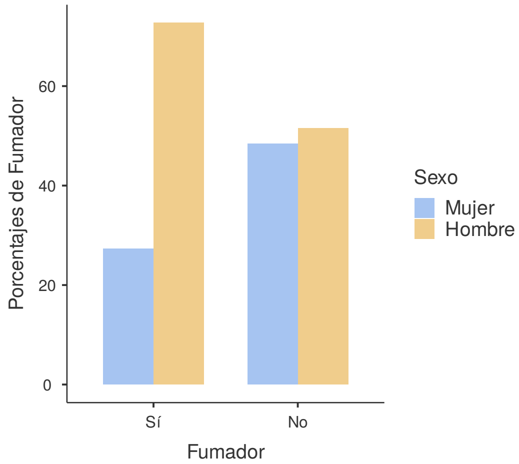


Figura 3.10: Gráfico de barras de frecuencias porcentuales del sexo por hábito de fumar.

Para evitar sesgos debidos a tamaños de muestra desiguales, se utilizan las **distribuciones condicionadas**, que permiten comparar proporciones de manera justa.

La idea de las distribuciones condicionadas es establecer una comparación justa independientemente del número de hombres y mujeres de la muestra. Es, de hecho, la técnica estándar para resumir la relación entre dos variables cualitativas, ya que así se evitan problemas derivados de posibles desbalances en el número de observaciones en cada grupo.

Para calcular estas frecuencias, primero debemos preguntarnos si existe una relación causal o un orden natural en la observación de las variables. En este caso, ¿qué parece más razonable?

- ¿Estudiar si las personas fuman en función de su sexo?
- ¿O estudiar el sexo de las personas en función de si fuman?

Parece razonable pensar que factores psicosociales pueden propiciar un consumo diferente de tabaco en función del sexo.

Dado que es más razonable analizar cómo el sexo influye en el hábito de fumar, se calcularán las distribuciones condicionadas al `Sexo`, es decir, se expresarán los porcentajes dentro de cada categoría de sexo.

3.3.4. Cálculo de las frecuencias condicionadas y su representación

Para calcular frecuencias condicionadas, se han de seguir los mismos pasos que para calcular y representar frecuencias conjuntas mostrados en la Sección 3.3.1, pero seleccionando en el submenú de `Porcentajes` la opción `Filas` o `Columnas`, según interese. El resultado puede verse en la Tabla 3.3 y en la Figura 3.9.

Fumador	**Sí**	**No**	**Total**
Hombre	13.8 %	86.2 %	100 %
Mujer	6.0 %	94.0 %	100 %
Total	10.2 %	89.8 %	100 %

Tabla 3.3: Tabla de contingencia de frecuencias condicionadas entre `Fumador` y `Sexo`.

3.3.5. Interpretación de las frecuencias condicionadas

Ahora se puede observar que las filas (variable a la que se condiciona) suman el 100 %, lo que permite comparar adecuadamente ambos grupos. Dentro del grupo de hombres, el 13.8 % son fumadores, mientras que en las mujeres esta proporción es del 6 %, lo que indica que dentro de la muestra los hombres tienen una mayor tendencia al hábito tabáquico.

Además, en la última fila se encuentran las proporciones generales de la muestra: el 10.2 % de los participantes fuman, mientras que el 89.8 % no fuman.

Interpretación
El uso de distribuciones condicionadas permite comparar adecuadamente los grupos, evitando sesgos por diferencias en el tamaño muestral. En este caso, se observa que la proporción de fumadores es mayor entre los hombres que entre las mujeres.

Frecuencias condicionadas a la otra variable

En este caso, se ha condicionado a las filas, ya que se pretende estudiar el hábito de fumar por sexo. Sin embargo, también es posible invertir el análisis y estudiar la distribución del sexo dentro de cada categoría de fumador.

En la Figura 3.10 se presenta el gráfico de barras correspondiente a este análisis, donde se observa la proporción de hombres y mujeres dentro de los grupos de fumadores y no fumadores. La tabla correspondiente a este gráfico se muestra en la Tabla 3.4.

En este caso, los porcentajes dentro de cada columna suman 100 %, lo que permite responder a la pregunta: "¿Cómo se distribuye el sexo dentro de los fumadores y no fumadores?". Se observa que el 72.7 % de las personas fumadoras en la muestra son hombres, mientras que las mujeres representan solo el 27.3 %. Por otro lado, entre las personas no

Fumador	**Sí**	**No**	**Total**
Mujer	27.3 %	48.5 %	46.3 %
Hombre	72.7 %	51.5 %	53.7 %
Total	100.0 %	100.0 %	100.0 %

Tabla 3.4: Distribución del sexo dentro de cada grupo de fumador.

fumadoras, la proporción de hombres y mujeres es más equilibrada, con un 51.5 % y 48.5 %, respectivamente.

3.4. Análisis de dos variables cuantitativas

3.4.1. Diagrama de dispersión

La representación gráfica mediante diagramas de dispersión supone el primer paso a la hora de explorar la relación entre dos variables cuantitativas. Estos gráficos permiten identificar patrones, tendencias y posibles relaciones funcionales entre las variables, facilitando la interpretación y el análisis de los datos.

Importancia del diagrama de dispersión

Los diagramas de dispersión permiten visualizar la relación entre variables cuantitativas, detectar patrones o tendencias, validar supuestos y generar hipótesis para futuras investigaciones. Además, la inclusión de una recta de mínimos cuadrados o la adición de una tercera variable a través del color permite obtener una comprensión más profunda de los datos e hipotetizas con posibles relaciones entre ambas variables.

Ejemplo: Relación entre altura y peso

En este ejemplo, se analiza la relación entre la `Altura` y el `Peso` de los individuos del fichero `pulsoProcesado.omv`.

Antes de realizar el gráfico, es importante definir qué variable debe situarse en el eje X (independiente) y cuál en el eje Y (dependiente).

Escoger los ejes de las variables.

Si se tiene conocimiento previo sobre una posible relación causal entre las variables, o si una de ellas es más fácil de medir o controlar, esta información debe considerarse al elegir los ejes, ya que, en esencia, se trata de distinguir entre la variable explicativa y la explicada.

Siguiendo este criterio, se representa el `Peso` en función de la `Altura`, situando la

`Altura` como variable independiente (X) y el `Peso` como variable dependiente (Y).

3.4.2. Generación del diagrama de dispersión en *jamovi*

Para generar un diagrama de dispersión en *jamovi*, se deben seguir estos pasos:

1. Ir a `Exploración` → `Gráfico de dispersión`. Ver Figura 3.11.

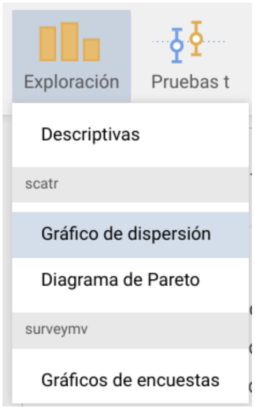


Figura 3.11: Menú para representar un diagrama de dispersión.

2. Arrastrar la variable `Altura` a la casilla `Eje X`.
3. Arrastrar la variable `Peso` a la casilla `Eje Y`. Ver Figura 3.12.

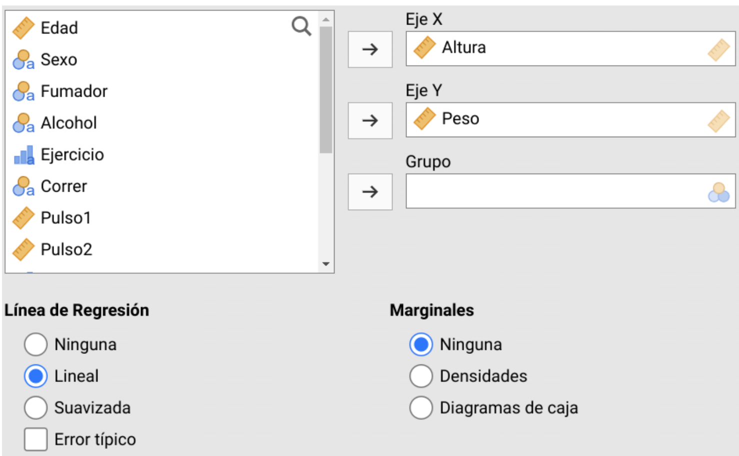


Figura 3.12: Elección de las variables del diagrama de dispersión.

4. Opcionalmente, se puede añadir una tercera variable en la casilla `Grupo` para diferenciar categorías por color.
5. En la sección `Línea de Regresión`, seleccionar la opción `Lineal` para visualizar la recta de regresión.

Ejemplo 1: Diagrama de dispersión con recta de mínimos cuadrados

En la Figura 3.13, se presenta el diagrama de dispersión resultante al representar el `Peso` en función de la `Altura`, incluyendo una recta de mínimos cuadrados.

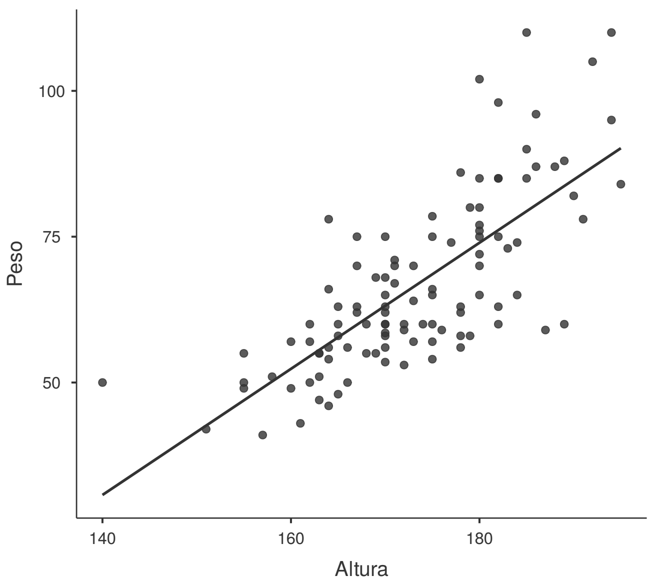


Figura 3.13: Diagrama de dispersión entre altura y peso con recta de mínimos cuadrados.

Como se observa en la Figura 3.13, los puntos tienden a agruparse alrededor de la recta de mínimos cuadrados, lo que sugiere una relación lineal positiva entre las variables: a mayor altura, mayor peso. Esta relación es coherente con lo esperado en un contexto antropométrico, ya que la altura y el peso suelen estar correlacionados positivamente.

Ejemplo 2: Diagrama de dispersión con diferenciación por sexo

Para profundizar en el análisis, se puede incluir una tercera variable en el gráfico utilizando el color para distinguir grupos. En este caso, se ha añadido la variable `Sexo` en la casilla `Grupo`, obteniendo el gráfico mostrado en la Figura 3.14.

En la Figura 3.14, los puntos están coloreados según el sexo, lo que permite visualizar posibles diferencias en la relación entre altura y peso para hombres y mujeres. Se observa que, en general, los hombres tienden a presentar valores de altura y peso mayores en comparación con las mujeres. Además, la recta de regresión sigue mostrando una relación lineal positiva en ambos grupos.

Nota 3.5. El uso del color en los diagramas de dispersión permite diferenciar grupos dentro de los datos y detectar posibles diferencias entre categorías. En este caso, la diferenciación por sexo facilita la identificación de tendencias particulares en cada grupo.

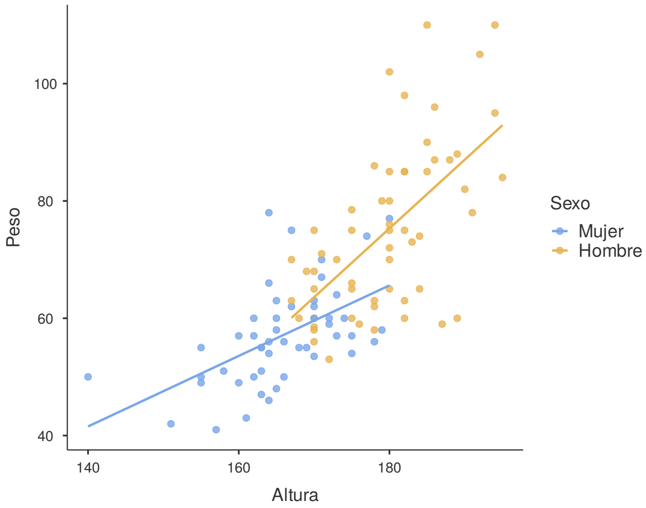


Figura 3.14: Diagrama de dispersión entre altura y peso diferenciando por sexo.

3.4.3. Coeficiente de correlación

El coeficiente de correlación es una medida estadística que indica la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. Su cálculo es fundamental para comprender el grado en que dos variables están relacionadas y evaluar si existe una posible relación entre ellas.

Para calcular la matriz de correlaciones en *jamovi*, se deben seguir estos pasos:

1. Ir a `Análisis` → `Regresión` → `Matriz de correlaciones.` (Figura 3.15).

Figura 3.15: Menú para el cálculo de la matriz de correlaciones.

2. Arrastrar las variables de interés a la casilla `Variables`. Pueden ser más de dos variables cuantitativas. Ver Figura 3.16.

3. Seleccionar el coeficiente de correlación a calcular, por ejemplo *Pearson* (Véanse las consideraciones adicionales a continuación).

4. Si se desean, se pueden habilitar las opciones `Mostrar significación`, `Marcar correlaciones significativas`, `N` e `Intervalos de confianza`.

5. Opcionalmente, también se puede habilitar la opción `Matriz de correlaciones` para obtener una representación gráfica.

Consideraciones adicionales.
Además del coeficiente de correlación de Pearson, que mide la relación lineal entre dos variables cuantitativas,es posible calcular otros coeficientes de correlación:

- **Coeficiente de correlación de Spearman**: Se utiliza cuando la relación entre las variables no es estrictamente lineal, sino que sigue una tendencia monótona (es decir, a medida que una variable aumenta, la otra también lo hace o disminuye de manera consistente). Es especialmente útil cuando los datos presentan asimetría o valores atípicos que podrían afectar a Pearson. También es utilizado cuando al menos una variable es ordinal.

- **Coeficiente Tau-b de Kendall**: Es otra alternativa no paramétrica que mide la asociación entre dos variables ordinales o cuantitativas. Su ventaja sobre Spearman es que penaliza menos los empates en los datos, por lo que es preferible en conjuntos de datos con muchas observaciones repetidas.

- **Intervalos de confianza**: Se pueden calcular al 95 % o a otro nivel especificado, proporcionando un rango en el que es probable que se encuentre el verdadero valor del coeficiente en la población.

- **Significación de la correlación**: Marca con asteriscos (*, **, ***) en función de la significación para facilitar la identificación de relaciones estadísticamente relevantes.

- **Matriz de correlaciones gráficas**: Permite visualizar la relación entre múltiples variables mediante diagramas de dispersión y densidad.

La elección del coeficiente adecuado depende del tipo de datos y del patrón de relación esperado entre las variables. En este caso, se ha utilizado el coeficiente de Pearson debido a que las variables analizadas son continuas y se asume una relación lineal entre ellas.

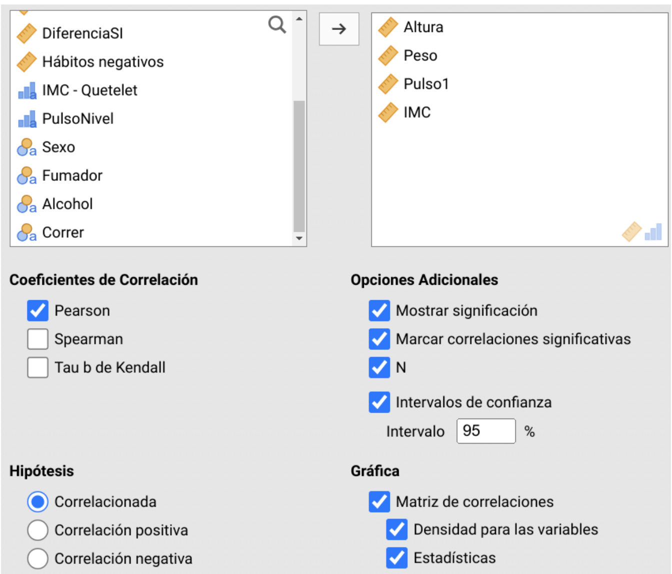


Figura 3.16: Opciones del menú de cálculo de la matriz de correlaciones.

Resultados

En la Tabla 3.5, se presenta la matriz de correlaciones obtenida, que incluye los coeficientes de correlación de Pearson, los valores de significación y los intervalos de confianza al 95 %.

Matriz de Correlaciones

		Altura	Peso	Pulso1	IMC
Altura	R de Pearson	—			
	gl	—			
	valor p	—			
	IC 95% Superior	—			
	IC 95% Inferior	—			
	N	—			
Peso	R de Pearson	0.741***	—		
	gl	104	—		
	valor p	<.001	—		
	IC 95% Superior	0.816	—		
	IC 95% Inferior	0.640	—		
	N	106	—		
Pulso1	R de Pearson	-0.181	-0.205*	—	
	gl	104	105	—	
	valor p	0.063	0.034	—	
	IC 95% Superior	0.010	-0.016	—	
	IC 95% Inferior	-0.360	-0.380	—	
	N	106	107	—	
IMC	R de Pearson	0.304**	0.860***	-0.178	—
	gl	104	104	104	—
	valor p	0.002	<.001	0.068	—
	IC 95% Superior	0.468	0.903	0.013	—
	IC 95% Inferior	0.120	0.801	-0.356	—
	N	106	106	106	—

Nota. * p < .05, ** p < .01, *** p < .001

Tabla 3.5: Tabla de la matriz de correlaciones entre `Altura`, `Peso`, `Pulso1` e `IMC`.

