METODOLOGÍA DE LA INVESTIGACIÓN EN CIENCIAS DE LA SALUD II

Zaira Torres, Adrián García-Mollá y Sara Martínez-Gregorio,

ARMAQOL

Metodología de la Investigación en Ciencias de la Salud II

ISBN: 978-84-09-70644-0

Metodología de la Investigación en Ciencias de la Salud II

Zaira Torres Romero
Adrián García Mollá
Sara Martínez Gregorio

Índice

TEMA 1

Introducción a Jamovi

Zaira Torres Romero
Adrián García Mollá
Sara Martínez Gregorio

Tema 1. Introducción a Jamovi

1.¿Qué es Jamovi?

Para poder realizar los análisis estadísticos que requiera cada tipo de investigación lo más usual es organizar la información en bases de datos y emplear algún programa estadístico. La obtención de una base de datos supone la codificación previa de las observaciones, la introducción de los datos en archivos informáticos, la depuración (detección y tratamiento de los errores y de los valores faltantes), y, a veces, la transformación y tratamiento de los ficheros para facilitar el análisis estadístico. Programas como SPSS, Jamovi, JASP, MPLUS o R son algunos ejemplos de herramientas que podremos emplear.

En este capítulo introduciremos el programa Jamovi, una Interfaz Gráfica de R creada por Jonathon Love, Damian Dropmann y Ravi Selker, que ofrece los últimos avances en metodología estadística y que tiene múltiples ventajas:

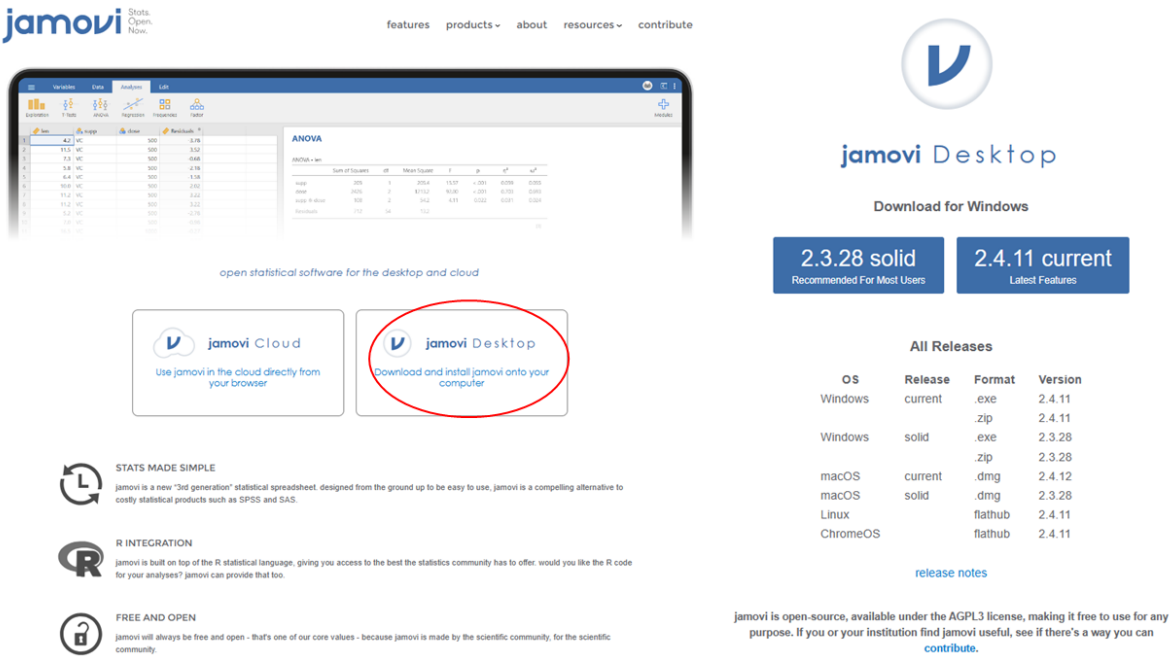
- Es de uso libre
- Permite usar muchas de las funciones de R de una forma sencilla.
- Permite realizar las pruebas y gráficos más comunes
- Es compatible con Windows, Linux y Mac
- Tiene una transferencia de archivos versátil y permite importar y exportar datos de múltiples programas
- Se puede extender y personalizar con múltiples módulos para análisis más específicos
- Ofrece la posibilidad de introducir código de R
- Continuamente se está actualizando

2. Instalación e Interfaz

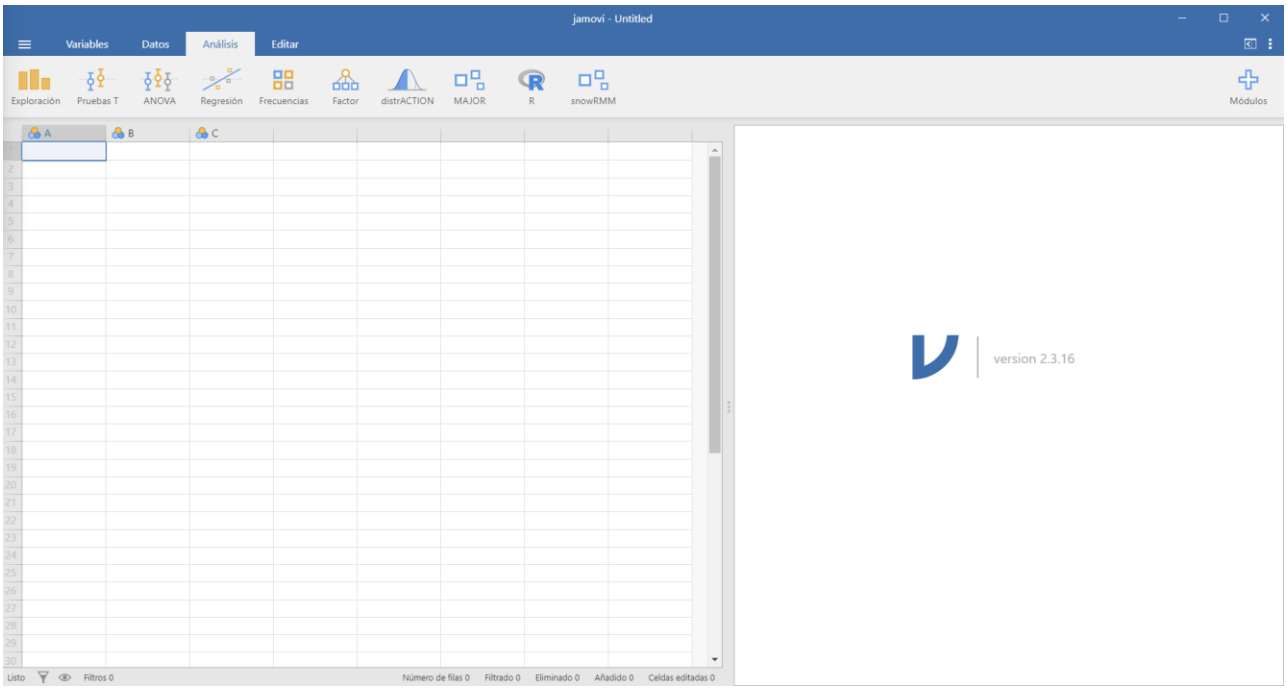
JAMOVI se puede descargar de manera gratuita en:

www.jamovi.org

La opción “Jamovi desktop” te lleva a una nueva pantalla donde debes elegir el instalador según tu sistema operativo y según si quieres la versión “solid”, ya probada y en la que se garantiza un uso estable y fiable de Jamovi o la versión “current”, que contiene nuevas funciones y puede contener errores menores. La versión “solid” es la recomendada para uso general y servirá para todas las funciones y análisis que veremos en este y próximos capítulos.

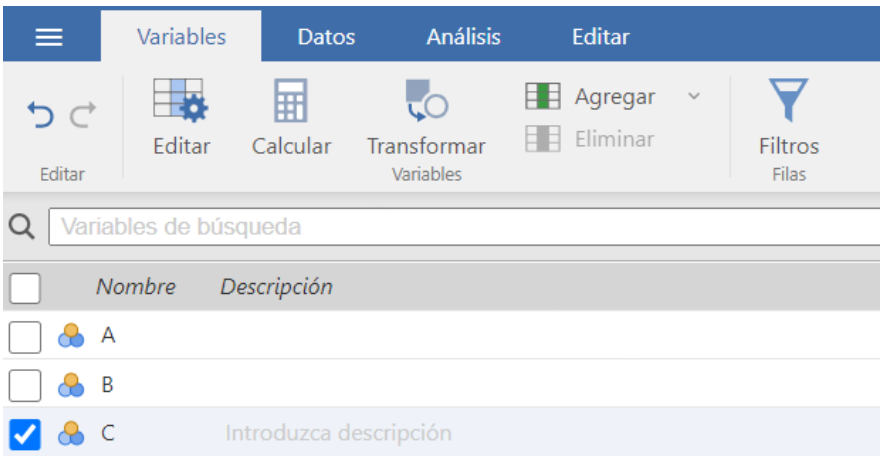


Nada más abrir Jamovi, la interfaz nos muestra dos ventanas y un menú principal con varias opciones.

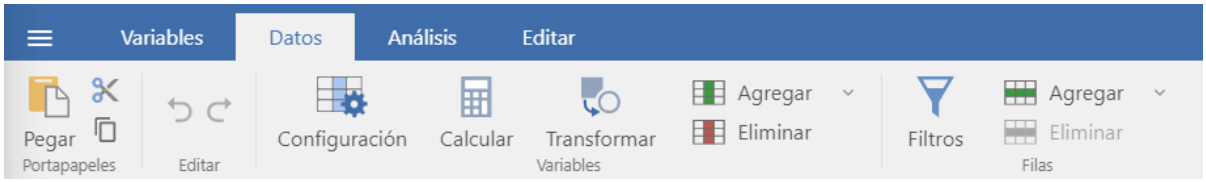


En la cuadrícula de la izquierda se pueden introducir los datos manualmente o importar ficheros con extensión ".csv", ".sav", ".txt" o ".xlsx". En la ventana de la derecha irán apareciendo los resultados generados, que se actualizarán automáticamente si realizamos alguna modificación. Podemos mover el tamaño de las ventanas según nos interese visualizar más grandes los datos o los resultados. El botón ☰ de la barra principal nos dará la opción de abrir o importar una base de datos, de guardar nuestra base y resultados y de exportar la base.

En la opción de "Variables" visualizaremos el nombre y descripción de las variables del conjunto de datos y tendremos las opciones de editarlas, calcular una nueva variable, transformar una variable, agregar una columna para una nueva variable o filtrar:



En el menú de “Datos” podremos realizar las siguientes operaciones:



- **Portapapeles:** Sirve para pegar, cortar o copiar observaciones.
- **Editar:** Sirve para deshacer o rehacer una acción.
- **Variables:** Configuración permite definir las propiedades de las variables, Calcular permite crear nuevas variables, Transformar permite hacer cambios en las variables, agregar y eliminar sirven para añadir o quitar una variable.
- **Filas:** Sirve para aplicar filtros y agregar o eliminar observaciones.

En el menú “Análisis” encontraremos distintas opciones para realizar los análisis estadísticos:

- **Exploración:** para tablas descriptivas y gráficos.
- **Pruebas T:** para comparar 2 muestras independientes y 2 muestras relacionadas, muestra frente a población de manera paramétrica y no paramétrica.
- **ANOVA:** para ANOVA de 1 factor, de varios factores, ANCOVA, MANCOVA, medidas repetidas con opción paramétrica y no paramétrica.
- **Regresión:** Correlaciones, modelos de regresión lineal y logística.
- **Frecuencias:** Tablas de contingencia, pruebas Chi-cuadrado y de McNemar.
- **Factor:** Análisis de fiabilidad de componentes principales y factorial.

Por último, en la derecha se encuentra la opción de 

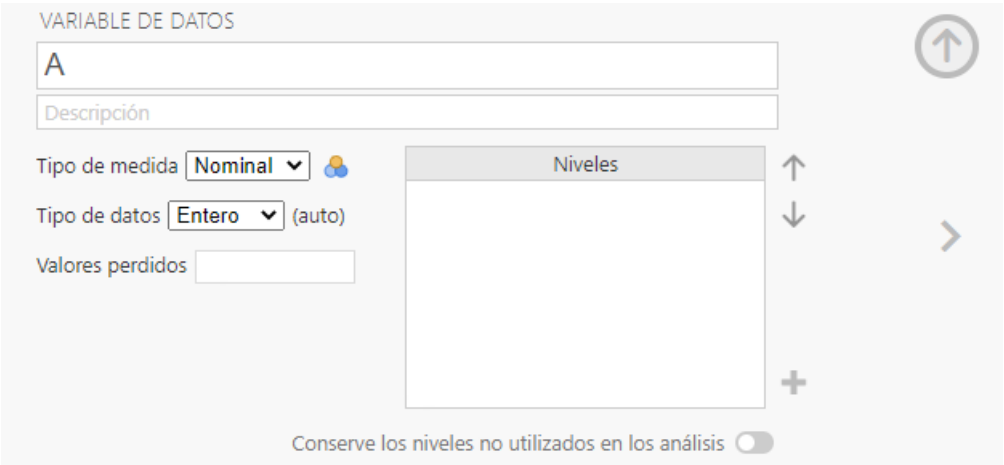
que sirve para incorporar módulos adicionales y poder realizar técnicas estadísticas más avanzadas.

3.Introducir y abrir datos

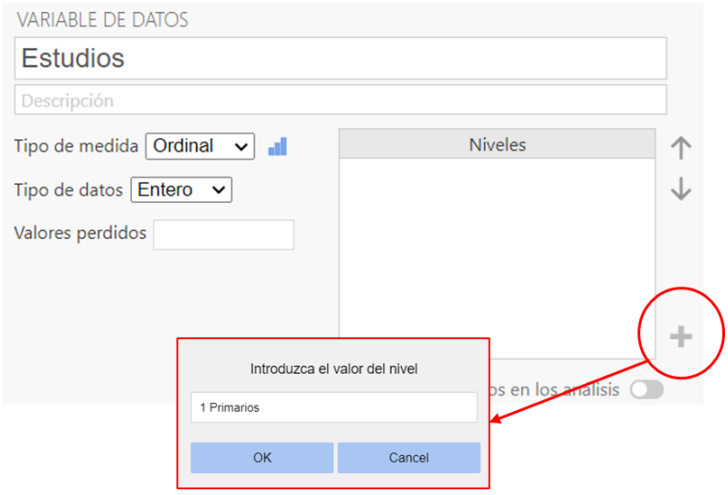
Una opción será introducir los datos manualmente, variable por variable, disponiendo los participantes en filas y las variables en columnas. Si alguna variable tiene decimales debemos tener en cuenta que estos deben separarse por puntos. Si tenemos datos faltantes podemos o bien dejar un espacio en blanco o bien darles un valor que especificaremos como faltante en la variable en la opción de “Valores perdidos”.