La Figura 3.17 muestra la representación gráfica de la matriz de correlaciones. Debajo de la diagonal podemos ver los diagramas de dispersión dos a dos junto con las rectas de mínimos cuadrados. En la diagonal un diagrama de densidad, que es el equivalente a un histograma *suavizado* o *idealizado*. Por encima de la diagonal ver los coeficientes de correlación calculados junto con sus marcas de significación.

A partir de los resultados obtenidos en la Tabla 3.5 y la Figura 3.17, se pueden extraer las siguientes conclusiones:

- **Altura y Peso** ($r = 0{,}741, p < 0{,}001$): Existe una correlación positiva fuerte entre la altura y el peso, lo que sugiere que, en general, las personas más altas tienden a pesar más.
- **Altura e IMC** ($r = 0{,}304, p = 0{,}002$): La correlación positiva débil sugiere que

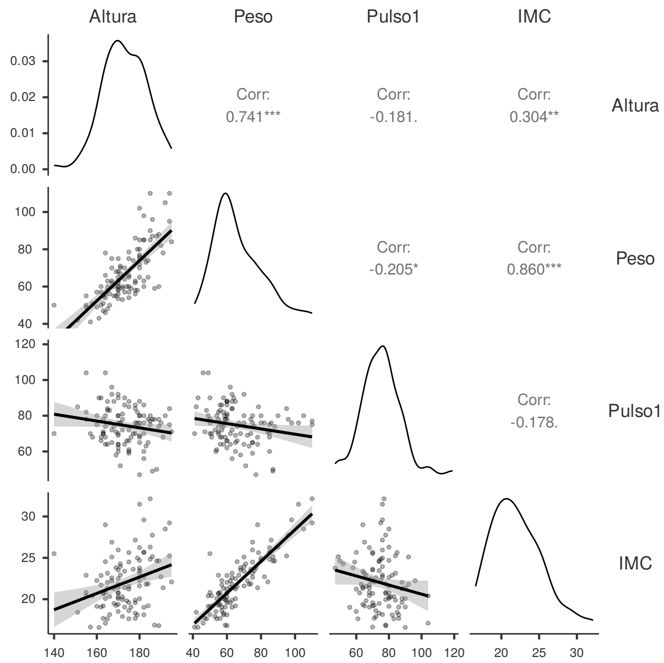


Figura 3.17: Representación gráfica de la matriz de correlaciones entre `Altura`, `Peso`, `Pulso1` e `IMC`.

existe una tendencia a que las personas más altas tengan un índice de masa corporal (IMC) ligeramente mayor.

- **Altura y Pulso1** ($r = -0{,}181, p = 0{,}063$): No se encuentra una relación estadísticamente significativa entre la altura y el pulso en reposo, lo que indica que la altura no parece influir de manera relevante en esta variable.
- **Peso e IMC** ($r = 0{,}860, p < 0{,}001$): Se observa una correlación positiva muy fuerte entre el peso y el IMC, lo que es esperable, dado que el IMC se calcula utilizando el peso.
- **Peso y Pulso1** ($r = -0{,}205, p = 0{,}034$): Existe una correlación negativa débil entre el peso y el pulso en reposo, lo que podría indicar que las personas con mayor peso tienden a tener un pulso ligeramente más bajo.
- **IMC y Pulso1** ($r = -0{,}178, p = 0{,}068$): No se encuentra una relación estadísticamente significativa entre el IMC y el pulso en reposo, lo que sugiere que no hay una asociación clara entre estas variables.

Resumen de resultados
Los resultados muestran relaciones significativas entre altura y peso, peso e IMC, y peso y pulso en reposo. La fuerte correlación entre peso e IMC es esperada, dado el método de cálculo del IMC. En contraste, la relación entre altura y pulso, así como la de IMC y pulso, no es significativa, lo que indica que no hay evidencia suficiente para afirmar una asociación entre estas variables.

3.5. Ejercicios

1. Compara el comportamiento de la variable `Pulso después` (`Pulso1`) del fichero `pulsoProcesado.omv` en el grupo de los que han corrido y en el grupo de los que no han corrido (variable `Correr`).

2. Muestra si hay diferencias de peso o de altura entre los hombres y las mujeres en los datos de la muestra del fichero `pulsoProcesado.omv`.

3. Se ha realizado un estudio para comprender el efecto del hábito de fumar en los patrones de sueño. La variable considerada es el tiempo que tarda una persona en quedarse dormida (en minutos). El fichero `sueno.sav` contiene los datos de los tiempos de personas fumadoras y no fumadoras.

   a) Realiza un estudio descriptivo de la variable `Tiempo`.
   b) Repite el estudio diferenciando entre personas fumadoras y no fumadoras.
   c) Utilizando los resultados obtenidos, analiza si la distribución del tiempo hasta quedarse dormido es la misma en ambos grupos. Justifica tu respuesta.

4. Se están estudiando dos medicamentos, A y B, para combatir el virus de la gripe. Se han administrado por vía oral dosis únicas de 100 mg a adultos sanos. La variable estudiada es el tiempo requerido (en minutos) para alcanzar la concentración máxima en plasma.

   a) A partir de los datos contenidos en el archivo `gripe.sav`, ¿encuentras algún dato atípico en esta variable?
   b) En caso afirmativo, justifica cuál o cuáles de ellos podrían eliminarse y/o corregirse.
   c) Realiza el estudio descriptivo de esta variable para cada uno de los medicamentos, una vez corregidos y/o eliminados los datos atípicos correspondientes.

5. Realiza un análisis de la relación entre el consumo de alcohol (`Hábito enólico`) y el `Sexo` de las personas encuestadas.

   a) Crea una tabla cruzada donde la variable `Hábito enólico` se sitúe en las filas y `Sexo` en las columnas.
   b) Calcula las distribuciones condicionadas al sexo, es decir, los porcentajes por columna.
   c) Interpreta los resultados. ¿Parece haber diferencias en el consumo de alcohol entre hombres y mujeres?
   d) Representa los resultados mediante un gráfico de barras y analiza si la visualización apoya la interpretación numérica.

   Nota 3.6. Si existieran diferencias en los porcentajes de consumo de alcohol entre hombres y mujeres, podríamos sospechar de una relación entre estas variables. En caso contrario, si los porcentajes fueran similares en ambos sexos, podríamos pensar que no hay una asociación clara.

6. Relación entre tabaco y alcohol.

   En este ejercicio, vamos a analizar la relación entre el `Consumo de tabaco` y el `Consumo de alcohol`.

   a) Construye dos tablas cruzadas:
      - Una con `Tabaco` en las filas y `Alcohol` en las columnas.
      - Otra con `Alcohol` en las filas y `Tabaco` en las columnas.

   b) Calcula los porcentajes por columna en ambas tablas.

   c) Compara los resultados de ambas tablas. ¿El análisis cambia dependiendo de qué variable se coloque en las filas y cuál en las columnas?

   d) Representa los datos mediante un gráfico de barras y analiza visualmente la relación entre ambas variables.

   Punto clave: En este caso, podemos estudiar tanto el consumo de alcohol en función del tabaco como el consumo de tabaco en función del alcohol. Dependiendo de la perspectiva elegida, las conclusiones pueden ser diferentes.

7. Relación entre IMC y consumo de tabaco.

   En este ejercicio, analizaremos si el `Índice de Masa Corporal (IMC)` está relacionado con el `Consumo de tabaco`. Para ello, se categorizará el `IMC` en tres niveles:

   - Bajo peso.
   - Peso normal.
   - Sobrepeso/obesidad.

   a) Construye una tabla cruzada donde la variable `IMC (en tres niveles)` esté en las filas y la variable `Tabaco` en las columnas.

   b) Calcula los porcentajes por columna para analizar la distribución del IMC en cada grupo de fumadores y no fumadores.

   c) Interpreta los resultados. ¿Se observan diferencias notables entre los grupos?

   d) Representa los datos gráficamente y analiza si el consumo de tabaco parece estar asociado con el peso corporal.

   Nota 3.7. Dado que el IMC puede estar influenciado por múltiples factores, este análisis debe interpretarse con cautela. Sin embargo, si se encuentran diferencias significativas en la distribución del IMC según el hábito tabáquico, podríamos estar ante una relación interesante para explorar.

8. Relación entre IMC y regularidad en el ejercicio.

   En este último ejercicio, analizaremos si la `Regularidad en el ejercicio` está relacionada con el `IMC (en tres niveles)`.

   a) Construye una tabla cruzada en la que la variable `IMC (en tres niveles)` esté en las filas y la variable `Regularidad en el ejercicio` en las columnas.

   b) Calcula los porcentajes por columna para identificar posibles patrones en la distribución del IMC según la frecuencia con la que las personas realizan ejercicio.

c) Interpreta los resultados. ¿Parece haber una relación entre estos factores? ¿Las personas que hacen ejercicio con mayor regularidad tienen una distribución de IMC diferente a quienes no lo hacen?

d) Genera un gráfico de barras para representar la información y apóyate en él para reforzar tu interpretación.

Idea clave: Si la regularidad en el ejercicio y el IMC están relacionados, podríamos encontrar diferencias en la proporción de personas con bajo peso, peso normal y sobrepeso/obesidad en función de su nivel de actividad física.

"La Estadística es la gramática de la ciencia."

Karl Pearson

4

Inferencia Estadística

4.1. Introducción a la Inferencia Estadística

La **inferencia estadística** es un conjunto de métodos utilizados para extraer conclusiones sobre una población utilizando la información contenida en una muestra de la misma. En el contexto de la investigación científica, el principal objetivo es tomar decisiones basadas en la evidencia obtenida de los datos observados, con la finalidad de generalizar estas conclusiones al conjunto de la población. Sin embargo, esto implica enfrentarse al problema de saber si los resultados obtenidos en una muestra son válidos para la población en su totalidad.

Existen dos herramientas principales en la inferencia estadística: la **estimación de parámetros** y los **contrastes de hipótesis**. La estimación de parámetros permite determinar, mediante intervalos de confianza, un rango de valores plausible para un parámetro poblacional, como la media o la proporción. Por otro lado, los contrastes de hipótesis se utilizan para evaluar si una afirmación sobre la población es consistente con los datos muestrales.

El problema central en la inferencia estadística es la **toma de decisiones** en condiciones de incertidumbre. Por ejemplo, si observamos que, en una muestra, las mujeres tienden a acumular más grasa que los hombres en los muslos, ¿podemos concluir que esta diferencia es significativa en toda la población? Este tipo de preguntas nos lleva a utilizar contrastes de hipótesis para decidir si los resultados son atribuibles al azar o reflejan una diferencia real.

4.2. ¿Cómo escoger el test adecuado?

El proceso de selección del test estadístico adecuado sigue un esquema lógico basado en el tipo de la, o las, variables de interés y la naturaleza de la comparación que se desea realizar. A continuación se propone un esquema de decisión para saber qué test es adecuado.

En primer lugar, es fundamental identificar la **naturaleza de la variable respuesta**. Si la variable es **cuantitativa**, el enfoque se centrará en analizar medias o medianas, mientras que si es **cualitativa**, el análisis se basará en proporciones o frecuencias. A partir de aquí, se abre un primer conjunto de decisiones.

Cuando la **variable de interés es cualitativa**, la pregunta central es si se desea comparar proporciones dentro de una única muestra o evaluar si existe relación entre dos variables categóricas. Si se está probando la igualdad de proporciones dentro de una sola variable (Sección 4.3), se usa un test de bondad de ajuste de proporciones, mientras que si se busca evaluar la independencia entre dos variables cualitativas, se recurre a un test de independencia de Chi-cuadrado (Sección 4.4).

Si se trata de una **variable cuantitativa**, el primer aspecto a evaluar es si la hipótesis implica la comparación de una única muestra con un valor teórico o si se están comparando dos o más muestras. En el caso de trabajar con una sola muestra y un valor de referencia, se debe comprobar si el tamaño de la muestra es suficientemente grande, o en su defecto, si los datos siguen una distribución normal (Sección 4.5). Si el tamaño es grande o la normalidad se cumple, se procede con un test paramétrico como el *t de una muestra*, mientras que si no se cumple, se opta por una prueba no paramétrica como el test de Wilcoxon para una muestra (Sección 4.6).

Cuando se pretende comparar **una variable cuantitativa en dos muestras independientes**, el primer paso es evaluar la normalidad en cada grupo, si los tamaños muestrales son pequeños, y la homogeneidad de varianzas. Si los datos son normales o se tiene un número grande de observaciones en cada grupo, se emplea un test t de muestras independientes. Además, si las varianzas no se pueden considerar iguales en el test de homogeneidad, entonces se ha de emplear la *corrección de Welch*. Por otro lado, si no se cumple la normalidad, se opta por el test de Mann-Whitney (Sección 4.7).

Si en lugar de dos grupos, para una **variable cuantitativa**, se comparan **tres o más muestras independientes**, el análisis se basa en el test ANOVA, de nuevo si se tiene un número mínimo de datos o si se pueden considerar normales en cada grupo. Además, si los grupos tienen varianzas homogéneas, al igual que para el test t se debería realizar la corrección de Welch. En caso contrario, se usa la prueba de Kruskal-Wallis. Estas pruebas son el análogo de las pruebas de dos muestras. En caso de encontrar diferencias significativas en estos tests, es necesario realizar análisis post-hoc para identificar qué grupos son diferentes (Sección 4.8).

En caso de querer evaluar la diferencia en **variables cuantitativas para muestras pareadas** (por ejemplo, mediciones antes y después en los mismos individuos), la decisión sigue un camino similar: si la diferencia entre pares de observaciones sigue una distribución normal o se tiene un tamaño de muestra adecuado, se aplica un test t para muestras dependientes; en caso contrario, se usa el test de rangos con signo de Wilcoxon, o sus alternativas para más grupos (Sección 4.9).

Finalmente, si el objetivo es analizar la **asociación entre dos variables cuantitativas u ordinales**, la decisión depende de la distribución de los datos. Si ambas variables son cuantitativas y relativamente simétricas, se emplea la correlación de Pearson; si alguna de ellas no lo es, la correlación de Spearman o la τ de Kendall son las alternativas habituales. Calcular el test de correlación, el intervalo de confianza del mismo, y su interpretación, ya se ha mostrado en la Sección 3.4.3.

4.3. Una Variable Cualitativa

4.3.1. Prueba binomial o de χ^2

Para comparar proporciones entre grupos en *jamovi*, se puede realizar una prueba de bondad de ajuste χ^2 (leído como chi-cuadrado) o prueba binomial. A continuación, se describe el procedimiento paso a paso.

Supongamos que queremos comprobar si el número de mujeres en la población se puede suponer del 50 %. Este mismo problema se resolverá con una opción más versátil en la Sección 4.3.2. No obstante, la ventaja de este método es que *jamovi* proporciona directamente el *Intervalo de Confianza* (IC) de la proporción.

En particular, queremos realizar el siguiente test de hipótesis para la proporción de mujeres que llamamos p:

- $H_0 : p = 0{,}5$.
- $H_1 : p \neq 0{,}5$.

1. Ir a `Análisis` → `Recuento` → `2 Resultados - Prueba binomial` (Figura 4.1).

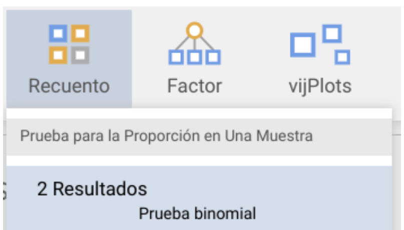


Figura 4.1: Interfaz de *jamovi* para realizar una prueba de proporciones.

2. Seleccionar la variable categórica de interés (en este caso `Sexo`).

3. Definir el **valor de prueba** de la proporción esperada, generalmente 0.5 si se quiere contrastar igualdad entre dos grupos (Figura 4.2).

4. Activar la casilla `Intervalos de confianza` para calcular un intervalo al 95 %.

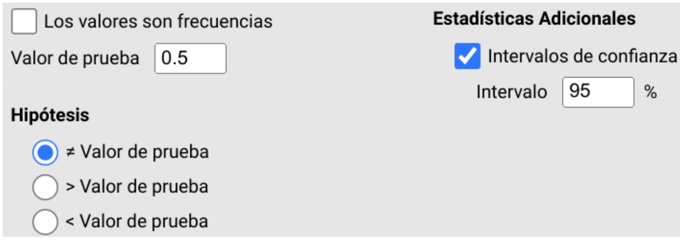


Figura 4.2: Opciones de la prueba de proporciones.

En la Tabla 4.1 se presentan los resultados obtenidos:

Nivel	**Frecuencia**	**Total**	**Proporción**	***p*-valor**	**Inferior**	**Superior**
Mujer	50	108	0.463	0.501	0.367	0.562
Hombre	58	108	0.537	0.501	0.438	0.633

Tabla 4.1: Resultados de la prueba binomial para la variable `Sexo`.

- Se presentan las frecuencias y proporciones observadas para cada categoría.

- El intervalo de confianza al 95 % para la proporción de mujeres en la muestra es $[0{,}367, 0{,}562]$. Esto significa que, con un 95 % de confianza, la proporción real de mujeres en la población se sitúa en este rango. Dado que el valor 0.5 está dentro del intervalo, no se puede concluir que la proporción de mujeres en la población sea significativamente distinta del 50 %.

- La prueba de bondad de ajuste χ^2 evalúa si la distribución observada difiere de la esperada. En este caso, el valor $p = 0{,}501$ indica que no hay evidencia suficiente para rechazar la hipótesis nula de igualdad de proporciones. Este resultado ha de estar en consonancia con la conclusión anterior razonando a partir del intervalo de confianza.

4.3.2. Test de bondad de ajuste

El **test de bondad de ajuste**, comúnmente basado en el estadístico χ^2, permite evaluar también si la distribución observada de una variable cualitativa se ajusta a una distribución teórica esperada. En términos generales, este test nos ayuda a determinar si las proporciones de las distintas categorías de una variable son iguales o siguen una distribución específica.

Algunas preguntas que podemos abordar con este test son:

- ¿Podemos asumir que hay igualdad entre el número de hombres y mujeres en la población?
- ¿Es uniforme la distribución de la variable `actividad física`?
- ¿Hay el doble de individuos con actividad moderada que en las categorías de actividad baja y alta?

Ejemplo: paridad entre hombres y mujeres

Como primer ejemplo, queremos comprobar si la distribución de la variable `sexo` es equitativa en la población. Para ello, formulamos las siguientes hipótesis:

- H_0: $p_{\text{mujer}} = p_{\text{hombre}} = 0{,}5$ (las proporciones son iguales).
- H_1: $p_{\text{mujer}} \neq p_{\text{hombre}}$ (las proporciones son diferentes)

Para realizar este test en *jamovi*, se sigue el siguiente procedimiento:

1. En la pestaña `Recuento`, seleccionar la opción `N resultados -` χ^2 `de bondad de ajuste` (Figura 4.3).

Figura 4.3: Selección del test de bondad de ajuste en *jamovi*.

2. Arrastrar la variable de interés, `Sexo`, al cuadro `Variable`.
3. Establecer las proporciones esperadas:
   - Si queremos contrastar igualdad de proporciones, asignamos la misma razón a todas las categorías (Figura 4.4).
   - Si queremos evaluar una distribución específica (ver el siguiente ejemplo, que la categoría `moderada` tenga el doble de frecuencia que las otras dos), ajustamos las razones proporcionalmente (Figura 4.5).

Los resultados obtenidos en la muestra de `pulsoProcesado.omv` se pueden ver en las Tablas 4.2 y 4.3

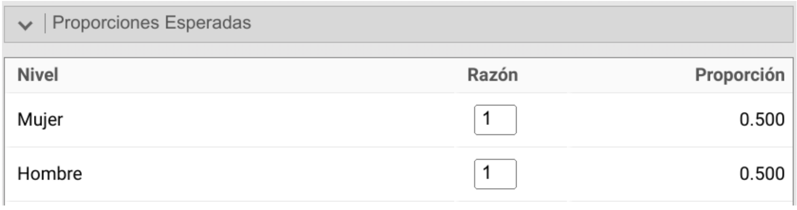


Figura 4.4: Configuración del test para proporciones iguales en *jamovi*.

Nivel	Frecuencia	Proporción
Mujer	50	0.463
Hombre	58	0.537

Tabla 4.2: Proporciones de la variable `sexo`.

χ^2	gl	p
0.593	1	0.441

Tabla 4.3: Resultados del test χ^2 de bondad de ajuste para la variable `sexo`.

El test χ^2 arroja un p-valor mayor que 0,05, indicando que no hay diferencias significativas, por lo que no se rechaza la hipótesis nula. Esto implica que no hay evidencia suficiente para afirmar que la distribución de sexo en la población no es distinta de 50 %-50 %.

Ejemplo: uniformidad en una distribución de frecuencias.

El test χ^2 de bondad de ajuste también se puede aplicar para evaluar si la distribución de la variable `actividad física` sigue una distribución uniforme en la población; o escritas formalmente las hipótesis:

- H_0: $p_{\text{moderada}} = p_{\text{baja}} = p_{\text{alta}}$ (las personas se identifican por igual en cada categoría).
- H_1: las proporciones no son iguales.

En la Tabla 4.4 se presentan las frecuencias y proporciones observadas en la muestra para cada nivel de actividad física, mientras que en la la Tabla 4.5 se puede observar el resultado del test.

Nivel	Frecuencia	Proporción
Baja	37	0.343
Moderada	57	0.528
Alta	14	0.130

Tabla 4.4: Proporciones de la variable `actividad física`.

χ^2	**gl**	**p**
25.722	2	$< 0{,}001$

Tabla 4.5: Resultados del test χ^2 de bondad de ajuste para la variable `actividad física`.

En este caso, como el p-valor es menor que el umbral de significación habitual de 0.05, se rechaza la hipótesis nula de igualdad de proporciones entre las tres categorías de actividad física. Esto sugiere que la distribución de la actividad física en la muestra no es uniforme, con una mayor proporción de individuos reportando una actividad moderada en comparación con los otros dos niveles.

Ejemplo: comparación de una distribución específica

Dado que la distribución de la variable `actividad física` no es uniforme, ahora evaluamos si la proporción de individuos con actividad `moderada` es el doble que la de las categorías `baja` y `alta`. Las hipótesis del test se pueden escribir de la siguiente manera:

- H_0: $p_{\text{moderada}} = 2p_{\text{baja}} = 2p_{\text{alta}}$ (la categoría `moderada` tiene el doble de individuos que las otras dos).
- H_1: la distribución de proporciones no sigue este esquema.

Para realizar este test, se han de seguir las instrucciones anteriores, pero indicando la proporción de interés, $1:2:1$, tal y como se muestra en la Figura 4.5. De esta manera, la hipótesis nula sostiene que la categoría `moderada` debería contar con el doble de individuos que las categorías `baja` y `alta`.

Figura 4.5: Configuración del test para proporciones específicas en *jamovi*.

El resultado del test de bondad de ajuste muestra un p-valor de $p = 0{,}006$ (Tabla 4.6).

χ^2	**gl**	**p**
10.130	2	0.006

Tabla 4.6: Resultados del test χ^2 de bondad de ajuste con distribución esperada $1:2:1$.

Puesto que el p-valor es menor que el umbral de significación de 0.05, se rechaza la hipótesis nula, lo que indica que la proporción observada en la muestra no sigue exactamente la razón 1 : 2 : 1. En otras palabras, aunque la actividad `moderada` sigue siendo la categoría más frecuente, su proporción relativa, rechazamos que sea exactamente el doble de las categorías `baja` y `alta`.

4.4. Independencia entre dos variables cualitativas: Test Exacto de Fisher

En la Sección 3.3 se presentó el análisis de tablas de contingencia en *jamovi*, mostrando cómo obtener frecuencias absolutas, relativas y condicionadas. En este apartado, además de revisar estas herramientas, exploraremos su uso para realizar inferencia estadística sobre la relación entre dos variables cualitativas.

Ejemplo: Asociación entre el sexo y el hábito de fumar

Utilizando la base de datos `pulsoProcesado.omv`, se desea comprobar si existe asociación entre la variable `sexo` y el hábito de `fumador`.

En este contexto, consideramos `sexo` como la **variable explicativa** y `fumador` como la **variable respuesta**, dado que queremos examinar si el hecho de ser hombre o mujer influye en la probabilidad de fumar.

Para evaluar si estas diferencias observadas en las proporciones son estadísticamente significativas, se establece el siguiente test de hipótesis:

- H_0: No hay relación entre `sexo` y `fumador`, es decir, la proporción de fumadores es la misma en hombres y mujeres.
- H_1: Existe asociación entre `sexo` y `fumador`, es decir, la proporción de fumadores es diferente entre hombres y mujeres.

Interpretación: RechazarH_0 indicaría que la proporción de fumadores varía según el sexo. En caso contrario, no se puede concluir que exista una relación significativa entre ambas variables.

Para realizar el análisis en *jamovi*, se comienza calculando las frecuencias condicionadas tal y como se realizó en la Sección 3.3 para así conocer cómo son los datos:

1. Ir a `Análisis` → `Recuento` → `Muestras Independientes - Prueba de independencia de` χ^2 (Figura 4.6).

Figura 4.6: Menú de selección de la prueba de independencia en *jamovi*.

2. Arrastrar `sexo` al cuadro de `Filas` y `fumador` al de `Columnas` (Figura 4.7).

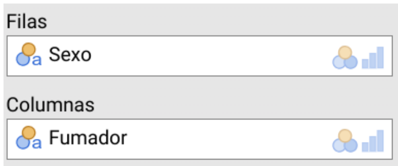


Figura 4.7: Selección de las variables para la tabla de contingencia.

Nota 4.1. Se pueden intercambiar las variables entre filas y columnas, pero en el paso posterior se deberán calcular las frecuencias condicionadas correspondientes.

3. En la pestaña `Celdas`, seleccionar la opción `Porcentajes` → `Fila`, para comparar la proporción de fumadores dentro de cada categoría de `sexo` (Figura 4.8).

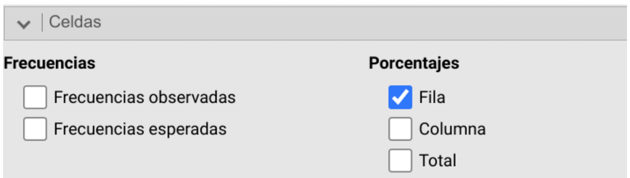


Figura 4.8: Configuración de la pestaña `Celdas` para visualizar los porcentajes por fila.

La tabla de contingencia resultante se presenta en la Tabla 4.7. En este caso, solo con fines ilustrativos, se han incluido tanto las frecuencias absolutas como los porcentajes por fila. De esta manera, visualizamos la distribución de `fumador` en cada categoría de `sexo`.

Sexo	**Sí**	**No**	**Total**
Mujer	3 (6.0 %)	47 (94.0 %)	50
Hombre	8 (13.8 %)	50 (86.2 %)	58
Total	11 (10.2 %)	97 (89.8 %)	108

Tabla 4.7: Distribución de la variable `fumador` según `sexo`, mostrando frecuencias absolutas y porcentajes por fila.

Los valores entre paréntesis corresponden a los porcentajes por fila, lo que permite comparar directamente la proporción de fumadores dentro de cada grupo de `sexo`. Se observa que el 6.0 % de las mujeres son fumadoras, mientras que en los hombres esta proporción asciende a 13.8 %.

El test de Fisher y su alternativa.

Para determinar si la diferencia observada en las proporciones es estadísticamente significativa, se debe realizar una prueba de independencia. En *jamovi*, existen varias opciones en la pestaña `Pruebas`:

- **Test de χ^2**: Adecuado para grandes muestras, pero su precisión disminuye si hay celdas con valores esperados bajos.

- **Test exacto de Fisher**: Proporciona resultados exactos sin aproximaciones. Gracias al avance en la capacidad de cómputo, actualmente este test puede aplicarse sin problemas salvo que se trabaje con bases de datos grandes, en cuyo caso podría no llegar a calcular una respuesta y bloquearse temporalmente el ordenador. Se recomienda, usarse por defecto si el ordenador puede soportar el cálculo. En caso contrario, utilizar el test χ^2.

4. Seleccionar las pruebas estadísticas adecuadas en la pestaña `Pruebas`, marcando χ^2 o `Test exacto de Fisher`.

5. En la sección `Medidas Comparativas`, activar `Diferencia de proporciones` e `Intervalos de confianza` con un nivel del 95 %, asegurándose de que la comparación se haga por filas (Figura 4.9).

   Nota 4.2. Se indica la comparación por filas porque la variable explicativa, el `sexo`, está en las filas.

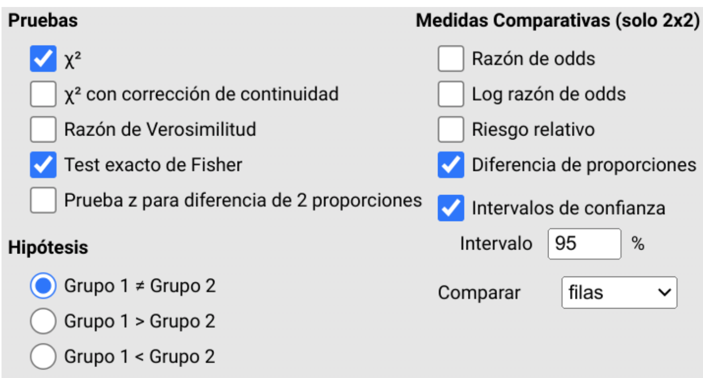


Figura 4.9: Selección de pruebas y medidas comparativas en *jamovi*.

Los resultados del test se presentan en la Tabla 4.8.

Prueba	***p*-valor**
Test exacto de Fisher	0.217

Tabla 4.8: Resultados del Test exacto de Fisher.

Además, los intervalos de confianza al 95 % para la diferencia de proporciones se presentan en la Tabla 4.9.

Medida	**Valor**	**Inferior**	**Superior**
Diferencia entre 2 proporciones	-0.078[a]	-0.188	0.033

Tabla 4.9: Intervalo de confianza del 95 % para la diferencia entre proporciones de fumadores por sexo.

Interpretación de los resultados

El valor $p = 0{,}217$ obtenido en el Test exacto de Fisher indica que no hay suficiente evidencia estadística para rechazar la hipótesis nula de independencia entre `sexo` y `fumador`. Es decir, la diferencia observada en las proporciones de fumadores entre hombres y mujeres podría explicarse por el azar y no por la existencia de un efecto real.

En consonancia, el intervalo de confianza del 95 % para la diferencia entre proporciones ($[-0{,}188, 0{,}033]$) contiene el valor 0, indicando que la ausencia de diferencias es creíble

Por lo tanto, aunque la proporción de fumadores es mayor en hombres (13.8 %) que en mujeres (6.0 %), la diferencia no es lo suficientemente grande como para considerarla estadísticamente significativa. En otras palabras, con solo la evidencia de esta muestra, no podemos concluir que el hábito de fumar esté asociado de manera con el sexo.

Reporte de resultados en un artículo científico

Se evaluó la asociación entre el sexo y el hábito de fumar mediante el Test exacto de Fisher. La proporción de fumadores fue del 13.8 % en hombres y del 6.0 % en mujeres, sin encontrar diferencias estadísticamente significativas entre ambos grupos (p = 0.217, diferencia = $-0{,}078\,\%$, IC 95 % $[-0{,}188\,\%, 0{,}033\,\%]$).

4.5. Evaluación de la normalidad

En estadística inferencial con variables cuantitativas, muchas pruebas paramétricas, como la *t de Student* o el *ANOVA*, asumen que los datos siguen una distribución normal. Por tanto, evaluar la normalidad de una variable cuantitativa en una muestra es esencial, porque el incumplimiento de este supuesto puede llevar a aplicar test sin sus hipótesis necesarias y estos a conclusiones erróneas.

Para ello, se plantean las siguientes hipótesis:

- H_0: La distribución de la variable es normal.
- H_1: La distribución de la variable no es normal.

Es importante notar que **no rechazar** la hipótesis nula no implica que la variable sea normal, sino que los datos no son lo suficientemente distintos de una distribución normal como para detectarlo con la muestra disponible. En este caso, pueden aplicarse técnicas adecuadas para variables normales.

Existen diversos contrastes para evaluar la normalidad de una variable. La recomendación habitual es utilizar el test de Shapiro-Wilk, especialmente en muestras pequeñas.

En *jamovi*, el uso del test de Shapiro-Wilk ya se ha ilustrado en la Sección 3.2 para el peso en función del sexo, y se puede realizar en el ménu de `Descriptivas`, seleccionando la opción de `Normalidad - Shapiro Wilk`.

En ese caso, la Tabla 3.1 muestra los resultados de ese test de normalidad para dos muestras: el peso de los homrbres, con un p-valor de 0,357, y el peso de las mujeres con p-valor igual a 0,073. Nótese, por tanto, que no podemos rechazar la normalidad ni para hombres ni para mujeres. Además, gráficamente se podía observar en la Figura 3.3 que las distribuciones no parecen muy diferentes de la normal, en concordancia con el resultados de los test.

Nota 4.3. Aquí, hemos ilustrado directamente el uso del test de Shapiro-Wilk para una variable cuantitativa en dos grupos. No obstante, también se puede hacer sin distinguir entre grupos si no se completa la opción de `Separa por`. Para ver un ejemplo, véase la Sección 2.6.3.

Adicionalmente, también se pueden emplear herramientas gráficas, como el *Gráfico Q-Q* o *Gráfico Cuantil-Cuantil*.

- `Análisis` → `Exploración` → `Descriptivas`
- Seleccionar la variable de interés (por ejemplo, `peso`), separada por grupos si se desea (por ejemplo, `sexo`) y activar, en la sección de *Gráficos*, la opción `Q-Q Plot` (Figura 4.10).

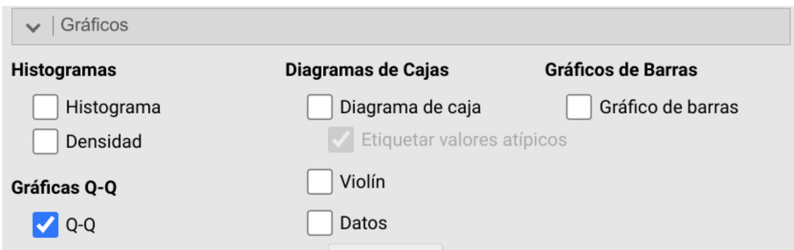


Figura 4.10: Menú de selección de gráficos en *jamovi*. Para evaluar la normalidad se debe activar la opción `Q-Q`.