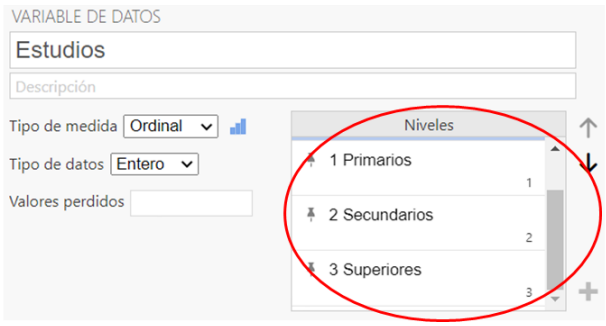
Por defecto el programa al abrirse siempre nos mostrará tres variables nominales sin datos llamadas “A”, “B” y “C”. Si clicamos encima de alguna de estas variables, se abre una pestaña que nos permite modificarla. Dar un nuevo nombre a la variable, poner una descripción (si fuera necesario), especificar el tipo de medida (nominal, ordinal o continua), si los datos de esa variable son números enteros o decimales.

Si clicamos en la flecha se minimizará la ventana y la flecha nos hará pasar a la siguiente variable.

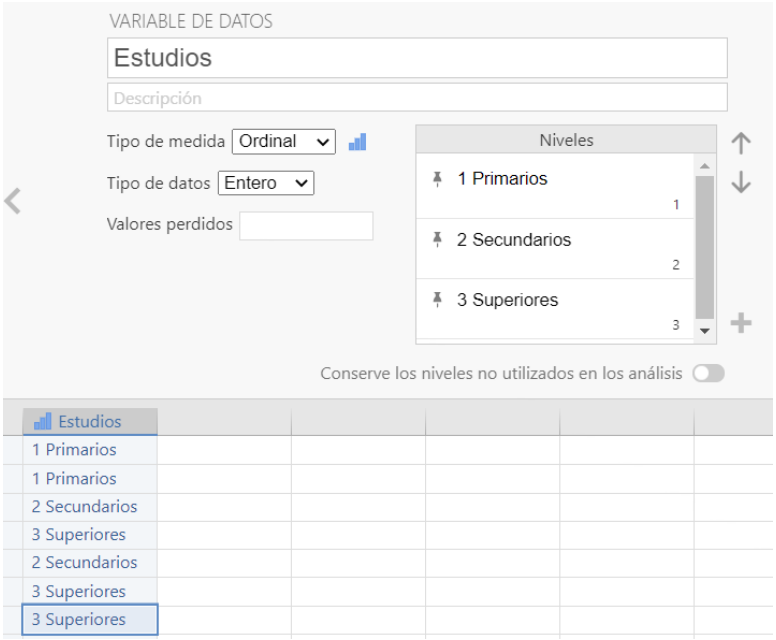


En el caso de que queramos definir variables nominales u ordinales deberemos añadir las categorias en el apartado “Niveles”. Donde primero introduciremos el valor numérico del nivel y luego la etiqueta correspondiente. En el ejemplo de abajo hemos especificado tres niveles para la variable “Estudios” 1= Primarios, 2= Secundarios y 3= Superiores.

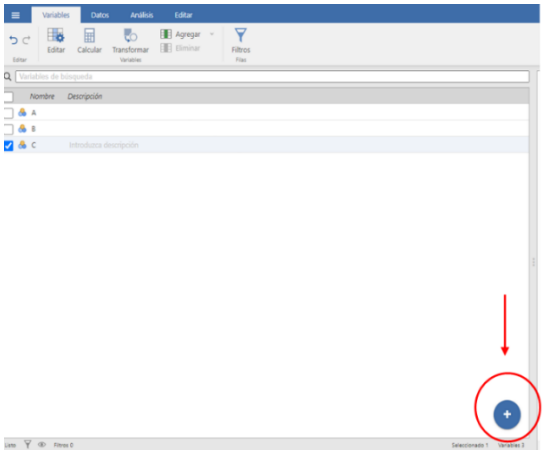


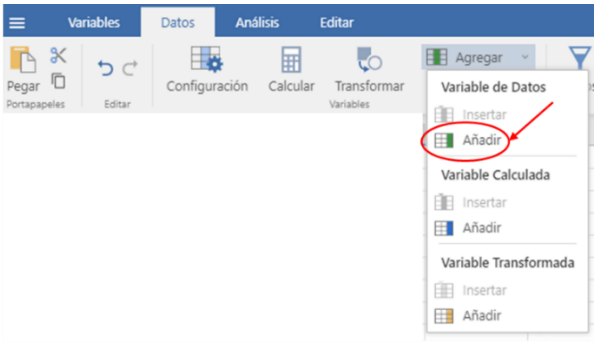


Si se insertan los datos de la variable después de definir sus niveles, las etiquetas aparecerán directamente en la cuadrícula de datos en vez de los valores numéricos asociados. Por ejemplo, para la variable “Estudios” si introducimos el número 3 directamente se cambiará por la etiqueta asociada “3 Superiores”



Si quisieramos añadir más variables podríamos hacerlo o bien desde el menú de variables, clicando en el símbolo + o bien desde la opción datos clicando en “Agregar”.



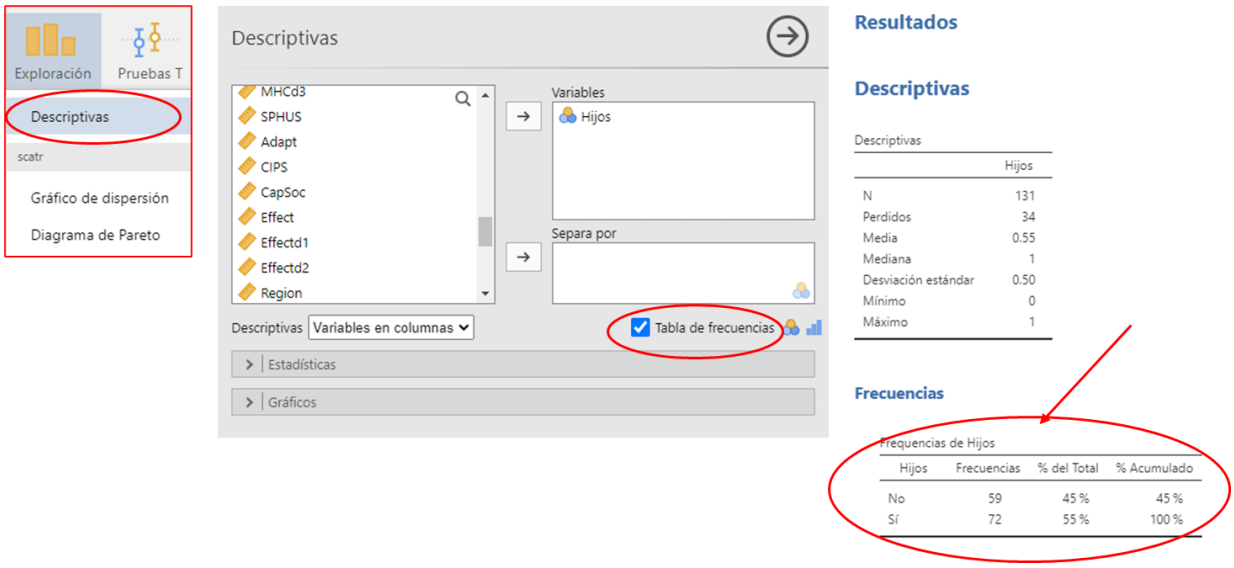


Para abrir un conjunto de datos Jamovi “.omv” o importar una base de Excel, SPSS, SAS, Stata, R clicaremos en ☰, en la opción “Abrir” y nos dará la opción de Navegar para buscar el lugar donde tengamos almacenada la base de datos.

4. Depurar datos

Depurar una base de datos consiste en detectar los valores anómalos y tratarlos de forma que no perturben los resultados de los análisis.

Antes de realizar cualquier análisis tenemos que revisar que no hay errores en la base de datos. Deberemos comprobar que no hay valores imposibles, que todos los valores estén codificados correctamente, que no queda ningún valor sin codificar y que los valores perdidos se corresponden realmente con respuestas incompletas. Para ello, podemos examinar cada variable con estadísticos descriptivos o con tablas de frecuencias (en caso de variables categóricas). Por ejemplo, si en nuestra base de datos tenemos una variable que hemos definido como “Hijos” y codificado como 1= Sí, 0=No, podemos pedir una tabla de frecuencias para ver que no haya ningún otro valor que no corresponda.

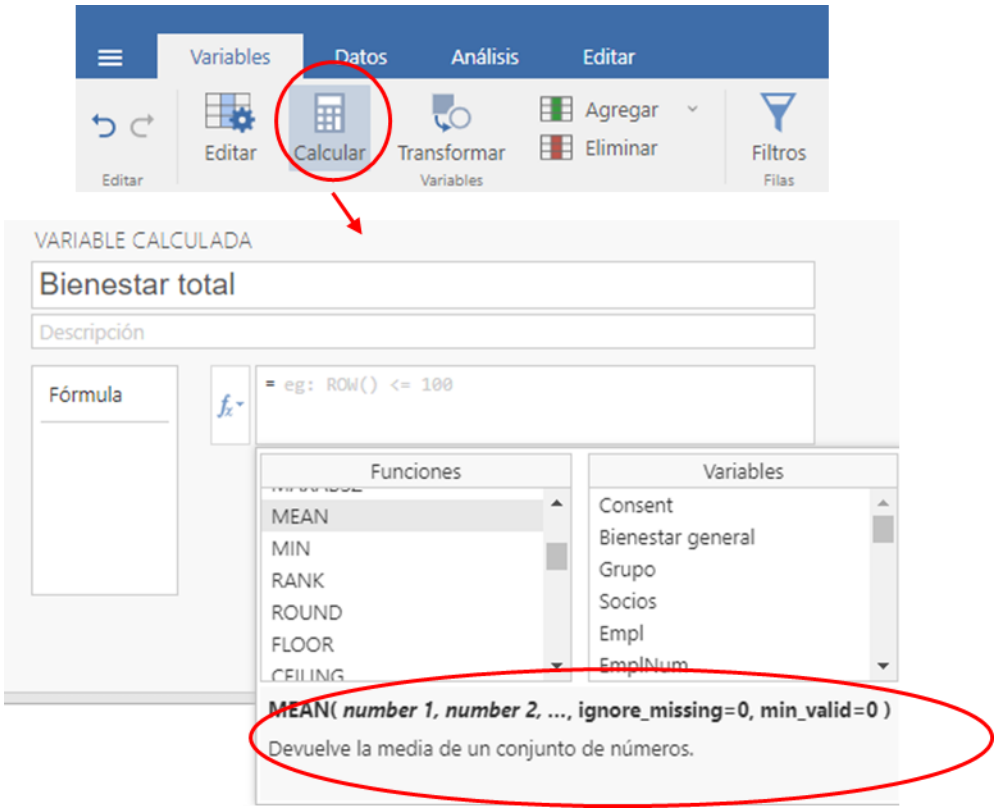


Además, deberemos asegurarnos de que las variables tienen la medida (nominal, ordinal, continua) que les corresponda ya que Jamovi es sensible a cómo declaremos la medida de nuestras variables y si las tenemos mal declaradas no nos dejará introducirlas en ciertos análisis.

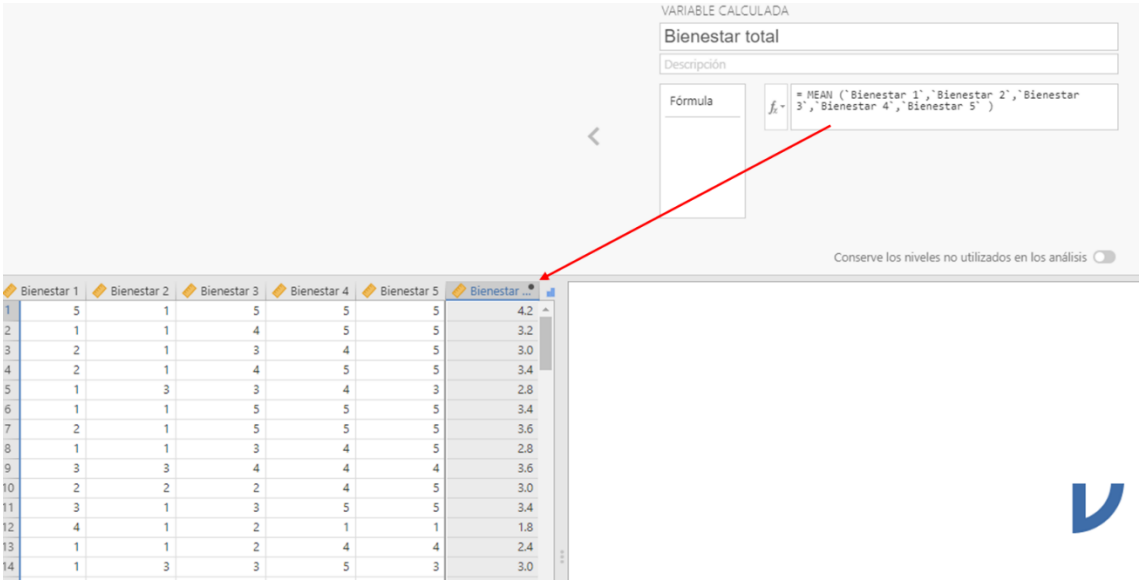
5. Calcular y recodificar variables

Muchas veces nos interesará calcular variables nuevas a partir de las que tenemos en nuestra base de datos. Por ejemplo, si tenemos una base de datos que contiene cinco variables que son ítems de una escala de Bienestar y sabemos que la escala se genera haciendo la media de estos ítems, podemos calcularla. Para calcular esta nueva variable de “Bienestar total” debemos ir a “Variables” y clicar en “Calcular”.

Inmediatamente aparecerá una nueva variable a la que podemos dar nombre, si clicamos en f_x, nos aparecerán las posibles funciones matemáticas y como debemos escribirlas, en este caso como queremos hacer la media buscaremos MEAN. Otro calculo común es querer sumar variables, este se busca como SUM.