- Si los puntos del `Q-Q Plot` se alinean con la diagonal, la distribución es aproximadamente normal. Véase como ejemplo la Figura 4.11.

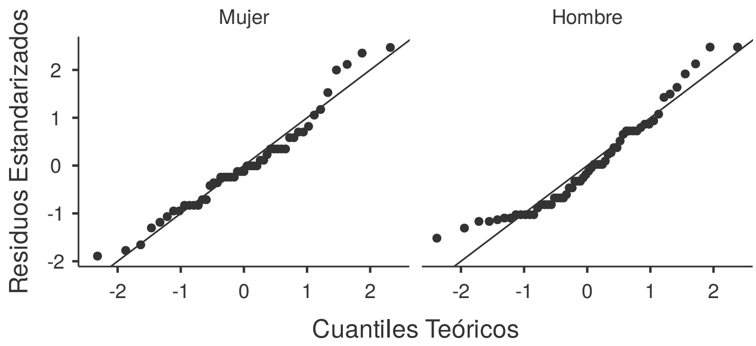


Figura 4.11: Gráficos Q-Q para evaluar la normalidad en función del sexo. Se presentan los residuos estandarizados frente a los cuantiles teóricos para hombres y mujeres. La proximidad de los puntos a la línea diagonal sugiere normalidad en los datos.

4.6. Una Variable Cuantitativa

En análisis de datos, es muy frecuente contrastar si la media de una variable cuantitativa en una muestra difiere de un valor teórico predefinido. Sin embargo, como se ha explicado en la introducción, comprobar las hipótesis de los test es lo que garantiza la validez de sus resultados.

Las pruebas paramétricas, como la prueba t de Student, asumen que la distribución de los datos es normal y, adicionalmente, que las varianzas entre los grupos son iguales si se trabaja con dos muestras independientes como en la Sección 4.7. Estas pruebas se basan en modelos matemáticos de distribuciones de probabilidad para estimar parámetros poblacionales. Cuando se cumplen estos supuestos, las pruebas paramétricas ofrecen el máximo poder estadístico para detectar diferencias en las medias.

En contraste, las alternativas no paramétricas, como el test de Wilcoxon, utilizan enfoques basados en rangos y no requieren asumir una distribución específica de los datos. Estas pruebas son más robustas frente a violaciones de normalidad, aunque presentan menor sensibilidad para detectar efectos sutiles.

Por ello, en esta sección se describe el flujo de trabajo para realizar este análisis en *jamovi*, desde la evaluación de supuestos hasta la interpretación de los resultados.

- Si los datos siguen una distribución normal, o la muestra es grande ($n > 30$), se emplea la **prueba t de una muestra**.
- Si los datos no son normales y la muestra es pequeña ($n < 30$), se utiliza la **prueba de rangos de Wilcoxon**.

4.6.1. Test t de una muestra

El test t permite contrastar si la media poblacional, μ, de una variable cuantitativa difiere de un valor teórico dado. Las hipótesis del contraste son, para un valor de referencia cualquiera μ_0:

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$

En particular, este test se calcula a partir del estadístico

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

Este estadístico se distribuye según la distribución t *de Student* (de ahí el nombre del test) cuando la variable sigue una distribución normal o el tamaño muestral es suficientemente grande. Nótese que se calcula simplemente a partir de la media muestral ($\bar{X}$), el tamaño de la muestra (n) y la desviación típica muestral (S), cantidades que ya han sido halladas en el Capítulo 2.

Implementación

0. Comprobar, usando las herramientas de estadística descriptiva (Capítulo 2) y el test de normalidad (Sección 4.5), si se cumplen las hipótesis para realizar este test[1]. Si no se cumplen, hay que aplicar la alternativa no paramétrica (Sección 4.6.3).
1. Ir a `Análisis` → `Pruebas t` → `Prueba t para una muestra`. Ver Figura 4.12.

Figura 4.12: Menú para ejecutar el test t de una muestra.

2. Arrastrar la variable cuantitativa al cuadro de `Variables de prueba`.
3. En `Valor de prueba`, ingresar el valor teórico a comparar.
4. Opcionalmente, activar `Descriptivas` para obtener medidas de tendencia central y dispersión.

[1] En este test, otra opción es, después del paso 2, seleccionar las casillas de `Prueba de normalidad` y `Descriptivas`

En la Tabla 4.10 se muestran los resultados de test t aplicado a la muestra de alturas con un valor teórico de 170 cm.

Variable	Estadístico t	gl	p-valor
Altura	3.247	105.000	0.002

Tabla 4.10: Resultados de la prueba t de una muestra para la variable Altura con valor de prueba 170.

Puesto que el p-valor es menor a 0,05, se rechaza la hipótesis nula en favor de la alternativa de que la media poblacional es diferente de 170 cm.

4.6.2. Cálculo del intervalo de confianza de la media

Para calcular el intervalo de confianza de la media en *jamovi*, se deben seguir los mismos pasos que para aplicar el test t (Sección 4.6.1), utilizando, en `Valor de prueba` el valor 0. Además, en la sección **Estadísticas Adicionales**, marcar (Figura 4.13):

- `Diferencia de medias`
- `Intervalo de confianza` (ajustado al nivel deseado, por defecto 95 %).
- `Descriptivas` para obtener medidas adicionales de tendencia central y dispersión.

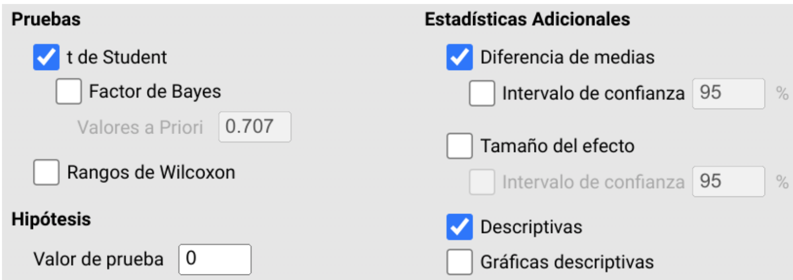


Figura 4.13: Opciones en *jamovi* para calcular el intervalo de confianza de la media.

Los resultados obtenidos en la prueba t de una muestra se presentan en la Tabla 4.11.

Variable	Estadístico t	gl	p-valor	Diferencia de medias	Inferior	Superior
Altura	174.811	105.000	$< .001$	173.217	171.252	175.182

Tabla 4.11: Resultados de la prueba t de una muestra con intervalo de confianza del 95 % y valor de prueba igual a 0.

Interpretación: El intervalo de confianza del 95 % para la media de la altura es [171,252, 175,182]. Esto significa que, con un 95 % de confianza, la media poblacional se encuentra dentro de este rango.

4.6.3. Prueba de Wilcoxon e Intervalo de Confianza para la mediana

Cuando la variable no sigue una distribución normal y el tamaño muestral es pequeño, la prueba t de una muestra no es apropiada. En estos casos, se utiliza la prueba de rangos con signo de Wilcoxon, que es una alternativa no paramétrica para comparar la mediana de la muestra con un valor teórico.

Las hipótesis del contraste son, para un valor de referencia cualquiera M_0:

- H_0 : Mediana $= M_0$
- H_1 : Mediana $\neq M_0$

Para calcular el intervalo de confianza o test utilizando la prueba de Wilcoxon, se siguen los mismos pasos que en la prueba t (Sección 4.6.1), pero seleccionando la opción `Rangos de Wilcoxon` en lugar de `t de Student` en la sección `Pruebas` (Figura 4.14).

Para el intervalo de confianza, en la sección `Estadísticas Adicionales` se deben marcar las siguientes opciones:

- `Valor de prueba` igual a 0.

  Nota 4.4. Si por el contrario se desea contrastar si la mediana vale, por ejemplo, 170 cm, utilizar `Valor de prueba` igual a 170. Utilizar 0 sirve para para calcular el IC.

- `Diferencia de medias`
- `Intervalo de confianza` (ajustado al nivel deseado, por defecto 95 %).
- `Descriptivas` para obtener medidas adicionales de tendencia central y dispersión.

Figura 4.14: Opciones en *jamovi* para calcular el intervalo de confianza con la prueba de Wilcoxon.

Los resultados obtenidos en la prueba de Wilcoxon se presentan en la Tabla 4.12.

Nota 4.5. Importante. Aunque en la Tabla 4.12 se hable de *Diferencia de medias*, en realidad lo que se trabaja con este test son las medianas.

Variable	Estadístico W	p-valor	Diferencia de medias	Inferior	Superior
Altura	5671.000	$< .001$	173.500	171.000	175.000

Tabla 4.12: Resultados de la prueba de Wilcoxon para una muestra con intervalo de confianza del 95 %.

Interpretación:

- El IC del 95 % para la mediana de la altura es $[171, 175]$. Esto significa que, con un 95 % de confianza, la mediana poblacional de la altura se encuentra dentro de este rango.
- Como en el caso de la prueba t, si el intervalo no contiene el valor teórico, hay evidencia para rechazar H_0. En este caso, el intervalo no incluye, por ejemplo, a 170, indicando que la mediana de la altura es significativamente distinta de este valor.
- Se puede comprobar que si utiliza un valor de prueba de 170, el resultado arroja un p-valor igual a 0,001, en consonancia con la conclusión extraída a través del intervalo de confianza.

4.7. Comparación de Dos Muestras Independientes

Otro de los problemas más comunes en el análisis de datos es determinar si existen diferencias significativas entre dos grupos. Cuando la variable de interés es cuantitativa, se emplean pruebas estadísticas específicas que comparan medidas resumen, como la media o la mediana, para evaluar dichas diferencias.

Sin embargo, para elegir el test estadístico más adecuado[2], es fundamental verificar los siguientes supuestos:

- **Normalidad en cada grupo:** Se evalúa mediante el test de Shapiro-Wilk o gráficos Q-Q. Ver Sección 4.5.
- **Homoscedasticidad (dispersión igual en ambos grupos):** Se verifica con el test de Levene para igualdad de varianzas (en otros software también está disponible el test de Bartlett).
- **Escala de Medición:** Las pruebas paramétricas requieren datos continuos, mientras que en las pruebas no paramétricas, como el test de la U de Mann-Whitney, basta con que los datos sean ordinales, manejando igualmente variables cuantitativas altamente asimétricas.

[2] Un test se considera adecuado cuando sus supuestos se cumplen y, además, maximiza la potencia estadística de la prueba.

Si se incumple alguno de estos supuestos, la práctica habitual suele ser navegar en en el siguiente árbol de decisión, esquematizado en la Figura 4.15:

1. Si el tamaño de las muestras (ambas) es grande ($n > 30$ en cada grupo) o si se puede asumir normalidad en ambos grupos, se recomienda realizar el **test *t*** de muestras independientes (Sección 4.7.1).
2. Si el tamaño de alguna muestra es pequeño, pero se puede asumir la normalidad (Sección 4.5), también se recomienda realizar el **test *t***.
3. Si se realiza el **test *t***, se ha de comprobar la igualdad de varianzas con el test de Levene (Sección 4.7.1). Si se puede asumir la igualdad de varianzas, continuar con el test t. En caso contrario, utilizar el **test *t* con la correción de Welch**.
4. Si falla la hipótesis de normalidad en algún grupo y alguna muestra es pequeña, se recomienda constrastar la igualdad de medianas con el **test de la U de Mann-Whitney** (Sección 4.7.2).

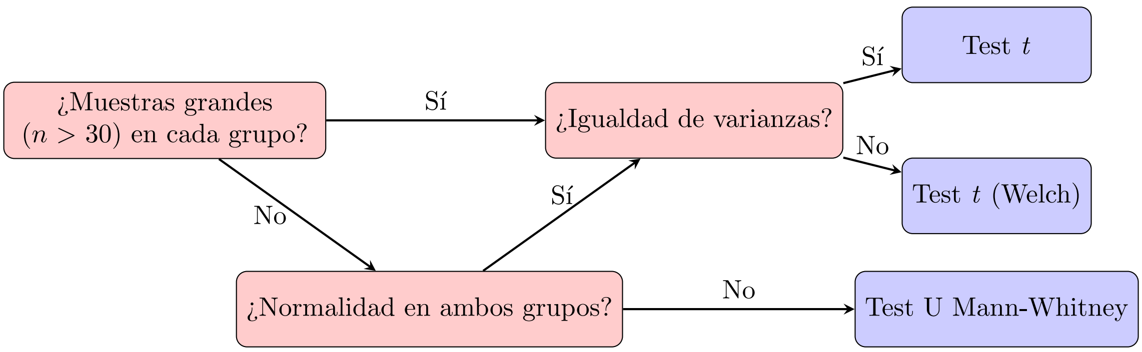


Figura 4.15: Esquema de selección del test estadístico adecuado en la comparación de una variable cuantitativa en dos muestras independientes.

4.7.1. Test t de Muestras Independientes

A continuación, se presenta el planteamiento del test t, uno de los más utilizados, junto con el cálculo de su estadístico. En la actualidad, estas pruebas suelen aplicarse mediante software especializado, como *jamovi*. Sin embargo, como se mostrará, todas las cantidades involucradas pueden calcularse manualmente de manera sencilla. Además, en muchos artículos o reportes, es común encontrar las medidas resumen necesarias para aplicar el test sin necesidad de disponer de los datos originales, bastando con realizar algunos cálculos aritméticos básicos. En estos casos, si no se cuenta con un software estadístico, los valores obtenidos deben compararse con las tablas de la distribución de la t para interpretar los resultados.

De este modo, dadas dos muestras aleatorias simples $(X_1, X_2, \dots, X_m)$ e $(Y_1, Y_2, \dots, Y_n)$ provenientes de dos poblaciones independientes con medias μ_X y μ_Y, respectivamente, el objetivo es contrastar la hipótesis nula:

$$H_0 : \mu_X = \mu_Y$$

frente a la hipótesis alternativa:

$$H_1 : \mu_X \neq \mu_Y.$$

Si se asume igualdad de varianzas ($\sigma_X^2 = \sigma_Y^2$), el estadístico de contraste, que tiene una distribución de t de Student, se define como:

$$t = \frac{\bar{X} - \bar{Y}}{S_C\sqrt{\frac{1}{m} + \frac{1}{n}}},$$

donde

$$S_C = \sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}}.$$

Cuando no se puede asumir igualdad de varianzas, el estadístico se calcula como:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}.$$

En este caso, los grados de libertad se ajustan mediante la corrección de Welch.

Para aplicar el test en *jamovi*, se siguen los siguientes pasos:

1. Seleccionar `Pruebas t` → `Prueba t para Muestras Independientes`, como se muestra en la Figura 4.16.

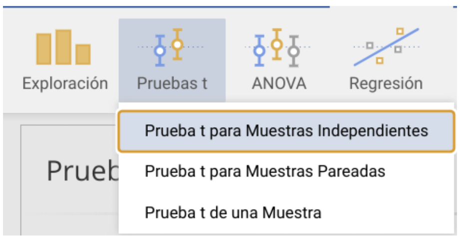


Figura 4.16: Selección de la prueba t para muestras independientes en *jamovi*.

2. Arrastrar la variable dependiente `Altura` al cuadro de `Variables Dependientes` y la variable categórica `Sexo` al cuadro de `Variables de agrupación`, tal como se observa en la Figura 4.17.

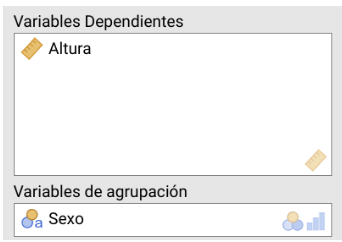


Figura 4.17: Asignación de variables dependientes y de agrupación en *jamovi*.

3. Configurar la prueba `t de Student`, activar la opción de `Diferencia de medias` e `Intervalo de confianza`, y comprobar la igualdad de varianzas con la `Prueba de homogeneidad` (Figura 4.18).

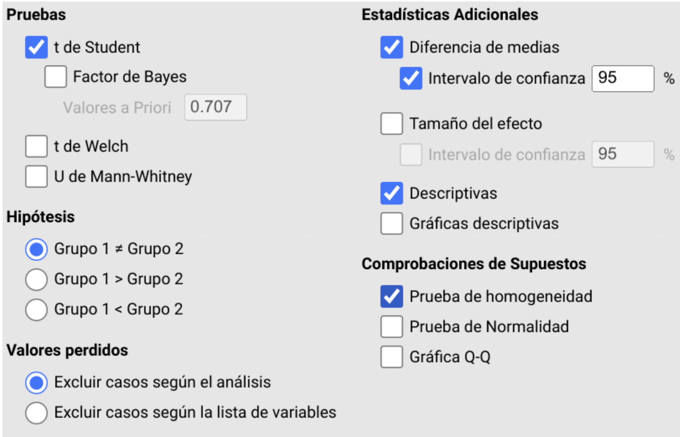


Figura 4.18: Configuración del test t en *jamovi*.

4. Si se desea realizar la *corrección de Welch*, en casos en los que no se puede asumir la igualdad de varianzas, tan solo hay que marcar dicha opción. Asimismo, para aplicar la técnica no paramétrica del test de Mann-Whitney, de nuevo, basta con marcar la opción correspondiente.

Ejemplo de test t de dos muestras e interpretación

Supongamos que se pretende contrastar si la altura de hombres y mujeres es distinta en media.

Podemos notar en la Tabla 4.13 que cada grupo cuenta con un número no pequeño de observaciones ($n > 30$), lo que permite aplicar el test t.

Grupo	N	Media	Mediana	DE	EE
Mujer	49	165.776	165	7.592	1.085
Hombre	57	179.614	180	7.459	0.988

Tabla 4.13: Estadísticos descriptivos de la altura en cada grupo.

Adicionalmente, la prueba de Levene (Tabla 4.14) muestra que no hay evidencia suficiente para rechazar la igualdad de varianzas ($p = 0{,}747$), lo que respalda el uso del test t de Student sin corrección de Welch.

Variable	F	gl1	gl2	p
Altura	0.105	1	104	0.747

Tabla 4.14: Resultados de la prueba de Levene para homogeneidad de varianzas.

Finalmente, los resultados de la prueba t de muestras independientes se presentan en la Tabla 4.15. Se observa que la diferencia de medias entre hombres y mujeres es significativa ($p < 0{,}001$), con un intervalo de confianza del 95 % que no incluye el valor cero. En particular, la diferencia de medias estimada es de $-13{,}839$ cm, lo que indica que, en promedio, la altura de los hombres es mayor que la de las mujeres en esta muestra. Este resultado se corrobora con el intervalo de confianza, que va de $-16{,}744$ cm a $-10{,}933$ cm, descartando con nivel de confianza del 95 % la posibilidad de que no exista diferencia en la población.

Estadísticos de la prueba				**Intervalo de confianza (95 %)**			
Test	**Estadístico**	**gl**	p	**Diferencia de medias**	**EE de la diferencia**	**Inferior**	**Superior**
t de Student	-9.446	104.000	<0.001	-13.839	1.465	-16.744	-10.933

Tabla 4.15: Resultados del test t de muestras independientes con intervalo de confianza al 95 %.

Reporte de resultados en un artículo científico

A la hora de reportar los resultados, es práctica estándar reportar el IC de la diferencia de medias. Por ejemplo:

Los hombres presentaron una altura significativamente mayor que las mujeres con una diferencia de medias de $-13{,}84$ cm ($IC_{95\,\%}$: $[-16{,}74,\ -10{,}93]$; $p < 0{,}001$).

Ejemplo de test t con corrección de Welch

Si queremos analizar, por ejemplo, si el peso medio de hombres y mujeres es el mismo en media, es necesario comprobar primero el supuesto de igualdad de varianzas. Como se observa en la Tabla 4.16, la prueba de Levene arroja un valor de $p < 0{,}001$, lo que indica que no se puede asumir homogeneidad de varianzas.

Variable	**F**	**gl1**	**gl2**	p
Peso	14.777	1	105	<0.001

Tabla 4.16: Prueba de Levene para homogeneidad de varianzas en el peso.

En estos casos, es recomendable emplear el test t con la corrección de Welch en lugar del test t estándar, ya que esta corrección ajusta los grados de libertad y permite una estimación más robusta de la diferencia de medias.

Estadísticos de la prueba				**Intervalo de confianza (95 %)**			
Test	**Estadístico**	**gl**	p	**Diferencia de medias**	**EE de la diferencia**	**Inferior**	**Superior**
t de Welch	-7.889	94.931	<0.001	-17.604	2.232	-22.034	-13.174

Tabla 4.17: Resultados del test t con corrección de Welch para la variable peso.

En este caso, podríamos reportar el resultado de la siguiente manera: los hombres presentaron un peso significativamente mayor que las mujeres ($p < 0{,}001$), con una diferencia media de $-17{,}604$ kg (IC 95 %: $[-22{,}034,\ -13{,}174]$).

4.7.2. Alternativa no parámetrica: Test de la U de Mann-Whitney

Si queremos analizar si el pulso en reposo es diferente entre fumadores y no fumadores, es necesario verificar primero los supuestos del test t de muestras independientes. Como se observa en la Tabla 4.18, el tamaño del grupo de fumadores es pequeño ($n = 11$), mientras que el grupo de no fumadores cuenta con 97 observaciones.

Grupo	N	Media	Mediana	DE	EE
Sí	11	77.545	76.000	9.575	2.887
No	97	74.763	75.000	11.751	1.193

Tabla 4.18: Descriptivos de pulso en reposo según hábito de fumar.

En este caso, por tanto, la normalidad de la variable debe comprobarse en cada grupo antes de aplicar el test, lo que se hace en el apartado de análisis descriptivo y no en el menú del test, tal como se explicó en las Secciones 3.2 y 4.5.

Como se muestra en la Tabla 4.19, la prueba de Shapiro-Wilk indica que la normalidad no puede asumirse para el grupo de no fumadores ($p = 0{,}043$).

Fumador	W de Shapiro-Wilk	Valor p
Sí	0.916	0.284
No	0.973	0.043

Tabla 4.19: Prueba de Shapiro-Wilk para normalidad en pulso en reposo.

Ya que el tamaño del grupo de fumadores es pequeño y no se puede garantizar la normalidad en ambos grupos, el test t no es adecuado y se debe utilizar el test no paramétrico de Mann-Whitney.

Nota 4.6. Es frecuente, en una situación como esta, aplicar también el test t, dado que el grupo que viola normalidad es de tamaño grande.

Hipótesis del test de Mann-Whitney

Como se opta por el test de Mann-Whitney, los resultados deben interpretarse en términos de la **mediana** en lugar de la media, aunque *jamovi* reporte ambas en la tabla descriptiva y escriba medias en el resultados del test.

En realidad, las hipótesis que se contrastan son:

- **Hipótesis nula (H_0):** Las distribuciones de los dos grupos son iguales en términos de posición (mediana).
- **Hipótesis alternativa (H_1):** Las distribuciones de los dos grupos difieren en posición, es decir, existe una diferencia en la mediana entre ambos grupos.

Para realizar este test en *jamovi* se deben seguir los mismos pasos que para el test t estándar (Sección 4.7.1), marcando en el último paso la opción: `U de Mann-Whitney`.

Los resultados del test de Mann-Whitney se presentan en la Tabla 4.20. Se observa que la diferencia entre grupos no es significativa ($p = 0{,}404$), con un intervalo de confianza del 95 % para la diferencia de medianas que incluye el cero.

Estadísticos de la prueba				**Intervalo de confianza (95 %)**		
Test	**Estadístico**	p	**Diferencia de medianas**	**EE de la diferencia**	**Inferior**	**Superior**
U de Mann-Whitney	451.000	0.404	3.000	—	-4.000	10.000

Tabla 4.20: Resultados del test de Mann-Whitney para la variable pulso en reposo según el hábito de fumar.

Reporte de resultados en un artículo científico
No se observaron diferencias significativas en el pulso en reposo entre fumadores y no fumadores ($p = 0{,}404$), mostrando una diferencia de medianas de 3,0 ppm ($IC_{95\,\%}$: $[-4{,}0,\ 10{,}0]$).

Nótese que como el intervalo de confianza contiene al cero y en consecuencia el p-valor del test no es significativo.

4.8. Comparación de Tres o Más Muestras Independientes

En esta sección se aborda la comparación de una variable cuantitativa en tres o más grupos independientes. Este caso constituye una extensión natural de las técnicas presentadas en la Sección 4.7. Como se ilustra en el esquema de la Figura 4.19, los supuestos a verificar (normalidad, homogeneidad de varianzas y tipo de variable) y las decisiones a tomar siguen el mismo razonamiento que en la comparación de dos grupos.

En este contexto, la normalidad y el tamaño muestral deben evaluarse individualmente para cada grupo. El test ANOVA (*ANalysis Of VAriance*) estándar o de Fisher (Sección 4.8.1) es el equivalente al test t de dos muestras, y cuenta con una variante, el ANOVA de Welch (Sección 4.8.2), que se aplica cuando no se cumple el supuesto de homogeneidad de varianzas. De manera también análoga, cuando los datos no cumplen los requisitos para aplicar ANOVA, la alternativa no paramétrica es el llamado test de Kruskal-Wallis (Sección 4.8.3).

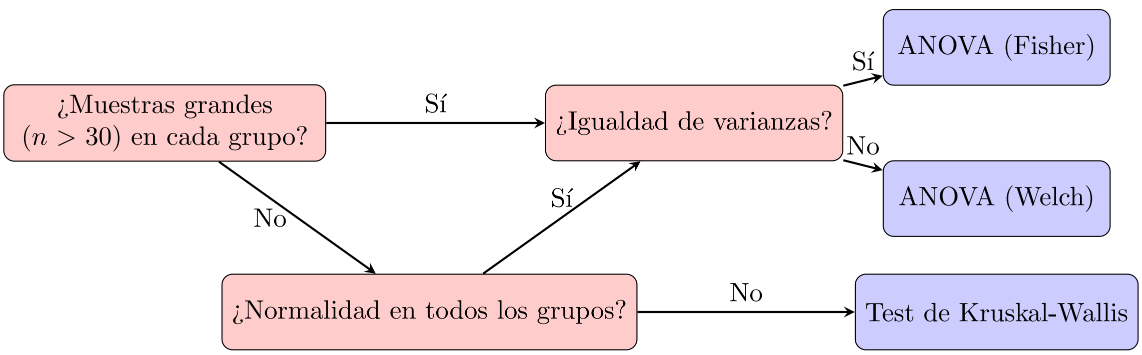


Figura 4.19: Esquema de selección del test estadístico adecuado en la comparación de una variable cuantitativa en tres o más muestras independientes.

Nota 4.7. Cuando se aplica ANOVA a dos grupos, el resultado es equivalente al test t, del mismo modo que el test de Kruskal-Wallis coincide con el test de Mann-Whitney en dicha situación.

¿Qué hacer si encontramos diferencias significativas?

Si el test ANOVA o el test de Kruskal-Wallis revelan diferencias significativas entre los grupos, es necesario realizar una prueba **post-hoc** para determinar específicamente entre qué grupos se encuentran dichas diferencias.

Las pruebas post-hoc permiten realizar comparaciones múltiples controlando el error de tipo I, ya que en un análisis con varios grupos el riesgo de obtener falsos positivos aumenta con cada comparación adicional.

La elección de la prueba post-hoc depende de la homogeneidad de varianzas. De los disponibles en *jamovi* se recomiendan:

- Si se ha asumido igualdad de varianzas en el ANOVA, se recomienda el **test de Tukey**.
- Si no se ha asumido homogeneidad, es preferible el **test de Games-Howell**.
- En el caso del test de Kruskal-Wallis, se suelen utilizar alternativas como el **test DSCF** (Dwass-Steel-Critchlow-Fligner).

4.8.1. ANOVA (Estándar de Fisher)

En esta sección vamos a ilustrar la aplicación del test ANOVA comparando si el nivel medio del `Pulso` se puede considerar igual en los tres niveles dados por el `IMC - Niveles`.

Las hipótesis del test son por tanto:

- **Hipótesis nula (H_0):** No hay diferencias significativas en la media del `Pulso` entre los niveles del `IMC - Niveles`, es decir, todas las medias poblacionales (μ) son iguales:
$$H_0 : \mu_{\text{Infrapeso}} = \mu_{\text{Normopeso}} = \mu_{\text{Sobrepeso}}.$$
- **Hipótesis alternativa (H_1):** Al menos una de las medias difiere de las demás, lo que indica la existencia de diferencias significativas entre los niveles del `IMC - Niveles`:
$$H_1 : \exists\, i, j \quad \text{tal que} \quad \mu_i \neq \mu_j.$$

El primer paso consiste en comprobar el tamaño de las muestras y la normalidad de la variable en cada grupo mediante el menú de análisis descriptivo en *jamovi*. La Tabla 4.21 muestra el número de observaciones en cada grupo y los resultados de la prueba de Shapiro-Wilk. Como se observa, hay dos grupos con un tamaño muestral pequeño ($n < 30$), sin embargo, no se rechaza la hipótesis de normalidad en ninguno de ellos, ya que los p-valores son mayores a 0,05. Por lo tanto, podemos aplicar el test ANOVA paramétrico.

Grupo	**N**	**W de Shapiro-Wilk**	**Valor p**
Infrapeso	15	0.964	0.769
Normopeso	72	0.993	0.958
Sobrepeso	19	0.957	0.514

Tabla 4.21: Estadísticos descriptivos y prueba de normalidad de Shapiro-Wilk para la variable Pulso1.

El segundo supuesto a evaluar para decidir si es necesario aplicar la corrección de Welch o no, es la homogeneidad de varianzas. En *jamovi*, esta comprobación se puede realizar directamente dentro del menú del test ANOVA.

1. Seleccionar `ANOVA` → `ANOVA de Un Factor`, como se muestra en la Figura 4.20.

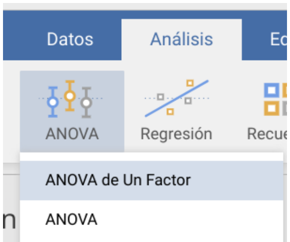


Figura 4.20: Selección del ANOVA de Un Factor en *jamovi*.

2. Arrastrar la variable dependiente `Pulso1` al cuadro `Variables Dependientes` y la variable categórica `IMC - Quetelet` al cuadro `Variables de Agrupación`, como se observa en la Figura 4.21.

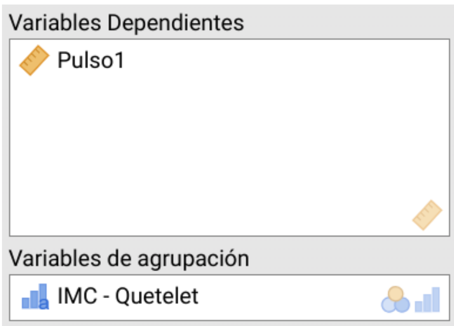


Figura 4.21: Asignación de variables en *jamovi* para ANOVA.

3. En las opciones del menú, se debe marcar `Prueba de homogeneidad` (Figura 4.22). Los resultados de la prueba de Levene se presentan en la Tabla 4.22, donde se observa un valor de $p = 0{,}695$, indicando que no hay evidencia suficiente para rechazar la igualdad de varianzas entre los grupos.

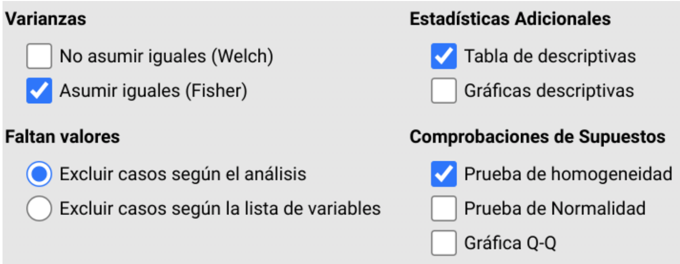


Figura 4.22: Configuración del ANOVA en *jamovi*.

Variable	F	gl1	gl2	p
Pulso1	0.365	2	103	0.695

Tabla 4.22: Prueba de Levene para homogeneidad de varianzas en la variable Pulso1.

Dado que se cumplen ambos supuestos (normalidad y homogeneidad de varianzas), se puede proceder con la aplicación del ANOVA estándar:

4. Marcar la opción `Asumir iguales (Fisher)`, ya que la prueba de Levene no rechazó la igualdad de varianzas.

Los resultados se presentan en la Tabla 4.23, donde se observa un valor de $p = 0{,}126$, lo que indica que no hay evidencia suficiente para concluir que el pulso medio difiere entre los distintos grupos de IMC.

Variable	F	gl1	gl2	p
Pulso1	2.11	2	103	0.126

Tabla 4.23: Resultados del ANOVA de Un Factor (Fisher) para la variable `Pulso1`.

Prueba Post-Hoc

En el anterior ejemplo, las diferencias no eran significativas, por lo que no tiene sentido realizar un análisis post-hoc.

Para ilustar esta sección, se puede comprobar que si se estudia el `Pulso1` en función del `Nivel del ejercicio`, el resultado del ANOVA indica diferencias significativas entre los grupos de niveles de ejercicio ($p = 0{,}048$), por lo que se debe proceder con un análisis post-hoc para identificar entre qué grupos se encuentran estas diferencias. En este caso, puesto que se asumió homogeneidad de varianzas (prueba de Levene con $p = 0{,}865$), se aplica la prueba de Tukey.

Para activarla en *jamovi*, se debe seleccionar la opción `Tukey (varianzas iguales)` en la sección de `Pruebas Post-Hoc` (Figura 4.23).

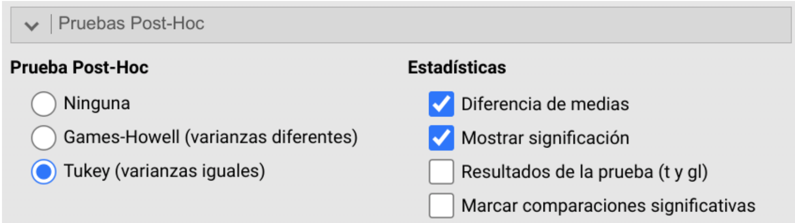


Figura 4.23: Selección de la prueba post-hoc de Tukey en *jamovi*.

Los resultados de la prueba de Tukey se presentan en la Tabla 4.24. Se observa que la única comparación con diferencia significativa es entre los grupos de *Baja* y *Alta* intensidad de ejercicio ($p = 0{,}019$), mientras que las demás comparaciones no alcanzan significación estadística.