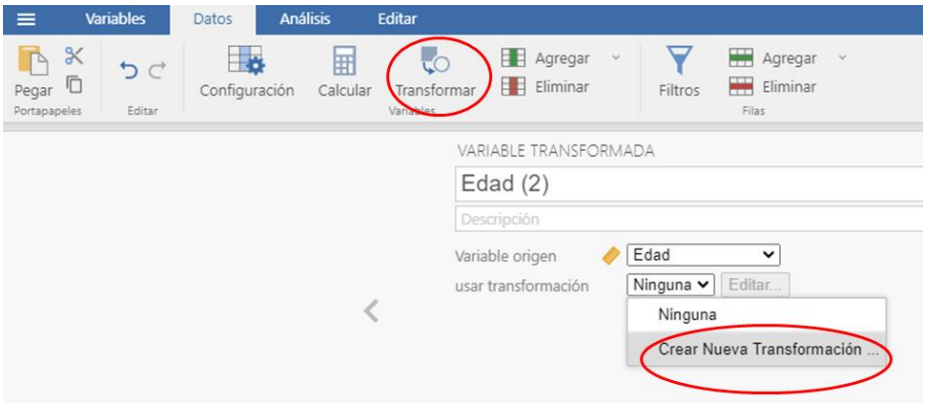


Una vez definido el cálculo de la variable, la nueva variable generada aparecerá en el conjunto de datos.

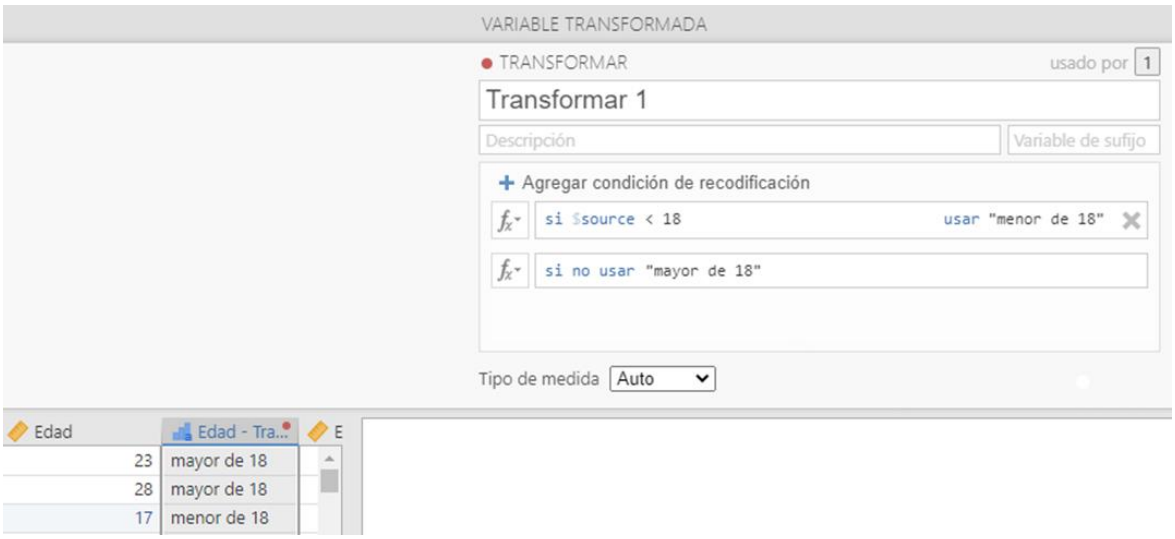


También es muy común recodificar variables, es decir, querer aplicar transformaciones a las variables que tenemos. Por ejemplo, puede ser que tengamos la

variable “Edad” medida en años que tiene la persona, pero queramos transformarla en dos grupos de edad, un grupo de menores de edad (menores de 18 años), y un grupo de mayores de edad (igual o mayor a 18 años). Para aplicar transformaciones a variables ya existentes podemos clicar encima de la variable que queramos transformar y en “Transformar”, por defecto se generará una copia de la variable original. En esta copia iremos a “usar transformación” e indicaremos “Crear Nueva Transformación”



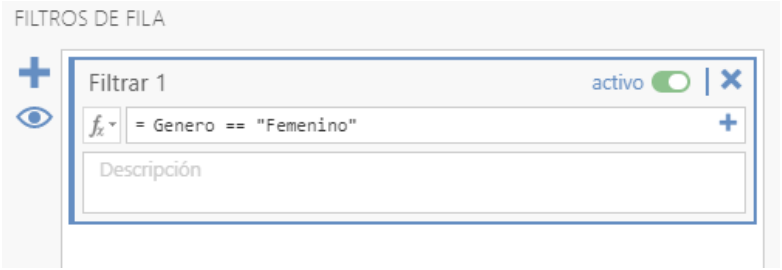
Se abrirá una nueva pestaña donde dejará que definamos la transformación y dará la opción de darle un nombre concreto a esta transformación, en nuestro caso hemos definido un grupo como <18 y el otro de mayores de edad para todos los casos donde no se cumpla esa condición, si hubiéramos necesitado definir más condiciones (por ejemplo, si quisiéramos definir tres grupos de rangos de edad) podríamos hacerlo clicando en “agregar condición de recodificación”.



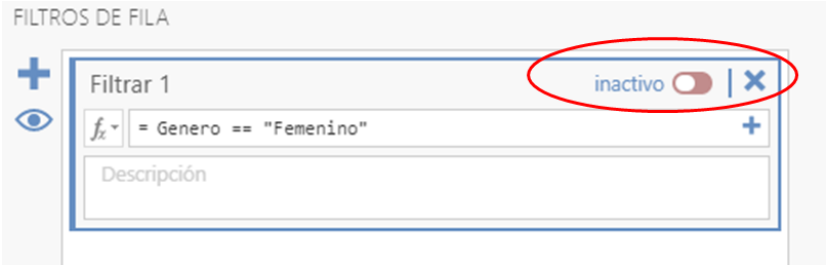
6. Aplicar filtros

Los filtros en Jamovi nos dan la posibilidad de estudiar un subconjunto de participantes que tenga alguna característica determinada. Los filtros se aplican clicando en “Datos” y en 

, generando una nueva variable “Filtrar 1”. Por ejemplo, si nos interesará estudiar al subconjunto de participantes de género femenino, aplicaríamos el siguiente código en el filtro:

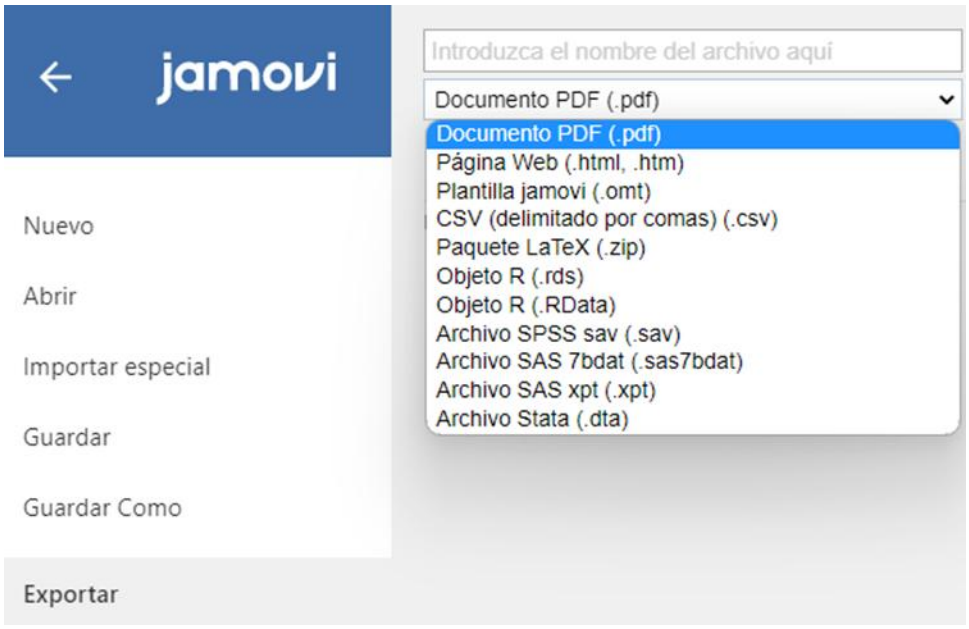


Y si en posteriores análisis queremos volver a usar el conjunto completo de datos podemos desactivar el filtro.



7. Guardar datos y resultados

La base de datos puede ser guardada como archivo Jamovi, clicando en  y en guardar. Si queremos guardar en otro programa, marcaremos la opción “Exportar” e indicaremos una de las opciones de la siguiente lista:


jamovi
Introduzca el nombre del archivo aquí
Documento PDF (.pdf)
Documento PDF (.pdf)
Página Web (.html, .htm)
Plantilla jamovi (.omt)
CSV (delimitado por comas) (.csv)
Paquete LaTeX (.zip)
Objeto R (.rds)
Objeto R (.RData)
Archivo SPSS sav (.sav)
Archivo SAS 7bdat (.sas7bdat)
Archivo SAS xpt (.xpt)
Archivo Stata (.dta)
Nuevo
Abrir
Importar especial
Guardar
Guardar Como
Exportar

Referencias

The jamovi project (2022). jamovi. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2022-01-01).

TEMA 2

Análisis descriptivo

Zaira Torres Romero
Adrián García Mollá
Sara Martínez Gregorio

Tema 2. Análisis descriptivo

1.Introducción

La Estadística se define como la ciencia que resume y organiza los datos con el objetivo de extraer información y elaborar conclusiones. Dentro de la Estadística, **la Estadística descriptiva** es la rama que se encarga de describir, analizar y representar datos relativos a una o varias muestras con el objetivo de resumir información. En este tema presentaremos distintos análisis descriptivos y cómo ejecutarlos e interpretarlos con Jamovi.

2.Tablas de frecuencias

Las frecuencias de una variable son una característica descriptiva que nos permite conocer cuántos casos hay en cada categoría o modalidad de la variable. Las tablas de frecuencias son adecuadas para variables cualitativas o semicuantitativas, en las tablas veremos representados estos tres tipos de frecuencias:

- Absolutas
- Relativas
- Relativas acumuladas

Frecuencias absolutas

Son el conteo de cada modalidad de la variable. El número de veces que se repite cada uno de los valores de una variable. La suma de todas las frecuencias absolutas representa el total de la muestra (n). Por ejemplo, si en una muestra de 500 personas que asisten a una clínica determinada tenemos 400 mujeres y 100 hombres. La frecuencia absoluta de hombres en la variable género es de 100, y la frecuencia absoluta de mujeres es 400.

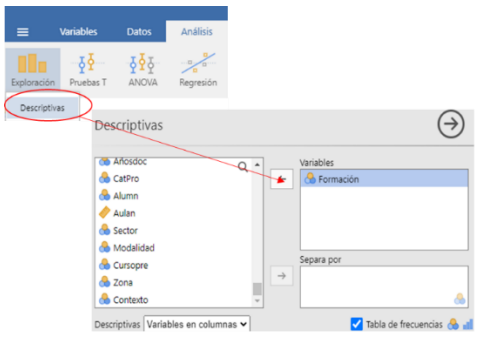
Frecuencias relativas

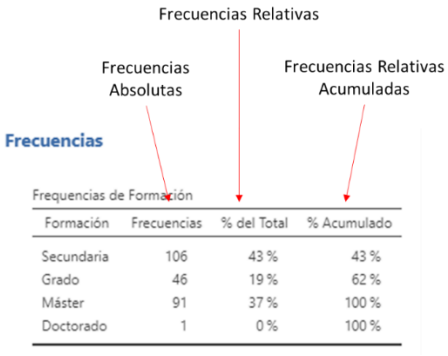
También se denominan proporción. Son el cociente entre la frecuencia absoluta de una modalidad de la variable y el total del tamaño de la muestra (n). Por ejemplo, siguiendo con el ejemplo anterior, la frecuencia relativa de hombres en la variable género sería de 100/500 = 0.2, y la frecuencia relativa de mujeres en la variable género sería 400/500 = 0.8. Si multiplicamos por 100 la frecuencia relativa obtendremos el **porcentaje**, tendríamos 20% de hombres y 80% de mujeres para la muestra de este ejemplo.

Frecuencias relativas acumuladas

También se denominan proporción acumulada. Son el cociente entre la frecuencia relativa absoluta de una modalidad de la variable y el total del tamaño de la muestra (n). Por ejemplo, si para la variable estado civil, tenemos 96 personas repartidas en cuatro categorías: casado (71 personas) soltero (5 personas), divorciado (10 personas) y viudo (10 personas) y queremos saber frecuencia acumulada de casados, solteros, y divorciados, tendremos que dividiremos la frecuencia absoluta acumulada de casados solteros y divorciados entre el total de la muestra 86/96= 0.896. Si multiplicamos por 100 la frecuencia relativa acumulada obtendremos el **porcentaje acumulado**, 89.6% de la muestra son casados solteros o divorciados.

Para calcular las tablas de frecuencias en Jamovi, iremos a “Análisis”, clicaremos en la opción “Exploración” y luego en “Descriptivos”. A continuación, introduciremos la variable que queramos en el cuadrado de “Variables” y seleccionaremos “Tabla de frecuencias”.





La variable que se ha incluido en el ejemplo indica el grado máximo de formación de las personas que se incluyen en la base de datos. Podemos ver el recuento o frecuencia absoluta de las personas en cada modalidad (por ejemplo, vemos que 106 personas tienen estudios secundarios y 46 de grado), el porcentaje relativo de cada modalidad (por ejemplo, un 19% de la muestra tiene estudios de un grado y un 37% de máster) y el porcentaje acumulado de estas frecuencias relativas acumuladas (por ejemplo, un 62% de la muestra tiene estudios secundarios o de grado).

3. Tendencia central

Las medidas de tendencia central resumen todos los datos de la muestra en un único valor que representa “el grueso” de los datos. Son medidas que sintetizan la posición que un grupo ocupa en el grado de posesión de una variable. Las más utilizadas son la Moda (Mo), la Mediana (Mdn o Md) y la Media Aritmética.

Moda

La Moda (Mo) es el valor de la variable que tiene mayor frecuencia, es decir, el valor que más se repite. Es un estadístico adecuado para representar la tendencia central de datos de variables con escalas cualitativas o cuantitativas y será el único recomendado para representar variables de escalas nominales.

Por ejemplo, si tenemos una variable X con los siguientes datos:

X= 4, 7, 5, 6, 5, 4, 5, 5, 5, 6, 5, 4, 4, 5, 6

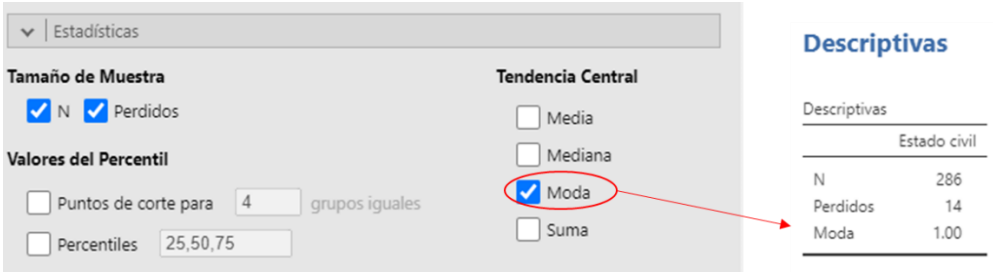
El valor con mayor frecuencia es 5, por lo que la moda sería 5, Mo = 5.