Comparación	**Diferencia de Medias**	**Valor p**
Baja - Moderada	3.878	0.236
Baja - Alta	9.708	0.019
Moderada - Alta	5.831	0.195

Tabla 4.24: Resultados de la prueba post-hoc de Tukey para la variable Pulso1 según los niveles de ejercicio.

Dado que solo la comparación entre los grupos de *Baja* y *Alta* intensidad de ejercicio muestra una diferencia significativa ($p = 0{,}019$), se concluye que el pulso medio es mayor en el grupo de menor ejercicio comparado con el de mayor ejercicio. No obstante, no se encuentran diferencias estadísticamente significativas entre los niveles de ejercicio *Moderado* y los otros dos grupos.

Reporte de resultados en un artículo científico

El ANOVA de un factor reveló diferencias significativas en el pulso en reposo entre los grupos de IMC ($p = 0{,}022$). La prueba post-hoc de Tukey mostró que el grupo con infrapeso presentó un pulso significativamente mayor que el grupo con sobrepeso ($p = 0{,}019$).

4.8.2. ANOVA de Welch

Si la homogeneidad de varianzas no se puede asumir (prueba de Levene con $p < 0{,}05$), se recomienda el ANOVA de Welch, que ajusta los grados de libertad para compensar las varianzas desiguales. Para activarlo en *jamovi*, se sigue el mismo procedimiento que el ANOVA estándar, pero marcando la opción `No asumir iguales (Welch)` en `Varianzas`, como se muestra en la Figura 4.24.

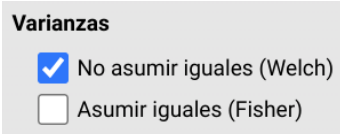


Figura 4.24: Selección del ANOVA de Welch en *jamovi*.

A diferencia del ANOVA estándar, este método permite realizar una comparación más robusta cuando los grupos presentan varianzas heterogéneas.

Puesto que en este contexto no se asume igualdad de varianzas, la prueba post-hoc recomendada es la de **Games-Howell**, en lugar de la de Tukey. Para activarla en *jamovi*, se debe seleccionar la opción `Games-Howell (varianzas diferentes)` dentro del menú de pruebas post-hoc, como se observa en la Figura 4.25.

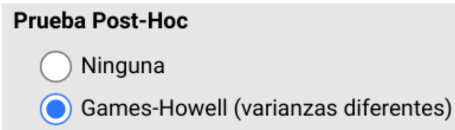


Figura 4.25: Selección de la prueba post-hoc de Games-Howell en *jamovi*.

La interpretación de los resultados del ANOVA de Welch, sus hipótesis y el resultado del test post hoc son completamente análogos a los del ANOVA estándar de Fisher.

4.8.3. Test de Kruskal-Wallis

En esta sección, se analiza si la edad difiere significativamente según el nivel de pulso (< 60, $60 - 80$, > 80), lo que puede servir, argumentando a la inversa, para discernir la influencia de la edad en el pulso.

En primer lugar, tal y como se observa en la Tabla 4.25, el tamaño muestral de algún grupo es muy pequeño ($n < 30$), y además, dado que los tres grupos presentan valores de $p < 0{,}001$, se rechaza la hipótesis de normalidad en todos ellos. Por lo tanto, no queda justificado el uso de un test ANOVA y se utilizará un test de Kruskal-Wallis como alternativa no paramétrica.

Grupo de Pulso	N	W de Shapiro-Wilk	Valor p
Baja	7	0.616	$< 0{,}001$
Media	64	0.572	$< 0{,}001$
Alta	37	0.851	$< 0{,}001$

Tabla 4.25: Prueba de normalidad de Shapiro-Wilk para la variable Edad según nivel de pulso.

Como se trata de un test no paramétrico, la comparación se realiza en términos de medianas en lugar de medias. Las hipótesis que se contrastan son:

- **Hipótesis nula:** H_0: La distribución de la variable `Edad` es la misma en todos los grupos, es decir, las medianas son iguales.
- **Hipótesis alternativa:** H_1: Al menos un grupo presenta una mediana diferente a la de los demás.

Para llevar a cabo este análisis en *jamovi*, se siguen los siguientes pasos:

1. Seleccionar `Análisis` → `ANOVA` → `ANOVA de Un Factor -- Kruskal-Wallis`.

Figura 4.26: Selección del test de Kruskal-Wallis en *jamovi*.

2. Arrastrar la variable dependiente `Edad` al cuadro `Variables Dependientes` y la variable categórica `PulsoNivel` al cuadro `Variables de Agrupación`.

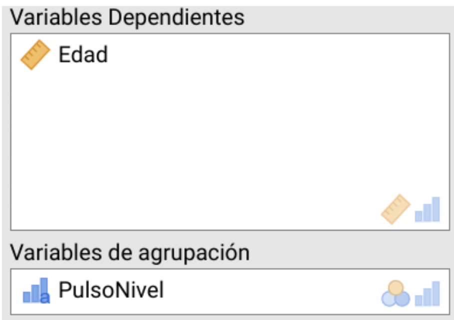


Figura 4.27: Asignación de variables en *jamovi* para el test de Kruskal-Wallis.

3. En el menú de configuración (Figura 4.28), se debe marcar la opción `Comparaciones por pares` si se desean realizar pruebas post hoc en caso de diferencias significativas.

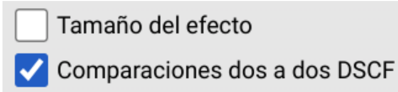


Figura 4.28: Menú para la realización de la prueba post hoc DSCF.

Los resultados del test de Kruskal-Wallis se presentan en la Tabla 4.26. Se observa un valor de $p = 0{,}468$, lo que indica que no hay evidencia suficiente para concluir que la edad difiera significativamente entre los grupos de pulso.

Variable	χ^2	**gl**	p
Edad	1.521	2	0.468

Tabla 4.26: Resultados del test de Kruskal-Wallis para la variable Edad según el nivel de pulso.

Además, dado que el resultado del test no es significativo ($p > 0{,}05$), no se requiere realizar comparaciones post hoc. No obstante, las realizamos a continuación solo a modo de ilustración. Los resultados se presentan en la Tabla 4.27, donde se observa que ninguna de las comparaciones entre parejas es significativa, lo que está en consonancia con el resultado global del test.

Comparación	**W**	p
Baja - Media	-0.979	0.768
Baja - Alta	-1.484	0.546
Media - Alta	-1.269	0.642

Tabla 4.27: Comparaciones por pares del test de Kruskal-Wallis para la variable `Edad` según `nivel de pulso`.

4.9. Comparación de Muestras Pareadas

4.9.1. Dos muestras relacionadas

En este apartado, analizamos la comparación de una variable cuantitativa en dos muestras pareadas. A diferencia de las muestras independientes, en este caso cada observación en un grupo tiene una correspondencia natural en el otro grupo (por ejemplo, mediciones antes y después de una intervención, lo que responde al esquema de pretest-postest).

El procedimiento a seguir depende de nuevo del tamaño muestral y la normalidad de la diferencia. La Figura 4.29 presenta un esquema de decisión para elegir la prueba estadística adecuada que recuerda inevitablemente a los flujos de decisión presentados en el resto de test t, pero sin necesidad de comprobar la homogeneidad de las varianzas ya que en realidad no se dispone de dos muestras sino de una. Asimismo, las interpretaciones y procedimientos son similares para los dos tipos de test pareados mencionados en el esquema, por lo que solo se mostrará un ejemplo detallado, siendo los demás análogos.

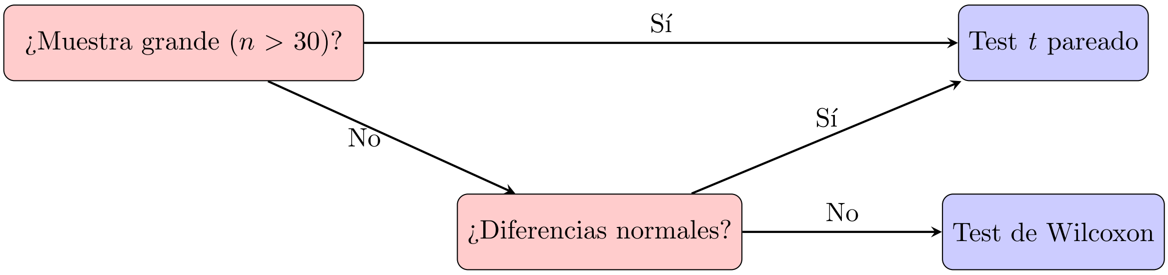


Figura 4.29: Esquema de selección del test estadístico adecuado en la comparación de una variable cuantitativa en dos muestras pareadas.

Ejemplo: Pulso antes y después de un minuto de reposo

En este ejemplo, queremos analizar si el pulso en reposo de las personas que no corren en el experimento varía significativamente tras un minuto de inactividad. Es decir, si la diferencia entre las mediciones del pulso antes (`Pulso1`) y después (`Pulso2`) es significativa.

Las hipótesis del test t pareado para evaluar si el pulso en reposo cambia tras un minuto de inactividad son:

- **Hipótesis nula (H_0):** No hay diferencia en la media del `Pulso` antes y después del minuto de reposo, es decir, las medias son iguales:

$$H_0 : \mu_{\text{Pulso1}=\text{Pulso2}}.$$

- **Hipótesis alternativa (H_1):** Existe una diferencia significativa entre el `Pulso` antes y después del minuto de reposo, es decir, las medias son distintas:

$$H_1 : \mu_{\text{Pulso1}\neq\text{Pulso2}}.$$

El primer paso es observar el tamaño de la muestra y los valores descriptivos de las mediciones. La Tabla 4.28 señala que el tamaño de la muestra es mayor a 30, por lo que el test t pareado puede aplicarse directamente sin necesidad de comprobar la normalidad.

Variable	Pulso1	Pulso2
N	63	63
Perdidos	0	0
Media	75.857	74.857

Tabla 4.28: Estadísticos descriptivos de la variable Pulso antes y después del reposo.

No obstante, como ilustración, se ha comprobado la normalidad de la diferencia en el menú de test t pareados dentro del apartado `Comprobaciones de Supuestos`. La Tabla 4.29 muestra los resultados de la prueba de Shapiro-Wilk para la variable diferencia `Pulso2-Pulso1` indicando que no hay evidencia para rechazar la normalidad.

Variable	W	p
Pulso2 - Pulso1	0.983	0.538

Tabla 4.29: Prueba de normalidad de Shapiro-Wilk para la variable `Pulso2-Pulso1`.

Para llevar a cabo el test t pareado en *jamovi*, se deben seguir estos pasos:

1. Seleccionar `Análisis` → `T-tests` → `Pureba t para Muestras Pareadas`, como se muestra en la Figura 4.30.

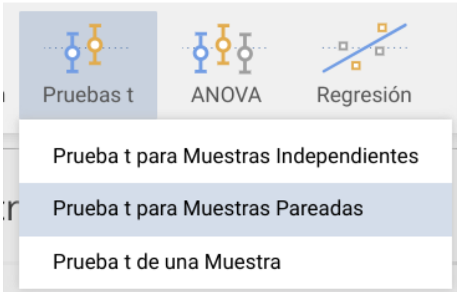


Figura 4.30: Selección del test t pareado en *jamovi*.

2. Arrastrar las variables `Pulso1` y `Pulso2` al cuadro `Variables Apareadas`, como se observa en la Figura 4.31.

Figura 4.31: Asignación de variables en *jamovi* para el test t pareado.

3. En el menú de configuración, marcar la opción `Intervalo de confianza del 95%` para obtener una estimación del efecto sobre las diferencias de las medias.

Los resultados del test t pareado se presentan en la Tabla 4.30. Se observa que el p-valor obtenido es $p = 0{,}047$, lo que indica que hay evidencia suficiente para rechazar la hipótesis nula al nivel de significación del 5 %. Esto sugiere que, en promedio, el pulso tras un minuto de reposo es significativamente diferente al pulso inicial.

Comparación	T	gl	p	Diferencia de Medias	IC 95 %
Pulso1 - Pulso2	2.023	62.000	0.047	1.000	(0.012, 1.988)

Tabla 4.30: Resultados del test t pareado para `Pulso1` y `Pulso2`.

Inevitablemente, al haber encontrado una diferencia significativa, el intervalo de confianza del 95 % para la diferencia de medias $(0{,}012, 1{,}988)$ no incluye el valor cero, confirmando la existencia de un efecto estadísticamente significativo, aunque pequeño, del reposo en la reducción del pulso.

Reporte de resultados en un artículo científico

Se realizó un test t pareado para evaluar si el pulso en reposo difería significativamente tras un minuto de inactividad. Los resultados del análisis mostraron una diferencia estadísticamente significativa entre ambas mediciones ($p = 0{,}047$, diferencia de medias = 1,000, IC 95 % [0.012, 1.988]).

Nota 4.8. Al igual que en secciones anteriores, el test no paramétrico no compara las medias sino las medianas.

4.9.2. Tres o más muestras relacionadas

En el caso de comparar tres o más medidas relacionadas en un mismo grupo de sujetos, las opciones disponibles en *jamovi* son el **ANOVA de medidas repetidas** para datos paramétricos y el **test de Friedman** como alternativa no paramétrica.

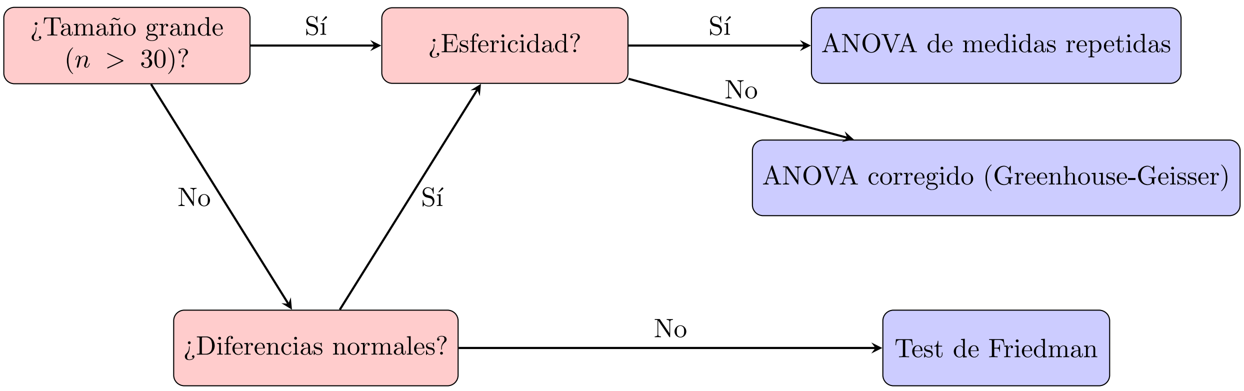


Figura 4.32: Esquema de selección del test estadístico adecuado en un diseño de medidas repetidas.

ANOVA de Medidas Repetidas

Si los datos cumplen los supuestos de normalidad o de un tamaño de muestra no pequeño ($n > 30$), se recomienda el uso del ANOVA de medidas repetidas.

Las hipótesis de este test, para k medidas sobre los mismos individuos, son:

- **Hipótesis nula (H_0):** No hay diferencias significativas en la media de la variable entre las mediciones repetidas, es decir, todas las medias poblacionales son iguales:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k.$$

- **Hipótesis alternativa (H_1):** Al menos una de las medias difiere de las demás, lo que sugiere cambios significativos en la variable analizada:

$$H_1 : \exists\, i, j \quad \text{tal que} \quad \mu_i \neq \mu_j.$$

En este ejemplo, compararemos dos medidas de la variable pulso(`Pulso1` y `Pulso2`) para las personas que no corrieron y así ilustrar el procedimiento en *jamovi*.

Nota 4.9. Aunque los test de esta sección están diseñados para analizar tres o más medidas relacionadas, en este caso ilustramos el procedimiento con solo dos medidas. De este modo, es más fácil ver el paralelismo con los resultados de la Sección 4.9.

A continuación, se detallan los pasos para realizar el ANOVA de medidas repetidas en *jamovi*.

Nota 4.10. Idealmente, para muestras pequeñas, habría que comprobar que las variables diferencia, en este caso solo `Pulso1 - Pulso2`, puedan considerarse normales. No obstante, suele ser práctica habitual comprobar solo la normalidad de los residuos del modelo. En este caso, se puede hacer a partir del resultado del QQ-plot en el tercer paso de las siguientes instrucciones (ver Figura 4.35).

1. Seleccionar `Análisis` → `ANOVA` → `ANOVA de Medidas Repetidas`, como se muestra en la Figura 4.33.

Figura 4.33: Selección del ANOVA de Medidas Repetidas.

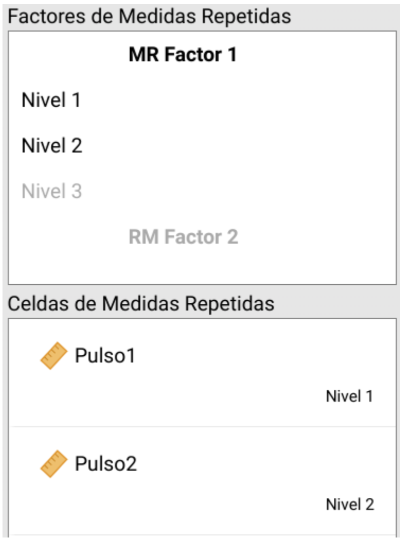


Figura 4.34: Definición del factor de medidas repetidas.