Si hay varios valores que se repiten un número igual de veces y son consecutivos, la Moda será la media de ellos. Por ejemplo, si tenemos una variable X con los siguientes datos:

X = 6, 4, 5, 3, 7, 7, 6.

Las modalidades más repetides son 6 y 7, como son consecutivas se calcula la media (6+7/2=6.5). Si estos valores no son consecutivos, por ejemplo, X = 6, 4, 5, 3, 3, 7, 6, tendremos una muestra bimodal o multimodal, en este caso la moda serían las modalidades representadas por el 3 y el 6.

Para calcular la Moda en Jamovi, iremos a “Análisis”, clicaremos en la opción “Exploración” y luego en “Descriptivos”. A continuación, introduciremos la variable que queramos en el cuadrado de “Variables” y seleccionaremos “Moda”.



En el ejemplo, hemos pedido la Moda para la variable nominal “Estado civil” en la que 1= casado, 2= soltero 3= divorciado y 4= viudo, la Moda de esta muestra es 1, por lo que la mayor parte de personas en esta muestra han indicado estar casadas.

Mediana

La Mediana es el valor que deja por debajo el 50% de los datos. La Mediana no utiliza todos los datos en su cálculo y se puede calcular en variables con escalas ordinales. Para calcular la Mediana se deben ordenar los datos de menor a mayor, calcular la posición que ocupa la mediana mediante la fórmula $Md = \frac{n+1}{2}$ y obtener el valor de la Mediana. Si el número de datos es impar la mediana es el valor central, si el número de datos es par se hace la media de las dos centrales.

Por ejemplo, si una variable se distribuye X= 4, 7, 5, 6, 5, 4, 5, 5, 5, 6, 5, 4, 4, 5, 6

Si los ordenamos tenemos: 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 7, el valor que deja por debajo el 50% de los datos, es decir que separa la muestra por la mitad es el 5, Md= 5.

Para calcular la Mediana en Jamovi, volveremos a hacer la misma ruta que en el apartado anterior (*Análisis, Exploración, Descriptivos*) introduciremos la variable y seleccionaremos “Mediana”.

En este caso la variable “Frecuencia alcohol” es una variable ordinal que hace referencia a la frecuencia con la que la persona consume alcohol (0= Nunca, 1= Excepcionalmente, 2= Mensualmente, 3= Semanalmente, 4=Diariamente). La Mediana en esta muestra es 2, lo que nos indica que un 50% de la muestra consume alcohol excepcionalmente o con menor frecuencia.

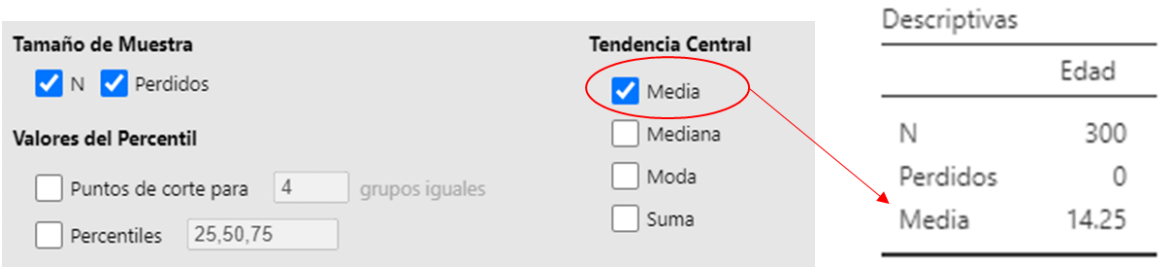
Media aritmética

La Media Aritmética es el centro de gravedad de la distribución. Es el valor resultante de sumar todos los datos que tenemos y dividir la suma entre el número total de datos (n), $\bar{X} = \frac{\sum n_i X_i}{n}$.

La Media es el estadístico de tendencia central, es un estadístico sensible ya que utiliza todos los valores para su cálculo y se ve afectado por todos los valores. La Media es el estadístico de tendencia central recomendado para variables cuantitativas, excepto si hay valores extremos o la distribución es muy asimétrica. En esos casos puede utilizarse la Mediana o la Media recortada, ya que están menos afectados por los valores extremos.

Por ejemplo, si con los datos X= 4, 7, 5, 6, 5, 4, 5, 5, 5, 6, 5, 4, 4, 5, 6 aplicamos la fórmula: $\bar{X} = \frac{\sum X_i}{n}$, obtenemos que la Media Aritmética para este conjunto de datos es $\bar{X} = \frac{76}{15} = 5.06$.

Como para el resto de los estadísticos de tendencia central, la Media Aritmética la encontraremos en Jamovi en *Análisis, Exploración, Descriptivos*, introduciremos la variable que queramos calcular y seleccionaremos “Media”. En este caso en el ejemplo se pide la media de la variable edad medida en años, la media es 14.25, lo que parece indicar que esta muestra contiene personas jóvenes.



4. Variabilidad

Para representar adecuadamente la totalidad de un grupo, no es suficiente con los valores de tendencia central, también necesitamos conocer como de diferentes son los datos de cada variable, para ello usaremos medidas de variabilidad. La variabilidad es la propiedad que nos informa del grado de heterogeneidad de un grupo. Permite describir el grado de variación o dispersión de unos datos, es decir, la similitud u homogeneidad que presentan. Una mayor dispersión indica mayores diferencias entre los datos. La variabilidad es independiente de la tendencia central y cuando se quieren describir variables se suele presentar una medida de tendencia central junto con una de variabilidad, por ejemplo, es común representar juntas la media aritmética y la desviación típica para variables cuantitativas.

Los estadísticos de variabilidad que ofrece Jamovi son la amplitud total, el rango intercuartílico, la varianza y la desviación típica.

Amplitud total, rango o recorrido

La amplitud total (AT), es la diferencia entre los valores extremos. Se calcula de una forma muy sencilla, restando a la puntuación máxima la mínima ($AT = X_{max} - X_{min}$). Es poco sensible a los valores centrales de la distribución y muy sensible a los de los extremos.

La Amplitud Total es solo una primera aproximación a la variabilidad que también se puede obtener para variables en escala nominal, indicando en ese caso la cantidad de categorías que presenta la variable.

Rango intercuartílico

El Rango intercuartílico (RIC) es la distancia entre el Percentil 25 (puntuación que deja por debajo el 25% de los datos de la muestra, también llamado cuartil 1, Q1) y el percentil 75 (puntuación que deja por debajo el 75% de los datos de la muestra, también llamado cuartil 3 Q3), se calcula con la diferencia entre el tercer cuartil y el primer cuartil dividido entre dos:

$$\text{RIC}= \frac{Q_3-Q_1}{2}$$

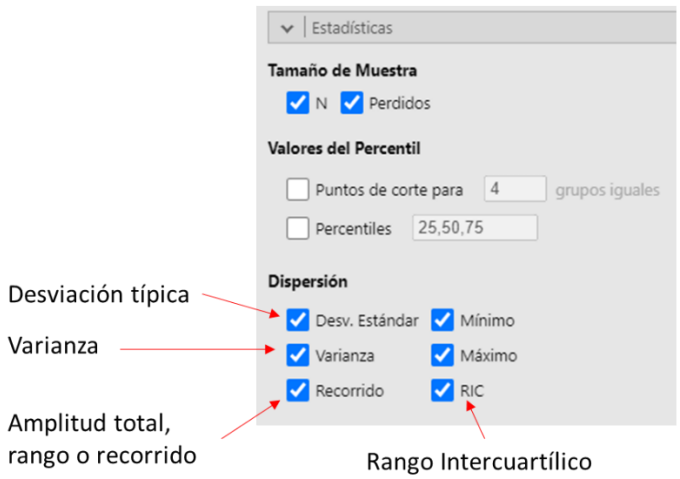
Es un estadístico robusto porque no depende de las colas de la distribución. Es una medida de variabilidad adecuada cuando la Mediana es la medida de tendencia central.

Varianza y desviación típica

La Varianza (VAR o s^2), es la media de las diferencias cuadráticas respecto a la media aritmètica, $s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$. Cuanto mayor es la magnitud de las diferencias entre los datos, más grandes son los valores de las diferencias respecto de la Media. La varianza está expresada en una escala de cuadrados.

La Desviación típica (DT o *s*) es la raíz cuadrada de la varianza, $s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}}$. La desviación típica tiene la ventaja frente a la Varianza de que tiene las mismas unidades de medida que los datos originales. Tanto la Varianza como la Desviación típica son valores siempre positivos que oscilan desde 0 a infinito. La Varianza y la Desviación típica son adecuadas para representar la variabilidad de una variable cuando lo es la media como estadístico de tendencia central.

En Jamovi todos estos estadísticos de variabilidad se encuentran en “Análisis”, “Exploración”, “Descriptivos” y en la opción “Dispersión” de la ventana emergente.



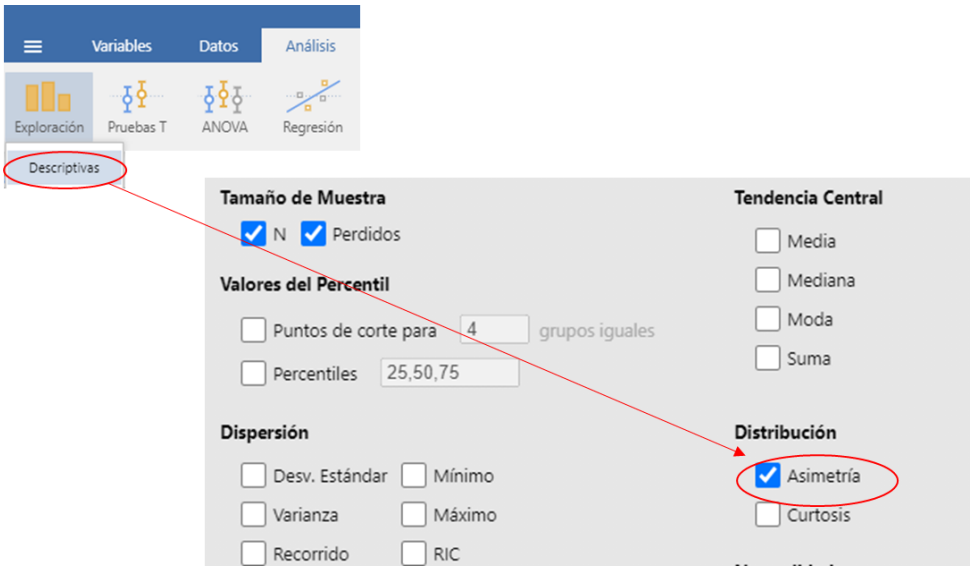
Descriptivas

	Edad
N	300
Perdidos	0
Desviación estándar	1.56
Varianza	2.43
RIC	2.00
Recorrido	6.00
Mínimo	12.00
Máximo	18.00

En este caso, para la variable “Edad” hemos calculado todos los estadísticos de variabilidad. La edad mínima de las personas de la muestra es de 12 años y la máxima de 18, la amplitud total o recorrido nos indica que son 6 años de diferencia lo que se lleva la persona más mayor de esta muestra con la más pequeña. También se puede observar que el rango intercuartílico es igual a 2, la varianza a 2.43 y la desviación típica a 1.56.

5. Asimetría

La Asimetría es un estadístico de distribución, mide el grado en que los datos se reparten equilibradamente en torno a la tendencia central. El análisis de la asimetría es adecuado para variables cuantitativas. En Jamovi, la Asimetría se encuentra en la opción de “Análisis”, “Exploración”, “Descriptivos” clicando en “Asimetría”.

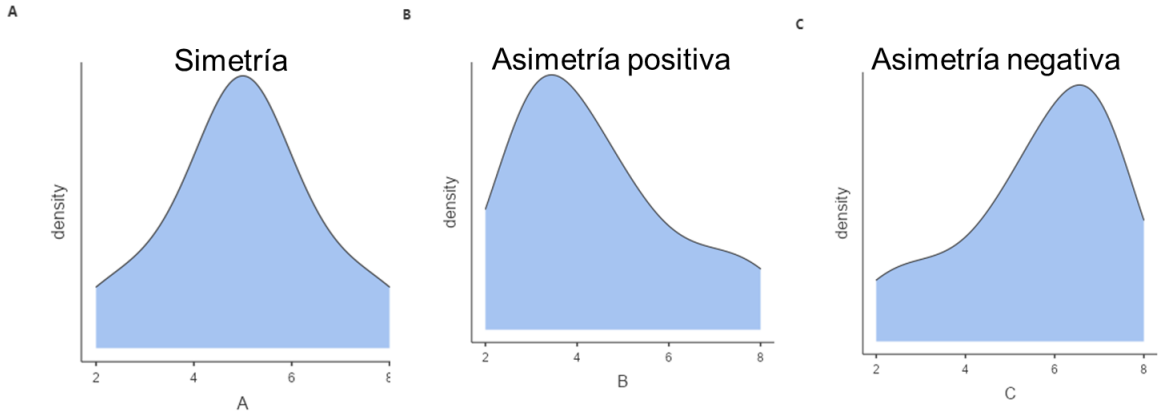


Cuando una distribución es **simétrica**, el valor de asimetría es 0 y los valores de la distribución se agrupan en valores centrales, en distribuciones simétricas la Moda, Mediana y Media son iguales. Cuando una distribución presenta **asimetría positiva**, el valor de asimetría es un número positivo, los datos tienden a estar concentrados en valores más pequeños de la variable y la Moda es más pequeña que la Mediana que a su vez es más pequeña que la Media. Por último, si la distribución presenta **asimetría negativa**, el valor de asimetría es un número negativo, los datos tienden a estar concentrados en valores más altos de la variable, en consecuencia, la Moda es más grande que la Mediana que a su vez es más grande que la Media en esa variable.

Por ejemplo, tenemos tres muestras (A, B y C) para una misma variable con los siguientes estadísticos descriptivos y que se distribuyen de la siguiente forma:

Descriptivas

	A	B	C
Media	5.00	4.38	5.62
Mediana	5	4	6
Moda	5.00	3.00	7.00
Asimetría	0.00	0.82	−0.82
Error est. asimetría	0.62	0.62	0.62



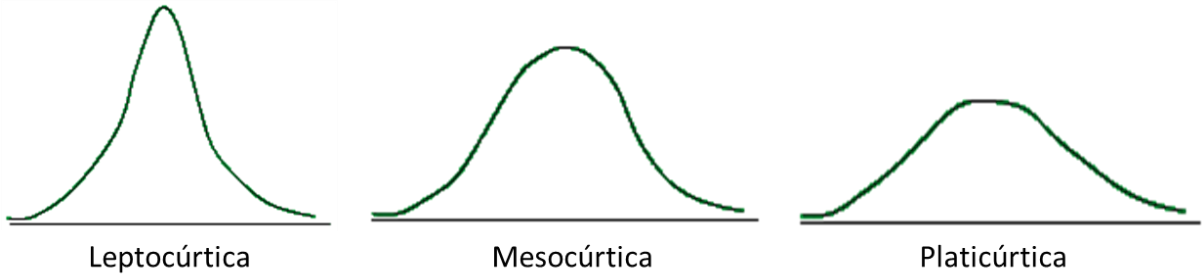
La muestra A presenta una asimetría positiva, la B una asimetría positiva y la C una asimetría negativa.

6. Curtosis

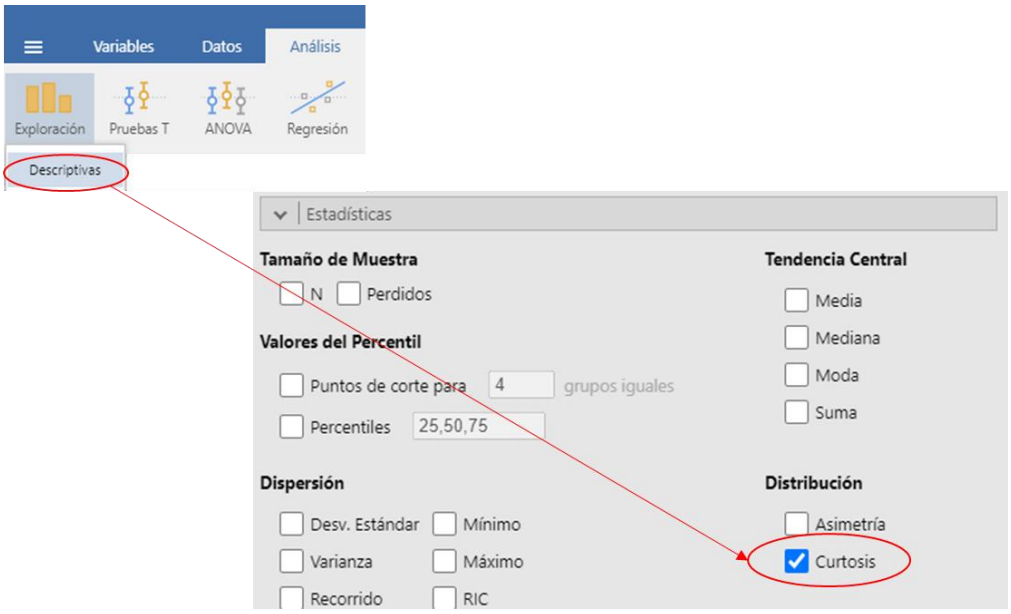
La Curtosis es el grado de apuntamiento o agudeza de una distribución respecto a la distribución normal. Al igual que la asimetría sirve para conocer cómo están distribuidos

los datos de la muestra y sólo se analiza en variables cuantitativas. La curtosis expresa el grado en que una distribución acumula casos en sus colas en comparación con los casos acumulados en las colas de una distribución normal con la misma varianza. Según la curtosis una distribución puede ser:

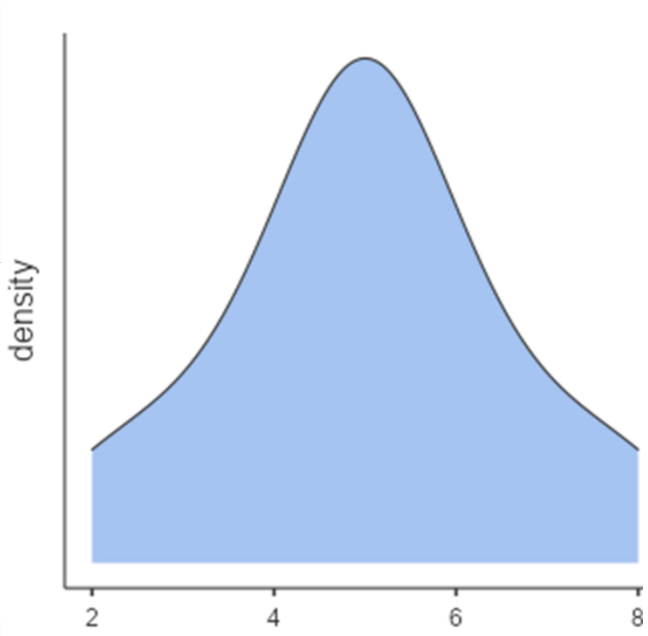
- **Leptocúrtica**: Curtosis positiva, más apuntada que la distribución normal
- **Mesocúrtica**: Curtosis 0 o alrededor de 0, igual de apuntada que la distribución normal
- **Platicúrtica**: Curtosis negativa, más aplanada que la distribución normal



En Jamovi, se encuentra siguiendo la ruta “Análisis”, “Exploración”, “Descriptivos” y clicando en “Curtosis”.



Descriptivas	
	A
N	13
Perdidos	1
Curtosis	0.44
Error est. curtosis	1.19



La curtosis en la variable del ejemplo es 0.44, por lo que sabemos que es un poco más apuntada que la distribución normal.

7. Representación gráfica

Los procedimientos gráficos tienen como finalidad facilitar la inspección visual de los datos. Hay muchos gráficos y algunas representaciones gráficas son más recomendables en función de la escala de medida de las variables. A continuación, presentaremos los gráficos que ofrece Jamovi.

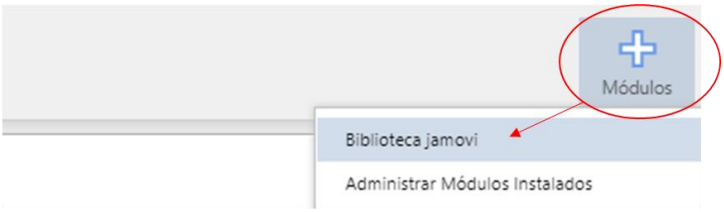
7.1. Gráficos para variables cualitativas

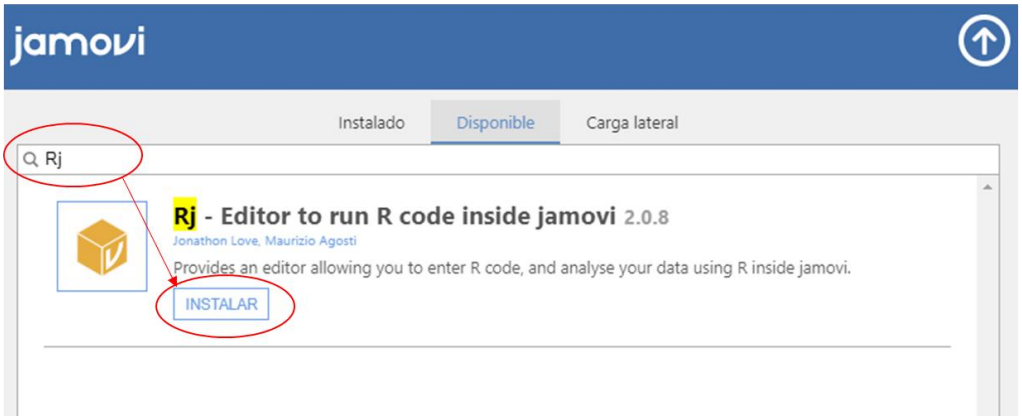
En aquellas variables que son cualitativas nos interesará representar las frecuencias o porcentajes de las diferentes categorías, para ello emplearemos diagrama de sectores o gráfico de barras.

Diagrama de sectores

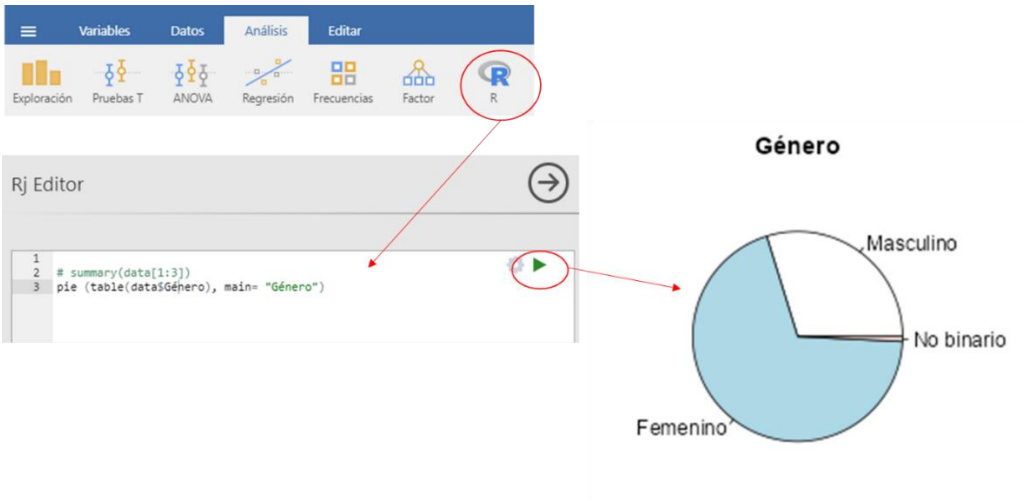
El diagrama de sectores o pictograma consiste en un círculo subdividido en áreas, cada una de las áreas es proporcional a la frecuencia de la modalidad que representa. Para hacer este gráfico debemos instalar el módulo Rj. Iremos a la sección de módulos,

entraremos a la biblioteca de Jamovi y buscaremos “Rj”, clicaremos en “instalar” una vez que nos aparezca este módulo.





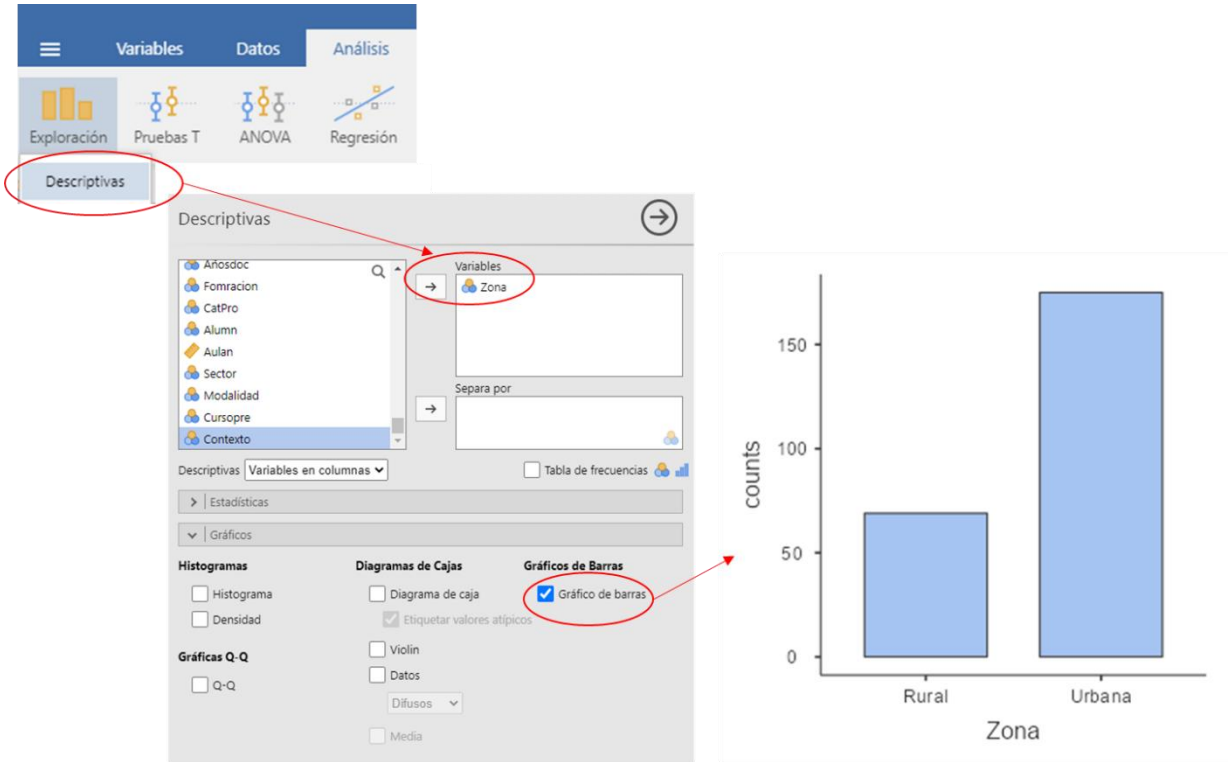
Una vez instalado, lo encontraremos en la barra de análisis. Para el diagrama de sectores, abriremos el módulo Rj y encontraremos una hoja para escribir código, el código que le debemos indicar **será *pie (table(data$nombredelavariable), main= " nombredelavariable")***. Una vez escrito clicaremos en la flecha verde para ejecutarlo y aparecerá el diagrama de sectores de la variable que le hayamos indicado en la hoja de resultados.



En este ejemplo hemos representado en un diagrama de sectores la variable “Género” medida en tres categorías (femenino, masculino, no binario). Aunque este gráfico no nos ofrece recuento de personas de la muestra dentro de cada categoría se puede observar que aproximadamente tres cuartas partes de la muestra han marcado la opción de género femenino.

Gráfico de barras

Este gráfico consiste en un conjunto de barras donde cada una representa una modalidad de la variable. La altura de cada barra es proporcional a la frecuencia de casos dentro de cada modalidad de la variable. Para representar una variable como diagrama de barras, deberemos ir a “Análisis”, clicar en la opción “Exploración” y luego en “Descriptivos”. A continuación, introduciremos la variable que queramos representar en el cuadrado de “Variables” y en la opción de “Gráficos” seleccionaremos “Gráfico de barras”.