2. En la ventana de configuración, agregar un nombre descriptivo al factor de medidas repetidas si se desea (en el ejemplo dejamos el nombre por defecto `MR Factor 1`) y definir los niveles correspondientes (En este caso dejamos dos niveles al haber solo medida antes y después. En caso de 3 o más niveles, activar el `Nivel 3` en el cuadro `Factores de Medidas Repetidas` y arrastra la tercera variable al cuadro `Celdas de Medidas Repetidas`). Ver Figura 4.34.

3. En la sección de opciones, activar la `Prueba de esfericidad` para verificar si se cumple este supuesto (Figura 4.35). Marcar la opción de `Greenhouse-Geisser` si se desea aplicar el test con corrección en caso de violar la esfericidad (para ello hay que ver el resultado de la prueba de esfericidad).

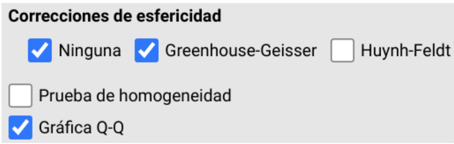


Figura 4.35: Prueba de esfericidad, correcciones del test y gráfica QQ.

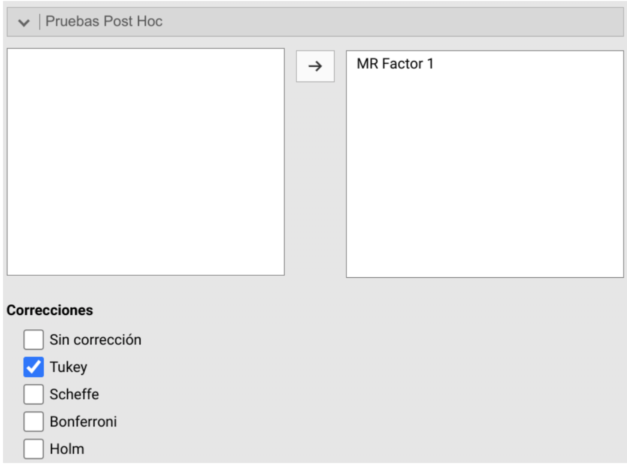


Figura 4.36: Configuración de pruebas post-hoc en *jamovi*.

4. En la sección de `Pruebas Post-Hoc`, seleccionar `Tukey` para realizar comparaciones entre niveles si se encuentran diferencias significativas en el ANOVA. (Figura 4.36).

Resultados del ANOVA de Medidas Repetidas

En primer lugar, dado que tenemos un tamaño muestral alto, solo quedaría comprobar la esfericidad. En este caso se obtiene un valor faltante (NaN), ya que el test no es aplicable para dos medidas. Con tres o más medidas debe arrojar un p-valor. Si este es menor que el nivel de significación se rechazaría la hipótesis de esfericidad y se debería hacer el test ANOVA de medidas repetidas con la correción de Greenhouse-Geisser.

En este caso, se realiza el test ANOVA de medidas repetidas estándar, cuyos resultados del análisis se presentan en la Tabla 4.31, observando un efecto significativo con un valor de $p = 0{,}047$.

Factor	Suma de Cuadrados	gl	F	p
MR Factor 1	31.500	1	4.094	0.047
Residual	477.000	62	7.694	-

Tabla 4.31: Resultados del ANOVA de medidas repetidas.

Nota 4.11. Cuando se analizan únicamente dos medidas repetidas, el ANOVA de medidas repetidas arroja el mismo resultado que el test t pareado. La diferencia radica en su capacidad para analizar tres o más mediciones sin necesidad de realizar múltiples comparaciones individuales.

La Figura 4.37 muestra la representación gráfica de las medias de cada medida.

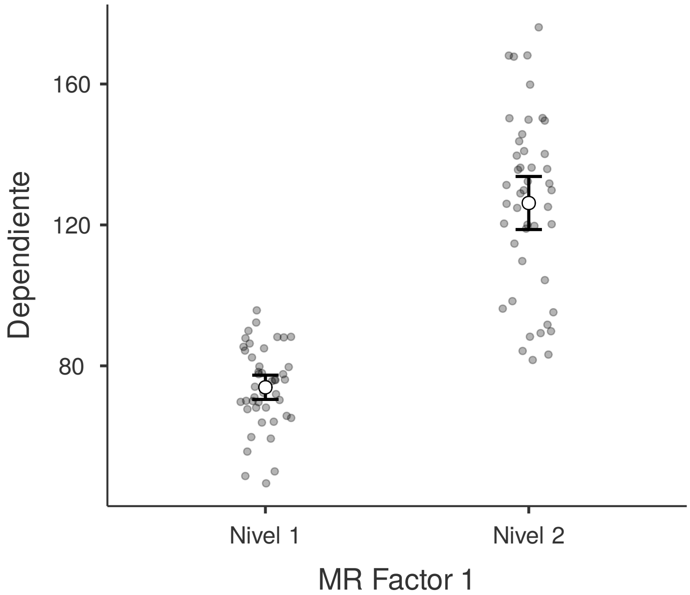


Figura 4.37: Gráfico de medias para el ANOVA de medidas repetidas.

Comparaciones Post-Hoc

En este caso, no cabría realizar un test post-hoc, puesto que solo disponemos de dos grupos. Solo por efectos ilustrativos, estos resultados se presentan en la Tabla 4.32. Se observa una diferencia significativa entre el **Nivel 1** y el **Nivel 2** ($p < 0{,}047$).

Comparación	**Diferencia de Medias**	**EE**	**gl**	**t**	**p**
Nivel 1 - Nivel 2	1.000	0.494	62.000	2.023	0.047

Tabla 4.32: Resultados de las comparaciones post-hoc (Tukey).

La Tabla 4.33 presenta las medias de cada medida estimadas junto con sus intervalos de confianza al 95 %.

Nivel	**Media**	**EE**	**Inferior**	**Superior**
Nivel 1	75.857	1.464	72.930	78.784
Nivel 2	74.857	1.368	72.123	77.591

Tabla 4.33: Medias marginales estimadas con intervalos de confianza al 95 %.

Reporte de resultados en un artículo científico

Antes del minuto de inactividad, la frecuencia cardíaca media era de 75.857 ppm (IC 95 %: [72.930, 78.784]) mientras que después del minuto de reposo la media fue de 74.857 ppm (IC 95 %: [72.123, 77.591]).
El ANOVA de medidas repetidas mostró una diferencia significativa entre ambas mediciones ($p = 0{,}047$); sin embargo, la magnitud del cambio es pequeña, sugiriendo que, aunque la frecuencia cardíaca se redujo ligeramente, el efecto del minuto de inactividad sobre el pulso es limitado.

Nota: Si se realiza un post-hoc entre tres o más grupos, habrá que describir también qué pares de grupos muestran diferencias significativas.

Test de Friedman

Cuando los datos no cumplen con los supuestos de normalidad y se tiene un tamaño de muestra reducido, se debe emplear el test de Friedman. Nótese que dicho test es una prueba no paramétrica que evalúa si hay diferencias significativas en la distribución de las mediciones en diferentes momentos.

- **Hipótesis nula (H_0):** No hay diferencias en la distribución de la variable entre las mediciones repetidas, es decir, las medianas son iguales en todos los momentos:

$$H_0 : \text{Mediana}_1 = \text{Mediana}_2 = \cdots = \text{Mediana}_k.$$

- **Hipótesis alternativa (H_1):** Al menos una de las mediciones tiene una mediana diferente a las demás, lo que indica cambios en la variable a lo largo del tiempo:

$$H_1 : \exists\, i, j \quad \text{tal que} \quad \text{Mediana}_i \neq \text{Mediana}_j.$$

Para aplicarlo en *jamovi*, se deben seguir los siguientes pasos:

1. Seleccionar `Análisis` → `ANOVA` → `ANOVA de Medidas Repetidas - Friedman`, como se muestra en la Figura 4.38.

2. Arrastrar las variables dependientes al cuadro de celdas de medidas repetidas, en este caso `Pulso1` y `Pulso2`.

3. Marcar la opción `Comparaciones entre parejas (Durbin-Conover)`, como se muestra en la Figura 4.39, para obtener un análisis post-hoc en caso de que el test sea significativo.

Figura 4.38: Selección del test de Friedman en *jamovi*.

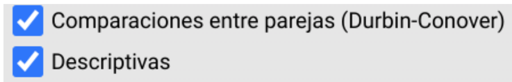


Figura 4.39: Activación del test post-hoc Durbin-Conover en *jamovi*.

Resultados del Test de Friedman

Los resultados del test de Friedman se presentan en la Tabla 4.34. Se observa un valor de $p = 0{,}225$, lo que indica que no hay evidencia suficiente para rechazar la hipótesis nula de que las distribuciones son iguales en las distintas mediciones.

Test	χ^2	**gl**	p
Friedman	1.473	1	0.225

Tabla 4.34: Resultados del test de Friedman.

Nota: en este caso el test no es significativo. Probablemente esto sea debido a la menor potencia del test paramétrico. Como en este caso se cumplían los supuesto para usar el test ANOVA de medidas repetidas, es ese el que se debería haber ejecutado.

Comparaciones Post-Hoc

Dado que el test de Friedman no mostró diferencias significativas entre los niveles ($p = 0{,}225$), el análisis post-hoc de Durbin-Conover se presenta solo con fines ilustrativos en la Tabla 4.35. Se observa que la comparación entre `Pulso1` y `Pulso2` arroja un valor de $p = 0{,}228$, lo que confirma que no hay diferencias significativas entre los niveles.

Comparación	**Estadístico**	p
Pulso1 - Pulso2	1.218	0.228

Tabla 4.35: Resultados de las comparaciones post-hoc (Durbin-Conover).

Interpretación de los resultados

El test de Friedman no reveló diferencias significativas en la distribución de la frecuencia cardíaca entre las dos mediciones ($p = 0{,}225$). La mediana del pulso antes del periodo de inactividad fue de 76.000 ppm, mientras que después del periodo de inactividad la mediana fue de 73.000 ppm.

A pesar de una ligera reducción en la mediana del pulso tras el minuto de inactividad, la ausencia de significación estadística sugiere que este cambio no es sistemático ni consistente en la muestra.

4.10. Ejercicios

1. Se desea analizar las variables `Peso` y `Pulso1` en el archivo `pulsoProcesado.omv`.

   a) Contrasta si la media de `Peso` es 65 kg y si la media de `Pulso1` es 75 pulsaciones por minuto.

   b) Evalúa si se cumplen las condiciones necesarias para aplicar estos contrastes en estas variables.

   c) Repite los contrastes anteriores separando por `sexo`.

2. Un investigador considera que el peso medio de los bebés recién nacidos ha cambiado respecto a hace treinta años, cuando era de 3.3 kg. En el archivo `bebes.sav` se dispone del peso de una muestra aleatoria de 50 recién nacidos del último año. Utiliza estos datos para contrastar la hipótesis del investigador.

3. El archivo `bilis.sav` contiene el pH de la bilis hepática de un conjunto de personas. Se quiere evaluar si, en base a esta muestra, el pH puede considerarse neutro ($pH = 7$). Realiza el contraste adecuado.

4. Se desea comprobar si las variables `Peso`, `Altura` y `Pulso1` en el archivo `pulsoProcesado.omv` tienen la misma media entre quienes han corrido y quienes no.

   a) Realiza los contrastes de igualdad de medias..

   b) Evalúa si se cumplen las condiciones necesarias para aplicar estos contrastes. Si no se cumplen, aplica una alternativa no paramétrica.

5. Un fabricante quiere evaluar la calidad de dos materiales para suelas de zapatos. Para ello, realiza el siguiente experimento: selecciona 30 niños y les coloca una suela con el material A en un zapato y una suela con el material B en el otro. Para evitar sesgos por lateralidad (diestros/zurdos), en la mitad de los casos el material A se coloca en la suela izquierda y en la otra mitad, en la derecha, asignándose al azar. Los datos están en el archivo `suelas.sav`, con las variables:

- `Niño`: Identificador de cada participante.
- `A`: Desgaste en milímetros del material A.
- `B`: Desgaste en milímetros del material B.
- `izquierda`: Indica si el material A estuvo en la suela izquierda (`1`) o en la derecha (`0`).

a) Contrasta si ambos materiales presentan el mismo desgaste medio.

b) Evalúa si el desgaste de los materiales A y B depende de si estuvieron en la suela izquierda o derecha.

6. El fabricante de suelas sospecha que la edad de los niños influye en el desgaste, por lo que realiza otro experimento. Se seleccionan 20 niños de 5 años, 20 de 6 y 20 de 7 años. Como en el experimento anterior, cada niño usa una suela con el material A y otra con el material B, asignando aleatoriamente si el material A va en la izquierda o en la derecha. Los datos están en el archivo `suelas2.sav`, con las siguientes variables:

- `A`: Desgaste en milímetros del material A.
- `B`: Desgaste en milímetros del material B.
- `izquierda`: Indica si el material A estuvo en la suela izquierda (`1`) o en la derecha (`0`).
- `edad`: Grupo de edad (1 = 5 años, 2 = 6 años, 3 = 7 años).

a) Contrasta si el desgaste de los materiales A y B depende de si estuvieron en la suela izquierda o derecha.

b) Evalúa si hay diferencias en los desgastes medios de los materiales A y B.

c) Examina si el desgaste medio es el mismo en cada grupo de edad para el material A y para el material B.

d) Si se calcula la diferencia de desgaste entre el material B y el material A para cada niño, ¿se puede rechazar la hipótesis de que la media de esta diferencia es la misma en los tres grupos de edad? ¿Se puede rechazar la hipótesis de igualdad de medias entre los niños de 5 y 6 años? ¿Cómo interpretarías estos resultados?

'Todos los modelos son erróneos, pero algunos son útiles."

George E.P. Box (Matemático)

5

Modelización lineal

5.1. El modelo lineal múltiple

La regresión lineal múltiple es una de las herramientas más populares en la modelización estadística debido a su capacidad para integrar simultáneamente múltiples variables predictoras para describir una variable respuesta.

Aunque pueda parecer una hipótesis muy fuerte, la linealidad es razonable en muchas situaciones, ya que en numerosos fenómenos naturales y sociales las relaciones entre variables pueden aproximarse adecuadamente mediante combinaciones lineales, tanto por motivos teóricos como por mera observación empírica. Además, aunque la relación exacta entre las variables rara vez es estrictamente lineal, esta suposición sigue siendo útil, pues permite capturar tendencias generales, simplificar la interpretación y proporcionar modelos fácilmente estimables y aplicables a la predicción y la inferencia.

El modelo de regresión lineal múltiple se expresa matemáticamente como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon.$$

Donde:

- Y es la variable dependiente o respuesta.
- $X_1, X_2, \ldots, X_k$ son las variables independientes o predictoras.
- β_0 es la ordenada al origen (intercepto).

- $\beta_1, \beta_2, \ldots, \beta_k$ son los coeficientes de regresión asociados a las variables independientes.
- ε es el término de error, que se asume normalmente distribuido con media cero y varianza σ^2.

5.2. Estimación del Modelo

Vamos a ilustrar el ajuste de un modelo de regresión lineal múltiple para predecir, en el fichero `pulsoProcesado.omv` para las personas del experimento que han corrido, cómo es el pulso después de correr (`Pulso2`) en función del pulso antes de correr (`Pulso1`). Además, vamos a incluir en el modelo otras variables antropométricas, como la altura y el peso, y variables sociodemográficas, como la edad, el sexo y los hábitos negativos[1] de la persona.

Ajustar o corregir por variables adicionales

En modelos de regresión, incluir variables adicionales en el análisis se conoce como *ajustar* o *corregir por* esas variables. Este proceso permite estimar el efecto de una variable específica (por ejemplo, `Pulso1`) sobre la variable dependiente (`Pulso2`), *controlando* la influencia de otras variables que podrían afectar la relación observada.

En este caso, al incluir variables como la edad, el peso, la altura, el sexo y los hábitos negativos, se obtiene una estimación más precisa del impacto del pulso inicial en el pulso final, eliminando, o al menos reduciendo, posibles *confusiones* o *sesgos* que podrían surgir si estas características influyeran en ambas variables.

Para estimar los parámetros del modelo de regresión lineal múltiple, lo que se llama técnicamente ajustar el modelo, se realiza una estimación máxima-verosimilitud. En *jamovi*, se deben seguir los siguientes pasos:

1. Seleccionar **Análisis** → **Regresión** → **Regresión Lineal**, como se muestra en la Figura 5.1.

Figura 5.1: Selección de regresión lineal en *jamovi*.

[1] Se recuerda que la variable `Hábitos negativos` toma el valor 2 si la persona fuma y bebe alcohol con regularidad, 1 si realiza solo una de las dos cosas, y 0 si no realiza ninguna.

2. Asignar la variable dependiente (`Pulso2`) al cuadro correspondiente y las covariables cuantitativas (`Pulso1`, `Edad`, `Altura`, `Peso`, `Hábitos negativos`) al cuadro de variables predictoras. Por otro lado, la variable cualitativa `Sexo` se debe colocar en el cuadro de `Factores`, ya que será tratada como una variable categórica en el modelo (Figura 5.2).