En el ejemplo, se ha representado en un gráfico de barras la variable zona donde reside la persona “Zona”, que tiene dos modalidades (rural y urbana). La altura de las columnas nos indica el recuento de las personas de la muestra en cada modalidad (la

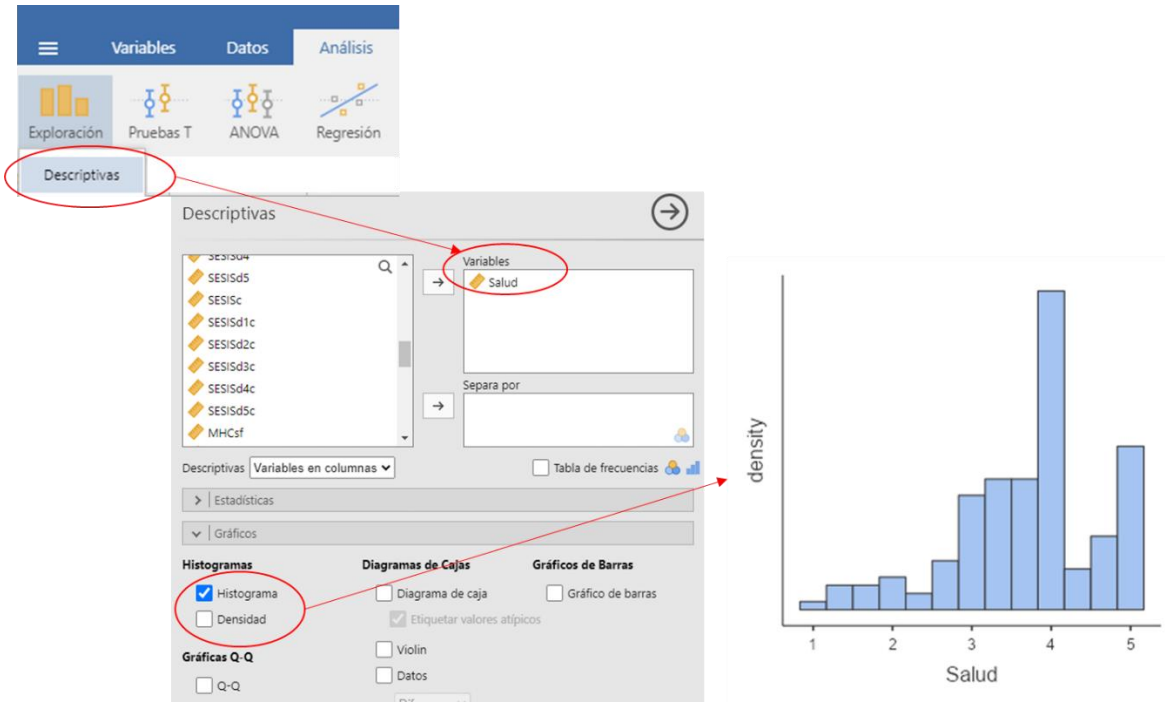
frecuencia absoluta). Podemos observar que, en esta muestra, la mayoría de las personas (más de 150) viven en zona urbana.

7.2. Gráficos para variables cuantitativas

Para variables cuantitativas, lo que nos resultará interesante es una representación gráfica que nos ayude a visualizar cómo está distribuida la muestra. Los gráficos recomendados serán el histograma y el diagrama de caja y bigotes.

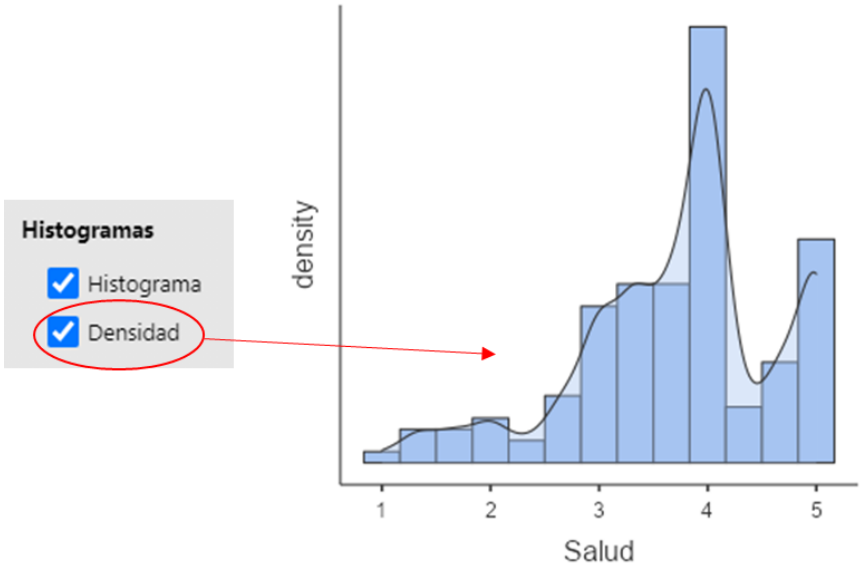
Histograma

Es en apariencia parecido al gráfico de barras, donde cada barra también indica la frecuencia de un valor numérico, pero en este caso las barras son adyacentes y suelen ser mucho más numerosas. Para representar una variable con el histograma, también deberemos ir a “Análisis”, clicar en la opción “Exploración” y luego en “Descriptivos”. A continuación, introduciremos la variable que queramos representar en el cuadrado de “Variables” y en la opción de “Gráficos” seleccionaremos “Histograma”.



En este caso hemos representado una escala general de salud, con valores entre el 1 y el 5 (con valores más altos indicando mejor salud). Para esta muestra vemos que

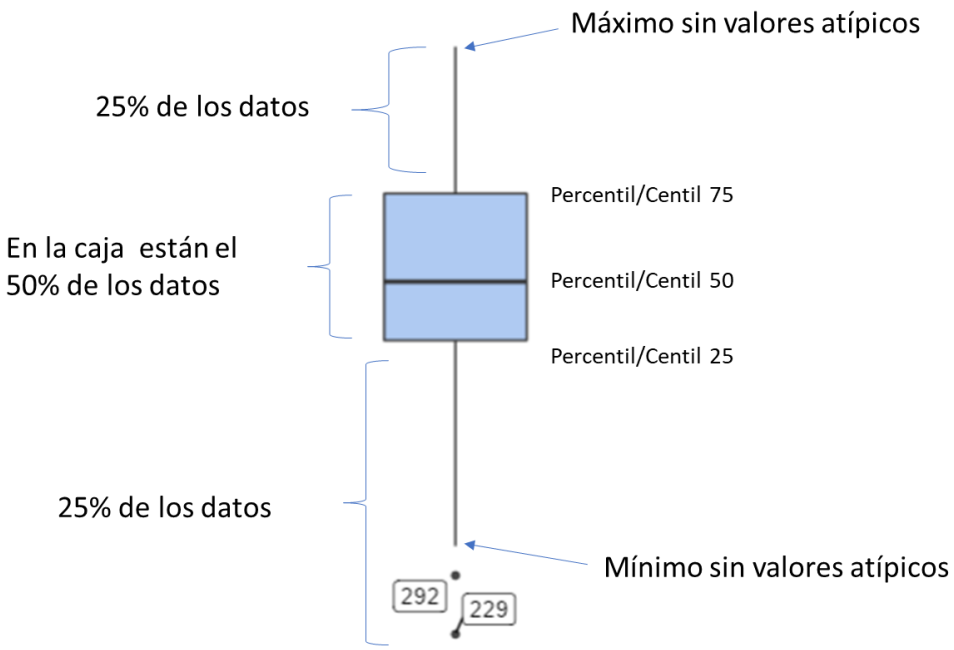
la distribución es asimétrica positiva, con la mayor parte de los valores se concentran en valores más altos de la escala y que el “pico” del histograma que representaría la Moda se encuentra alrededor del número 4.



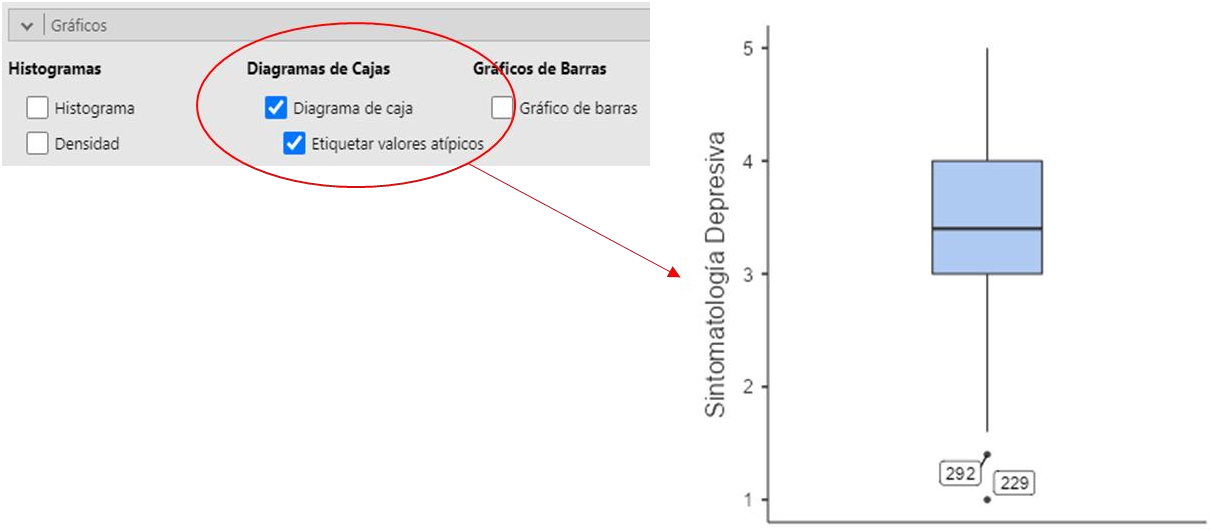
También podemos marcar la opción “Densidad” para que el histograma nos muestre la curva suavizada. Esta curva tiene como finalidad que se pueda visualizar mejor la forma de la distribución.

Diagrama de caja

El diagrama de caja también sirve para visualizar cómo está distribuida la muestra. Indica de manera proporcional la distancia entre el valor más alto y el más bajo (una vez excluidos los valores extremos). Da información de dónde se encuentran los percentiles 75, 50 y 25, que indican la puntuación que deja por debajo el 75%, 50% y 25% de la muestra. La Mediana coincide con el percentil 50 (la línea negra que parte la caja). Además, el diagrama de caja nos indica si hay casos atípicos en la variable (casos de puntuaciones muy extremas que representa como círculos).



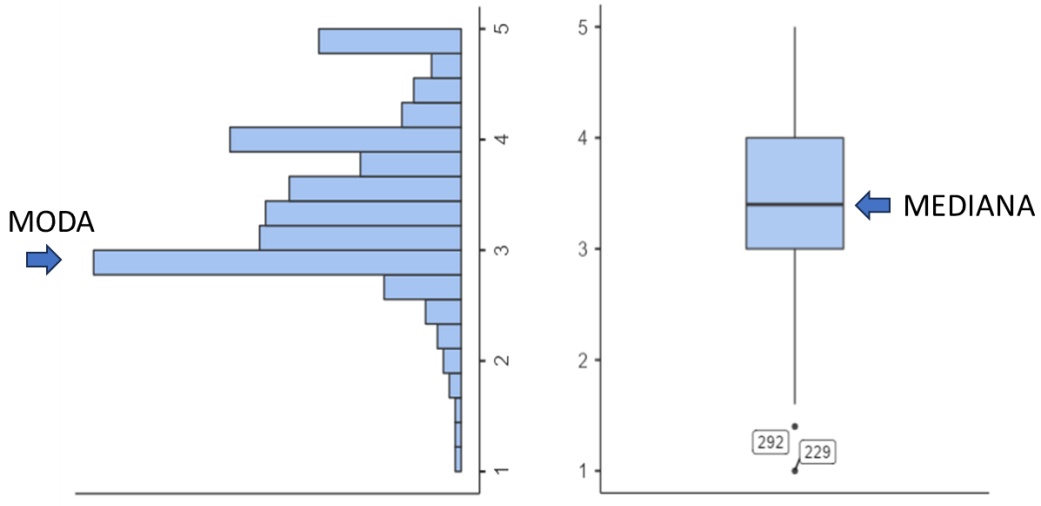
Para hacer el diagrama de caja, deberemos ir a “Análisis”, clicar en la opción “Exploración” y luego en “Descriptivos”. A continuación, introduciremos la variable que queramos representar en el cuadrado de “Variables” y en la opción de “Gráficos” seleccionaremos “Diagrama de caja”. La opción “Etiquetar valores atípicos” nos indicará qué número de caso en la base de datos es el que tiene el valor atípico (en caso de que la variable tenga valores atípicos).



En el ejemplo, está representada la variable “sintomatología depresiva”, es una escala que mide si la persona presenta o no ciertos síntomas de depresión, con valores entre 1 y 5 (valores más altos indicarían una mayor sintomatología depresiva). En este

caso la muestra presenta asimetría negativa y está concentrada entorno a valores en sintomatología depresiva más altos en la variable. El 50% de las puntuaciones centrales (la caja) se encuentran entre el 3 y el 4. El 75% de las puntuaciones son iguales o inferiores a 4 (límite superior de la caja) y el 25% de la muestra presenta puntuaciones inferiores a 3 (límite inferior de la caja). Podemos ver que hay dos casos de puntuaciones atípicas con valores muy inferiores al resto de la muestra (personas número 292 y 229 en la base de datos).

El histograma y diagrama de caja y bigotes representan la distribución de la muestra, aportan información sobre algunos marcadores de tendencia central, sobre la asimetría de la muestra y la variabilidad. A continuación, se presenta el histograma girado de la misma variable de sintomatología depresiva para que se pueda observar esta equivalencia en ambas representaciones gráficas.



Referencias

The jamovi project (2022). jamovi. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2022-01-01).

TEMA 3

Introducción análisis inferencial

Zaira Torres Romero
Sara Martínez Gregorio
Adrián García Mollá

Tema 3. Introducción análisis inferencial

1. Conceptos básicos

Hasta ahora nos hemos enfocado en la Estadística Descriptiva que se centra en describir las características de los datos disponibles. No obstante, muchas veces, el análisis de datos tiene como finalidad extraer conclusiones de datos no observados. En otras palabras, a inferir lo qué ocurrirá en la población a partir de los datos de las muestras que tenemos disponibles.

Los análisis en los que la finalidad es extrapolar los resultados de los datos disponibles a los resultados que obtendríamos si pudiéramos observar toda la población pertenecen a la **Estadística Inferencial**.

Como ya hemos visto en capítulos anteriores a partir del análisis de las muestras obtenemos estadísticos. Cualquier muestra puede contener cierto grado de error a la hora de representar a su población, por eso diremos que los estadísticos son estimadores o aproximaciones al verdadero valor de la población, al parámetro. Hay dos tipos de estimaciones, estimación puntual y por intervalo.

2. Estimación

Estimación puntual

La estimación puntual consiste en extraer un único valor, un estadístico, para estimar el valor del parámetro. Por ejemplo, la media de una muestra sería un estadístico que resume en un único valor la tendencia central de la distribución. Dependiendo del parámetro que queramos estimar existen diversos valores. Como norma general, se intentará que los estimadores puntuales sean:

Insesgados o sin sesgo: El sesgo seria la diferencia entre la esperanza (o valor esperado) del estimador y el verdadero valor del parámetro. Lo deseable es que el sesgo sea lo menor posible. Por ejemplo, la media es un estimador insesgado pero la varianza no.

Consistentes (Asintótico): Diremos que un estimador es consistente si a medida que aumentamos el tamaño de la muestra se aproxima al verdadero valor del parámetro.

Eficientes: Un estimador será eficiente si tiene poco error estándar, poca varianza. Por ejemplo, la media es más eficiente que la mediana.

Suficientes: Un estimador es suficiente si resume la información más relevante. Por ejemplo, la media de una muestra sería un estimador suficiente de la media de la población.

En el tema anterior de Estadística Descriptiva ya vimos como calcular los estadísticos más comunes con Jamovi (en el apartado “Exploración”, “Descriptivas”, y desplegando la pestaña de “Estadísticos”):

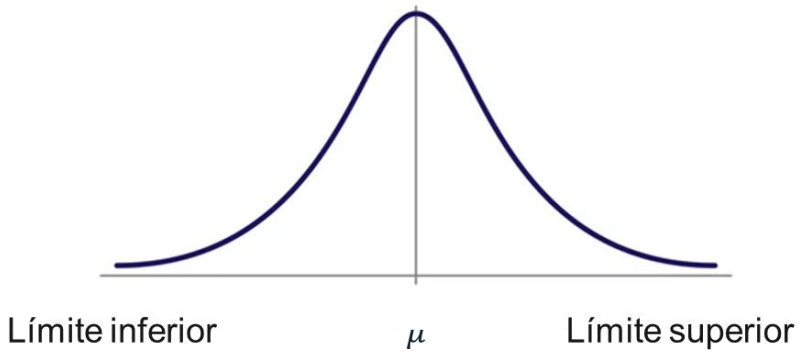
Estimación por intervalo

Aunque mediante estimación puntual elijamos el estimador de características más deseables, ofrecer un solo valor numérico, no da información sobre el error que puede contener esa estimación. La estimación por intervalo ofrece un rango de valores numéricos posibles para el parámetro marcando dos límites (el **Límite inferior (Li)** y el **Límite superior (Ls)**). Además, ofrece la probabilidad de que el valor del parámetro se encuentre dentro de ese rango de valores.

Esta probabilidad de que el valor del parámetro este dentro del rango de valores, viene determinada por la **confianza (C)** de nuestro intervalo. El parámetro poblacional pertenecerá al intervalo calculado con una confianza del C%. Generalmente se suele emplear una confianza del 95% o del 99%. La probabilidad de que el parámetro no se encuentre dentro del intervalo es complementaria al nivel de confianza, se le llama **error** y se simboliza como **α**. Para un **intervalo de confianza (IC)** del 95%, la probabilidad de error en un 5% y para un IC de 99% la probabilidad error es un 1%.

El nivel de confianza y la **amplitud** del intervalo varían conjuntamente. Un intervalo con un mayor nivel de confianza será más amplio y tendrá una mayor probabilidad de acierto del valor del parámetro. Mientras que, un intervalo con menor nivel de confianza será más pequeño, ofrecerá una estimación más precisa, pero con una mayor probabilidad de error. También deberemos tener en cuenta el tamaño muestral, cuanto más grande sea una muestra, más pequeño será el error estándar y mayor precisión tendremos.

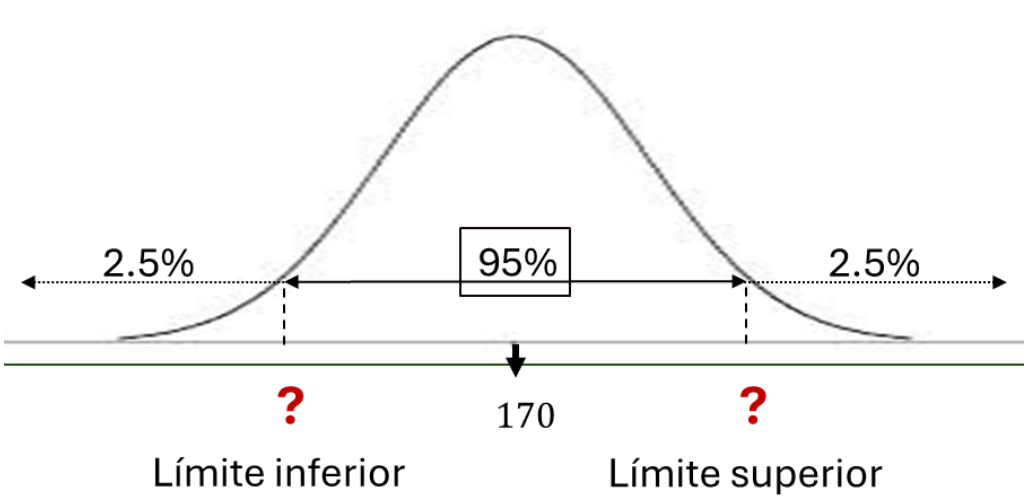
Para calcular un intervalo de confianza también es necesario conocer la distribución muestral. En el caso de la media esta suele seguir una distribución Normal como la que se muestra en la imagen:



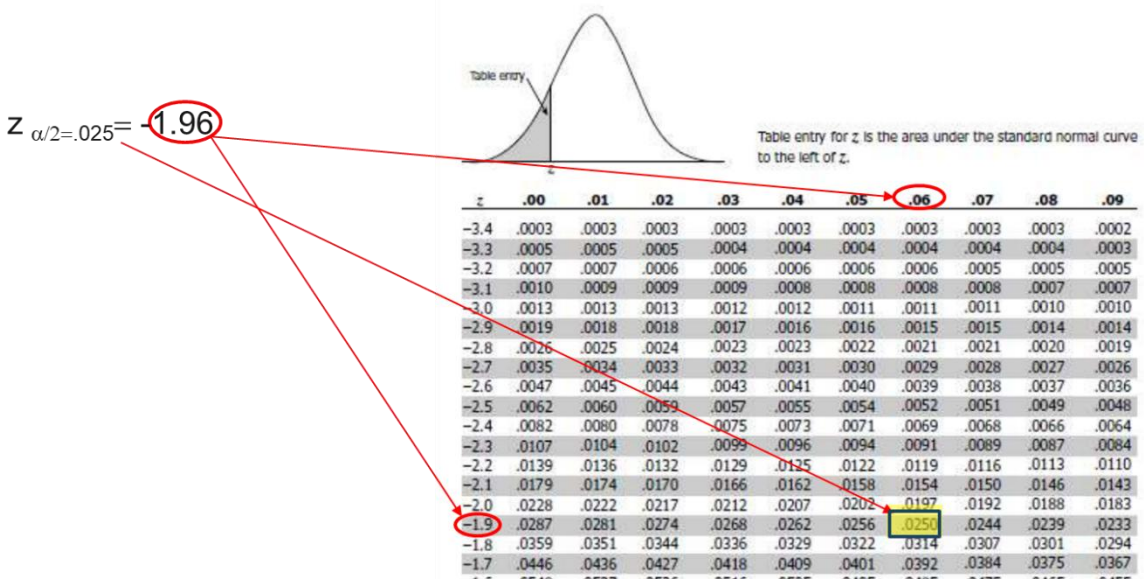
Los límites superiores e inferior señalan la puntuación por debajo o por encima de la cuál es muy improbable que se sitúe el valor de la media poblacional.

Veamos un ejemplo de estimación por intervalo, supongamos que queremos estimar un intervalo de confianza (IC) de la media de altura de la población española. Sabemos que la muestra que hemos recogido se distribuye según la distribución normal y sabemos que su media es de 170 cm; su desviación típica de 15 y el tamaño de la muestra es n= 38.

El primer paso para una estimación por intervalo será fijar el nivel de confianza o riesgo. Supongamos que en este caso queremos un IC del 95%, ósea un nivel de riesgo del 5% y un α= 0.05.



El segundo paso será buscar los dos límites en la distribución muestral. Como sabemos que la muestra sigue una distribución normal podríamos buscar en la tabla z. El error que estamos dispuestos asumir (α= 0.05) se reparte en cada límite, así que buscaremos los valores z $\alpha/2$=0.025 y z 1-$\alpha/2$=0.975. Estos valores corresponden con los valores -1,96 y 1.96.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367

Calcularíamos el error estándar (ES) y el error máximo (EM)

ES de la media= $\frac{\sigma}{\sqrt{n}}$= $\frac{15}{\sqrt{38}}$=2.43

EM=ES x z = 2.43 x 1.96= 4.76

Por último, calcularíamos los dos límites del intervalo de confianza:

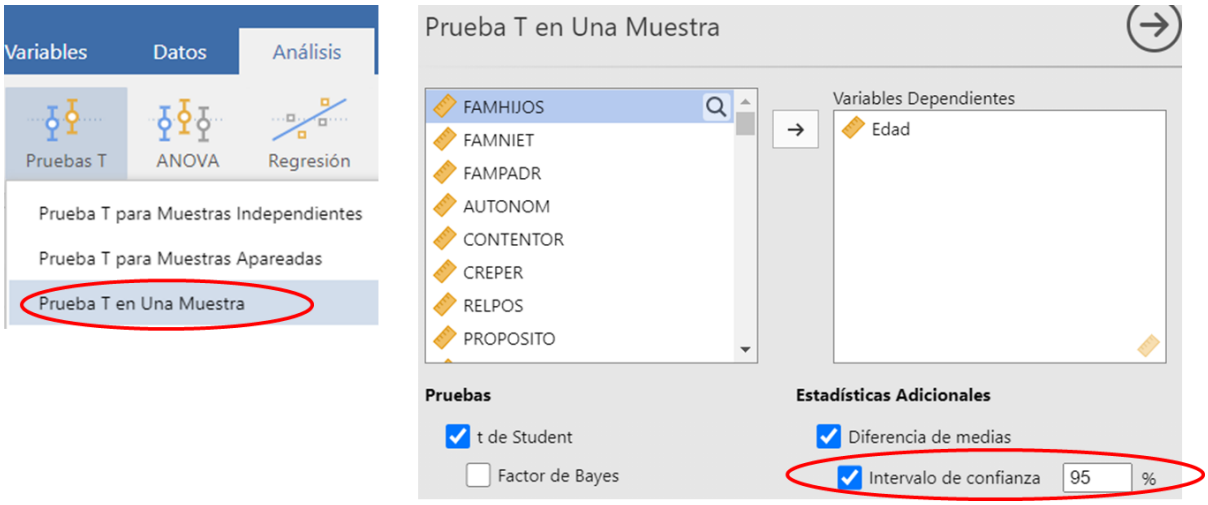
Límite Superior (LS) = 170 +4.76 = 174.76

Límite Inferior (LI) = 170 – 4.76 = 165.23

La media de la altura de la población española con un 95% de confianza está situada entre 165.23 y 174.76 cm.

Intervalo de confianza para una media en Jamovi

Jamovi nos permite obtener fácilmente los intervalos de confianza alrededor de una media para una variable cuantitativa. Debemos seleccionar el menú “Análisis”, “Pruebas T”, “Pruebas T para una muestra” y seleccionar la opción “Intervalo de confianza”.



Prueba T en Una Muestra

Prueba T en Una Muestra

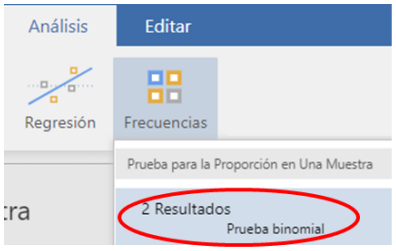
						Intervalo de Confianza al 95%	
		Estadístico	gl	p	Diferencia de medias	Inferior	Superior
Edad	T de Student	161.60	281.00	< .001	63.06	62.29	63.82

Por defecto nos dará los valores para un intervalo de confianza del 95% pero podemos cambiar este porcentaje al que queramos.

Además, con Jamovi, también podemos obtener los intervalos de confianza para las Pruebas T de comparación de dos medias que veremos en el punto 6. Siempre encontraremos la opción de intervalo de confianza en el apartado de “Estadísticas adicionales” de los análisis.

Intervalo de confianza para una proporción en Jamovi

En el caso de tener una variable categórica, podemos calcular el intervalo de confianza para las proporciones de sus categorías. Para ello deberemos clicar en “Análisis”, “Frecuencias” y en prueba para la proporción en una muestra seleccionar “2 Resultados”.



Prueba de Proporciones (2 Resultados)

Prueba Binomial

						Intervalo de Confianza al 95%	
	Nivel	Frecuencia	Total	Proporción	p	Inferior	Superior
GENERO	Hombre	84	286	0.29	< .001	0.24	0.35
	Mujer	202	286	0.71	< .001	0.65	0.76

Nota. H_a es proporción ≠ 0.5

Nos ofrecerá un intervalo de confianza para las proporciones de cada categoría de la variable.

3. Contraste de hipótesis

Definición y conceptos

Ya introdujimos en concepto de hipótesis en temas anteriores y sabemos que las hipótesis son enunciados verificables. En este tema veremos qué es el contraste de hipótesis y cómo resolverlo. En concreto, nos centraremos en las hipótesis estadísticas, aquellas que se pueden poner a prueba mediante análisis estadísticos.

En cualquier análisis debemos partir con la idea de que no hay ningún efecto (diferencia, relación…) mientras no se demuestre lo contrario. Esto se plasma en el contraste de hipótesis que comenzaremos formulando una primera hipótesis de nulidad, que se pone a prueba para mantener o rechazar.

En el contraste de hipótesis definiremos dos opciones, dos hipótesis contrapuestas:

H0 = Hipótesis nula, supone la ausencia de efectos, es la que contiene la igualdad de los estadísticos puestos a prueba

H1 = Hipótesis alternativa, contiene la diferencia entre los estadísticos puestos a prueba

Estas hipótesis deben ser **exhaustivas** (contener todas las posibilidades) y **excluyentes** (no pueden solapar opciones en común).