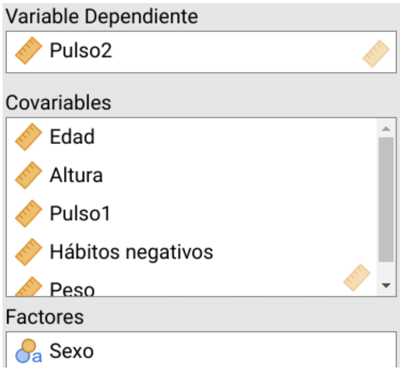


Figura 5.2: Especificación de variables en el modelo de regresión.

3. Establecer un nivel de referencia para la variable `Sexo`, seleccionando *Mujer* como categoría base (Figura 5.3). Esto es fundamental, ya que en un modelo de regresión con variables categóricas, los coeficientes se interpretan en relación con el nivel de referencia. En este caso, el coeficiente asociado a `Sexo` representará la diferencia en `Pulso2` entre los hombres y las mujeres, tomando a las mujeres como el grupo de comparación.

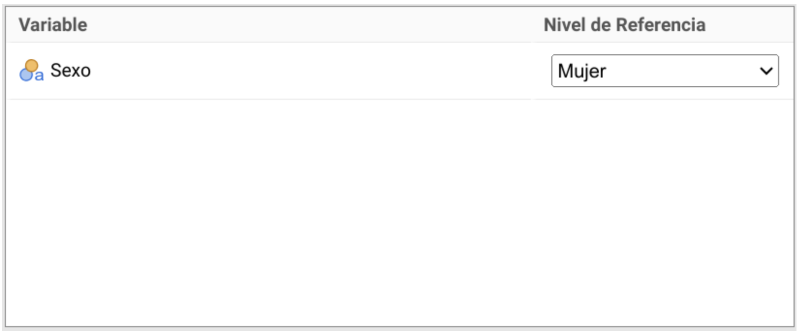


Figura 5.3: Selección del nivel de referencia para la variable `Sexo`.

4. En la sección de `Comprobaciones de Supuestos`, activar las opciones `Estadísticas de colinealidad`, `Prueba de normalidad`, `Gráfica Q-Q de residuos` y `Gráficas de residuos` para evaluar los supuestos del modelo (Figura 5.4).

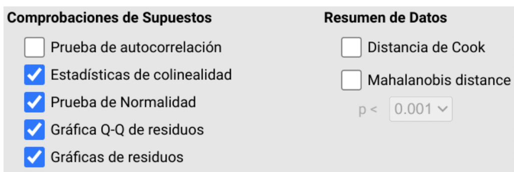


Figura 5.4: Selección de pruebas de diagnóstico para los supuestos del modelo.

5. En la sección de `Medidas de ajuste`, activar las opciones `R²` y `R² corregida` para evaluar la calidad del modelo y seleccionar la `Prueba F` para determinar la significación global del ajuste del modelo (Figura 5.5).

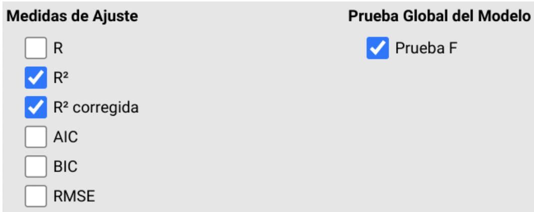


Figura 5.5: Selección de medidas de ajuste y prueba de significación global.

Resultados del Modelo

Los coeficientes estimados del modelo se presentan en la Tabla 5.1. Se observa que la variable `Pulso1` tiene un coeficiente positivo y significativo ($b = 1{,}510, p < 0{,}001$), lo que indica que un aumento en el pulso inicial se asocia con un incremento en `Pulso2`. Además, la variable `Altura` también muestra una asociación significativa ($b = 1{,}166, p = 0{,}035$). Sin embargo, el resto de predictores no presentan efectos estadísticamente significativos.

Predictor	**Estimador**	**EE**	**Inferior**	**Superior**	**t**	**p**
Constante	-186.227	102.922	-394.768	22.314	-1.809	0.079
Pulso1	1.510	0.282	0.938	2.081	5.354	< 0.001
Sexo (Hombre - Mujer)	-4.552	8.677	-22.134	13.030	-0.525	0.603
Edad	0.993	1.323	-1.686	3.673	0.751	0.457
Peso	-0.191	0.332	-0.864	0.482	-0.574	0.569
Altura	1.166	0.534	0.085	2.247	2.185	0.035
Hábitos negativos	-7.808	5.645	-19.247	3.630	-1.383	0.175

Tabla 5.1: Coeficientes estimados del modelo de regresión.

Interpretación de los Coeficientes

Cada coeficiente $\hat{\beta}_j$ en la ecuación de regresión representa el cambio medio en Y cuando X_j aumenta en una unidad, manteniendo constantes las demás variables.

Ejemplo: Un coeficiente estimado de $\hat{\beta}_1 = 1{,}510$ para `Pulso1` indica que, en promedio, un aumento de una unidad en `Pulso1` se asocia con un incremento de 1.510 latidos por minuto en `Pulso2`, manteniendo constantes el resto de las variables.

5.2.1. Bondad de Ajuste del Modelo

La calidad del modelo se mide con el coeficiente de determinación R^2 y su versión ajustada:

- **Coeficiente de Determinación R^2:** Representa la proporción de variabilidad de Y explicada por el modelo:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}.$$

Un $R^2 = 0{,}53$ significa que el 53 % de la variabilidad de Y es explicada por el modelo.

- **Coeficiente de Determinación Ajustado $R^2_{ajustado}$:** Penaliza la inclusión de predictores irrelevantes:

$$R^2_{ajustado} = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

En cuanto a la calidad del ajuste del modelo para el pulso (ver Tabla 5.2), el coeficiente de determinación (R^2) indica que el modelo explica aproximadamente el 46,8 % de la variabilidad en `Pulso2`, mientras que el R^2 corregido, que ajusta por el número de predictores, es de 38,1 %.

Medida	**Valor**
R^2	0.468
R^2 corregido	0.381

Tabla 5.2: Medidas de ajuste del modelo.

5.3. Pruebas de Significación

Para evaluar la validez del modelo, se realizan pruebas de hipótesis tanto a nivel global como individual.

Test de Significación Global (Test F) El test F evalúa si el conjunto de predictores explica una proporción significativa de la variabilidad en la variable respuesta. Se plantea el siguiente contraste:

Hipótesis del Test F:

- H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (ninguna variable independiente explica Y).
- H_1: Al menos un $\beta_j \neq 0$ (al menos una variable independiente contribuye significativamente al modelo).

El estadístico de prueba se calcula como:

$$F = \frac{\frac{\sum(\hat{Y}_i - \bar{Y})^2}{k}}{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-k-1}} \sim F_{k,n-k-1}.$$

Si el p-valor asociado a F es menor que 0.05, se rechaza H_0, concluyendo que el modelo tiene poder predictivo.

Los resultados de la prueba global del modelo se presentan en la Tabla 5.3. Se observa que el modelo tiene una significación estadística alta ($p < 0{,}001$), lo que indica que al menos una de las variables predictoras contribuye a explicar la variabilidad en `Pulso2`.

Modelo	R^2	R^2 corregido	F	**gl1**	**gl2**	**p**
1	0.468	0.381	5.416	6	37	< 0.001

Tabla 5.3: Prueba global del modelo.

Test de Significación Individual (Test t) Para cada coeficiente β_j, se evalúa si su contribución al modelo es significativa:

Hipótesis del Test t:

- H_0: $\beta_j = 0$ (la variable X_j no contribuye significativamente al modelo).
- H_1: $\beta_j \neq 0$ (la variable X_j tiene un efecto significativo en Y).

El estadístico de prueba se obtiene como:

$$t = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim t_{n-k-1}.$$

Si el p-valor es menor a 0.05, se rechaza H_0 y se concluye que la variable X_j tiene un efecto significativo sobre Y.

Los resultados de la Tabla 5.1 muestran que las únicas variables significativas son `Pulso1` y `Altura`, con $p < 0{,}001$ y $p = 0{,}035$, respectivamente. Esto sugiere que un mayor pulso inicial y una mayor altura están asociados con valores más altos en `Pulso2`, manteniendo constantes el resto de las variables.

5.4. Evaluación de Supuestos

Para validar la adecuación del modelo de regresión, es fundamental evaluar distintos supuestos, como la *colinealidad entre predictores*, la *normalidad de los residuos* y la *homocedasticidad*. A continuación, se presentan los diagnósticos realizados.

5.4.1. Colinealidad entre Predictores

Para evaluar la presencia de colinealidad entre los predictores, se presentan los factores de inflación de la varianza (VIF) en la Tabla 5.4. Se observa que la variable `Altura` tiene un VIF relativamente alto ($VIF = 3{,}867$), lo que indica cierta colinealidad, aunque no a niveles críticos. Suele considerarse problemáticos valores superiores a 5 e inadmisibles valores superiores a 10. En estos casos, es recomendable realizar un análisis de correlaciones para eliminar variables del modelo, ya que un VIF alto sugiere que varias variables aportan información redundante.

Variable	**VIF**
Pulso1	1.118
Sexo	2.090
Edad	1.556
Peso	2.601
Altura	3.867
Hábitos negativos	1.146

Tabla 5.4: Estadísticas de colinealidad del modelo.

5.4.2. Normalidad de los Residuos

Para verificar la normalidad de los residuos, se realiza la prueba de Shapiro-Wilk, cuyos resultados se presentan en la Tabla 5.5. Con un $p = 0{,}716$, no se rechaza la hipótesis de normalidad, lo que sugiere que los residuos del modelo siguen una distribución aproximadamente normal.

Estadístico	**p**
0.982	0.716

Tabla 5.5: Prueba de normalidad de Shapiro-Wilk aplicada a los residuos del modelo.

Sin embargo, en muestras grandes, la prueba de Shapiro-Wilk tiende a rechazar la hipótesis de normalidad incluso cuando la distribución de los residuos es visualmente muy similar a una normal. Por ello, es recomendable complementar este análisis con una inspección del gráfico Q-Q de los residuos para evaluar si las desviaciones de la normalidad son significativas.

En la Figura 5.6, se observa que la mayoría de los puntos se alinean con la diagonal teórica, lo que indica que los residuos siguen aproximadamente una distribución normal. Pequeñas desviaciones en los extremos pueden sugerir la presencia de valores atípicos o ligeras asimetrías.

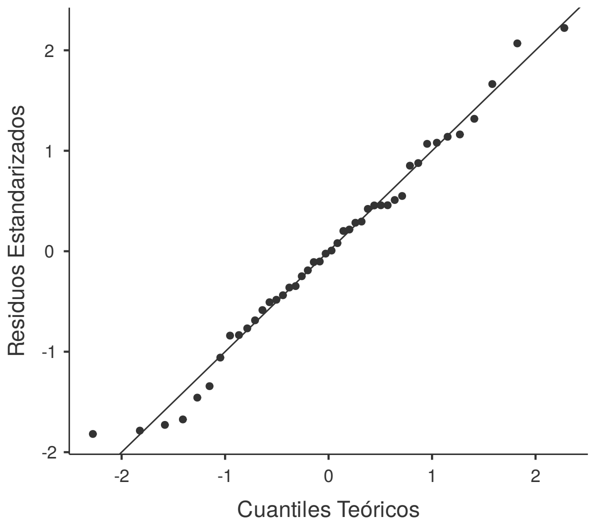


Figura 5.6: Gráfico Q-Q de los residuos estandarizados del modelo de regresión.

5.4.3. Homocedasticidad y Distribución de los Residuos

Otro supuesto fundamental es la *homocedasticidad*, es decir, que la varianza de los residuos se mantenga constante a lo largo de los valores predichos. Para verificar este supuesto, se analiza el gráfico de residuos frente a valores ajustados (Figura 5.7).

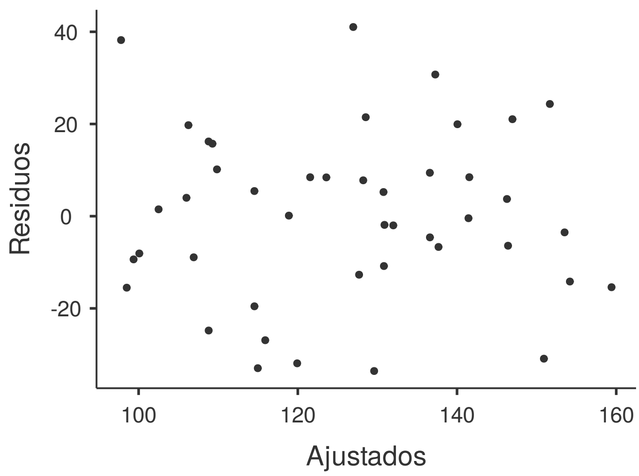


Figura 5.7: Gráfico de residuos vs valores ajustados del modelo.

En este gráfico, se observa que los residuos están distribuidos aleatoriamente en torno al eje horizontal sin mostrar un patrón sistemático. Esto indica que la homocedasticidad se cumple, es decir, la varianza de los residuos permanece constante a lo largo de los valores ajustados, lo que valida la robustez del modelo de regresión.

5.5. Selección de Variables

En la modelización de regresión, la selección de variables es un paso fundamental para encontrar un equilibrio entre un modelo explicativo y un modelo parsimonioso. Un modelo con demasiadas variables puede llevar a sobreajuste, mientras que uno con pocas variables podría no capturar adecuadamente la relación entre los predictores y la variable respuesta.

Uno de los métodos más utilizados para la selección de variables es la *selección paso a paso* (*stepwise selection*), que consiste en incluir o excluir predictores basándose en criterios estadísticos, como el criterio de información de Akaike (AIC) o el coeficiente de determinación ajustado (R^2 ajustado). Sin embargo, en *jamovi*, este procedimiento no está implementado de forma automática.

A pesar de esta limitación, podemos realizar la selección de variables de forma manual evaluando los cambios en el R^2 ajustado al incluir o excluir predictores en el modelo.

Evaluación de la inclusión de una variable

Para ilustrar este procedimiento, supongamos que incorporamos la variable `Fumador` en el modelo. Al hacerlo, el R^2 ajustado disminuye, lo que indica que la inclusión de esta variable no mejora el ajuste global del modelo. En la Tabla 5.6, se observa que el R^2 ajustado es de 0.377, en comparación con el R^2 ajustado de 0.381 del modelo original (Tabla 5.2).

Modelo	R^2	R^2 corregido
1	0.478	0.377

Tabla 5.6: Medidas de ajuste del modelo tras incluir la variable `Fumador`.

Dado que la incorporación de `Fumador` no contribuye a mejorar la capacidad explicativa del modelo y, de hecho, reduce el R^2 ajustado, se concluye que esta variable no aporta información relevante y debería ser excluida del modelo salvo que se tenga un motivo teórico de fuerza mayor para mantenerla.

Evaluación de la eliminación de una variable

De manera similar, podemos analizar el impacto de retirar del modelo la variable `Hábitos negativos`. Si eliminamos esta variable, el R^2 ajustado disminuye, pasando a un valor de 0.366, como se muestra en la Tabla 5.7. En comparación con el modelo original (Tabla 5.2), donde el R^2 ajustado era de 0.381, esta reducción sugiere que la variable `Hábitos negativos` contribuye al ajuste global del modelo, aunque su coeficiente individual pueda no ser significativo en la prueba t.

Modelo	R^2	R^2 corregido
1	0.440	0.366

Tabla 5.7: Medidas de ajuste del modelo tras eliminar la variable `Hábitos negativos`.

En este caso, aunque la variable `Hábitos negativos` no sea significativa de manera individual según su prueba t, su inclusión en el modelo mejora el ajuste global. Por lo tanto, podría ser conveniente mantenerla, ya que su presencia junto con otras variables contribuye a explicar mejor la variabilidad de `Pulso2`.

Conclusión sobre la Selección de Variables

La selección de variables es un proceso iterativo que debe basarse en múltiples criterios. Aunque el p-valor de la prueba t para cada predictor es una referencia útil, no debe ser el único criterio para decidir qué variables incluir o excluir. El R^2 ajustado proporciona una medida más robusta de la calidad del ajuste del modelo y permite evaluar el impacto de cada predictor en conjunto con los demás.

Bibliografía

[1] Álvarez Tena, B., Anglada Salvanés, S., Asín Lafuente, J., Badía Blasco, F. G., Berrade Ursúa, M. D., Castillo Mateo, J., Castillo Salazar, I. P., Galé Pola, M. C., Gracia Tabuenca, Z., Hernández González, A., Iranzo Sanz, J. A., Isidoro Ramírez, D., Lafuente Blasco, M., Lahoz Arnedo, D., Plo Alastrué, B. F., Sánchez-Valverde García, M. B., Sangüesa Lafuente, C. J., & Tejel Altarriba, F. J. (2025). *Prácticas de estadística básica con software específico y de propósito general* [Curso abierto]. Universidad de Zaragoza. https://ocw.unizar.es/

[2] Ferrándiz, J. A. (s.f.). *Tesis doctoral*. Universidad de Zaragoza.

[3] Lafuente Blasco, M. (2025). *Estadística Aplicada a la Investigación Social*. Prensas de la Universidad de Zaragoza. ISBN: 978-84-1340-977-1.

[4] R Core Team. (2024). *R: A language and environment for statistical computing* (Versión 4.4) [Software]. https://cran.r-project.org

[5] The jamovi project. (2024). *jamovi* (Versión 2.6) [Software]. https://www.jamovi.org

[6] Wilson, R. J. (s.f.). *Pulse Rates before and after Exercise* [Conjunto de datos]. The University of Queensland. http://www.statsci.org/data/oz/ms212.html