Pongamos un ejemplo, supongamos que un estudio clínico, se plantea si un medicamento nuevo para reducir la presión arterial tiene un efecto significativamente diferente al de un placebo. La hipótesis nula sugeriría que no hay diferencia en la reducción de la presión arterial entre el medicamento y el placebo, mientras que la hipótesis alternativa propondría que el medicamento tiene un efecto diferente al del placebo:

Hipótesis nula (H0): $\mu_{medicamento} = \mu_{placebo}$

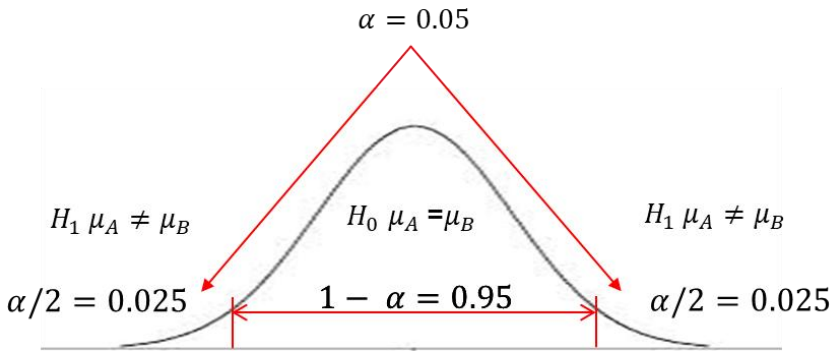
Hipótesis alternativa (H1): $\mu_{medicamento} \neq \mu_{placebo}$

También distinguiremos dos tipos de contrastes de hipótesis, el bilateral y el unilateral. Diremos que el contraste es bilateral si no importa el sentido de la diferencia. El contraste bilateral es el estándar, **en él,** la hipótesis nula se plantea de forma que si no es cierta las diferencias pueden estar en un sentido u otro. Por ejemplo, imaginemos que queremos saber si dos tratamientos (A y B) son igual de efectivos para tratar la depresión o hay alguno que sea más efectivo. Plantearíamos nuestro contraste de manera bilateral, ya que si hay diferencias nos interesan tanto si es más efectivo el tratamiento A como si lo es el tratamiento B.

$$\boldsymbol{H_0\ \mu_A = \mu_B}$$

$$\boldsymbol{H_1\ \mu_A \neq \mu_B}$$

En los contrastes bilaterales, el error que asumimos en el análisis (α) se divide en dos regiones. En **un IC del** 95%, **el** error del 5% **se repartiría en los dos sentidos**:



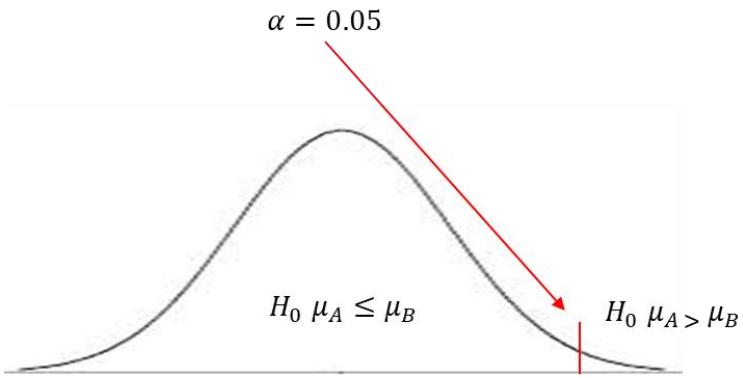
Por el contrario, el contraste será unilateral si lo planteamos de forma que interesa el sentido de la diferencia. Por ejemplo, en caso de que pensemos que las diferencias favorecen a un grupo o sólo nos interesa examinar las diferencias en un

sentido. Supongamos que un nuevo tratamiento B se plantea como más efectivo para disminuir la depresión que el tratamiento A. Aquí nuestras hipótesis plantearían si la media de depresión con el tratamiento A es mayor o igual que con el tratamiento B (H0), o si la depresión con el tratamiento B es menor que con el A (H1).

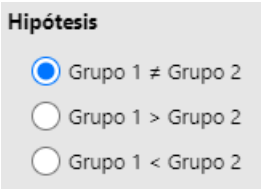
$$H_0\ \mu_A \leq \mu_B$$

$$H_0\ \mu_{A>}\ \mu_B$$

En los contrastes unilaterales el error cae todo en una única región de rechazo de la hipótesis nula:



En Jamovi podemos especificar si nuestro análisis es bilateral o unilateral en la opción de “Hipótesis”:



Estadístico de contraste

El estadístico de contraste indica la discrepancia entre los datos empíricos y la hipótesis nula. Nos da la información de cuánto se aleja el estadístico (muestra) del parámetro (población). Según el contraste de hipótesis que planteemos el estadístico de contraste será diferente, a continuación, se presenta una tabla con los estadísticos de contraste para las pruebas de una media y de contraste de dos medias independientes:

Prueba	Estadístico de contraste
Test para contrastar una media	$z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ varianza poblacional conocida
	$z=\frac{\bar{X}-\mu}{S/\sqrt{n}}$ varianza poblacional desconocida y n >100
	$t=\frac{\bar{x}-\mu_0}{(s/\sqrt{n})}$ varianza poblacional desconocida y muestra pequeña
Test para contrastar medias de dos poblaciones independientes	$z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_{10}-\mu_{20})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ varianza conocida
	$t=\frac{(\bar{x}_1-\bar{x}_2)-d_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2},$ iguales varianzas, varianzas desconocidas
	$t=\frac{(\bar{x}_1-\bar{x}_2)-d_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ varianzas desiguales y desconocidas

Jamovi, en cada análisis nos ofrece el estadístico correspondiente. Para el ejemplo anterior en el que queríamos comprobar si las medias de depresión eran iguales o diferentes en función de dos tratamientos (A y B) el estadístico sería la T de Student:

Prueba T para Muestras Independientes

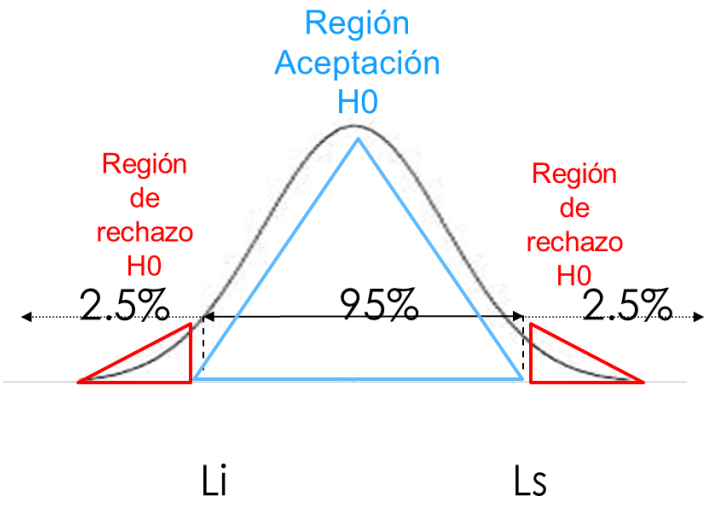
		Estadístico	gl	p	Diferencia de medias	EE de la diferencia
Depresión	T de Student	−3.26 [a]	273.00	0.001	−0.31	0.10

Regla de decisión y significación estadística

La regla de decisión es el criterio para decidir si la hipótesis nula (H0) se mantiene o se rechaza. Para aplicar la regla de decisión podemos dividir la distribución muestral en dos partes:

Zona de aceptación de la H0: donde se encuentra el rango de valores más probables si H0 es cierta, definida por el nivel de confianza.

Zona de rechazo de la H0: donde se encuentra el rango de valores menos probables si H0 es cierta, definida por el nivel de riesgo (α).



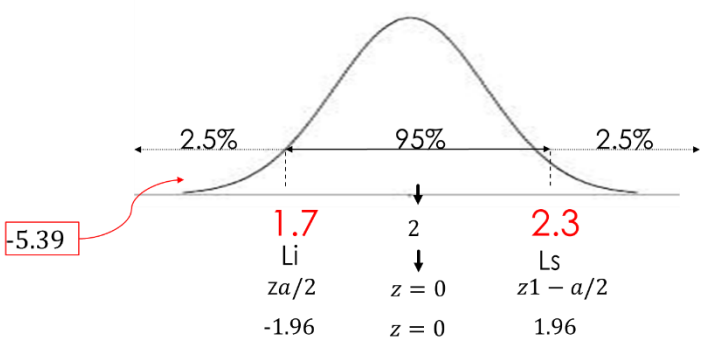
Pongamos un ejemplo sencillo, una prueba para una media. Supongamos que queremos probar si la población de personas mayores de Valencia tiene la misma percepción de su salud que la población de personas mayores españolas.

$$H_0: \mu_v = \mu_e$$

$$H_1: \mu_v \neq \mu_e$$

Sabemos que la media poblacional de los españoles es 2 con límites superior e inferior para un 95% de confianza de 1.7 y 2.3. Para comparar con estos datos poblacionales tenemos una muestra de personas mayores de Valencia de n=500, con media 1.88 y desviación típica 0.46. En nuestro caso la desviación poblacional es desconocida y la muestra grande, por lo que trabajaremos con la siguiente fórmula para calcular el estadístico de contraste:

$$z = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{1.88-2}{0.02} = -5.39$$



Al poner el estadístico de contraste calculado con los datos de la muestra, vemos que caería en a la región de rechazo de la H0. La decisión que deberíamos tomar será rechazar la hipótesis nula.

Aunque la visualización de si el estadístico de contraste cae dentro o fuera de la región de aceptación es una forma útil de saber si debemos mantener o rechazar la H0, esta decisión se puede tomar mucho más rápido si comparamos el nivel crítico (p) con la significación estadística del análisis.

El **nivel de significación (α),** es lo mismo que el nivel de riesgo o el error. Define la probabilidad de rechazar la hipótesis nula.

El **nivel crítico (p)** Es la probabilidad asociada a una región crítica limitada por el valor observado del estadístico, de suponer que H0 es cierta. Un nivel crítico muy pequeño se interpreta como una prueba muy significativa a favor de la hipótesis alternativa.

Al comparar el nivel crítico (p) con el nivel de riesgo (o nivel de significación) si:

p > α mantendremos la H0

p< α rechazaremos la H0

Cuanto más pequeño sea el p-valor mayor es la evidencia en contra de H0. Como generalmente el nivel de significación es un 5%, α=0.05, si **p < 0.05 diremos que hay significación estadística** en el análisis.

Prueba T para Muestras Independientes

		Estadístico	gl	p	Diferencia de medias	EE de la diferencia
Depresión	T de Student	−3.26	273.00	0.001	−0.31	0.10

En el ejemplo de arriba, concluiríamos que sí hay diferencias estadísticamente significativas en las medias.

Tamaño del efecto

El **tamaño del efecto** es una medida estandarizada que se utiliza para indicar la magnitud del efecto estadístico. Cuando se reportan resultados estadísticos, además de decir si son estadísticamente significativos y ofrecer el nivel crítico se recomienda acompañarlo con el tamaño del efecto. El problema del nivel crítico (p) es que depende del tamaño de la muestra. En muestras muy grandes, pequeñas diferencias pueden ser estadísticamente significativas, lo que implicaría que estamos rechazando H0 pero encontrando unas diferencias o una relación irrelevante en la realidad. Si las muestras son muy pequeñas, puede que grandes diferencias no lleguen a ser estadísticamente significativas. Por ello, es necesario distinguir entre la significación estadística y la significación práctica.

Para tomar decisiones reales, no nos interesará sólo mostrar que hay diferencias estadísticamente significativas, sino que estas pueden ser relevantes en la práctica. Por ejemplo, puede que sólo nos interese cambiar un tratamiento por otro nuevo en caso de que la mejora vaya a ser grande. Por ello, se recomienda analizar el **tamaño del efecto**.

Cada prueba tiene un tamaño del efecto medido de una forma distinta y se aplicarán criterios diferentes para saber si el tamaño del efecto es pequeño, mediano o grande.

Por ejemplo, para las pruebas de contraste de medias el estadístico del tamaño del efecto es **d de Cohen**, cuyos estándares de interpretación son:

- d entre 0 y 0.2 pequeño
- d entre 0.2 y 0.5 mediano
- d entre 0.5 y 0.8 grande

Para pruebas ANOVA de comparación de más de dos medias el tamaño del efecto se expresa con **eta parcial al cuadrado (η2)** donde consideraríamos:

- η2 alrededor de 0.01, pequeño
- η2 alrededor de 0.06, medio
- η2 de 0.14 o mayor, grande

Tipos de errores estadísticos

Como hemos visto, lo primero en el contraste de hipótesis sería establecer una H0. Esta hipótesis se se pone a prueba contra nuestros datos (se confronta) y se decide si tiene suficiente apoyo probabilístico o la rechazamos. No obstante, rechazar una hipótesis nula no prueba que sea falsa, sólo significa que no hay evidencia empírica para rechazarla. El contraste entre la realidad y lo que decidimos en nuestra investigación nos lleva a cuatro situaciones posibles:

			Realidad	
			Ho	
			Verdadera	Falsa (H1 verdadera)
Investigación	Ho	Aceptar	Decisión correcta **1 - α**	Decisión incorrecta Error de tipo II **β**
		Rechazar (Aceptar H1)	Decisión incorrecta Error de tipo I **α**	Decisión correcta **1 - β**

En cuanto a las decisiones incorrectas:

- El Error de tipo I, (α o nivel de riesgo) es la probabilidad de rechazar una hipótesis nula verdadera (detectar un efecto que NO existe).
- El Error tipo II (β) es la probabilidad de mantener una hipótesis nula falsa (NO detectar un efecto que SI existe).

En lo que respecta a las decisiones correctas:

- Nivel de confianza (1-α), es la probabilidad de mantener una hipótesis nula verdadera, no detectar un efecto cuando no existe.
- La Potencia (1-β) es la probabilidad de rechazar una hipótesis nula falsa, es decir detectar un efecto cuando existe.

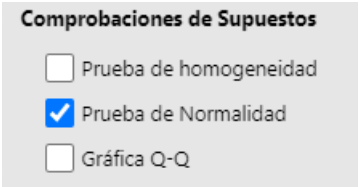
4. Pruebas de normalidad

Las pruebas de normalidad son métodos estadísticos utilizados para evaluar si un conjunto de datos sigue una distribución normal. El contraste de hipótesis que realiza esta prueba es el siguiente:

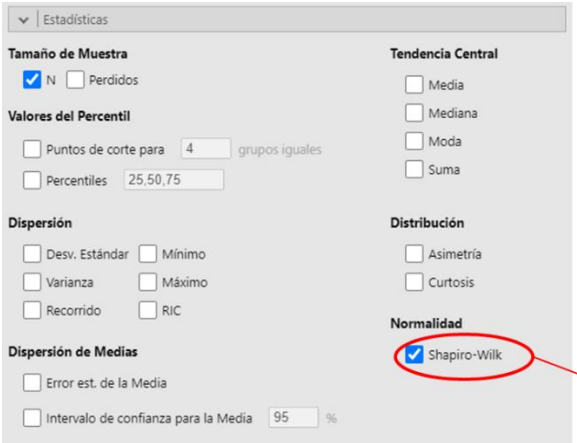
H0: la distribución se distribuye según la Normal

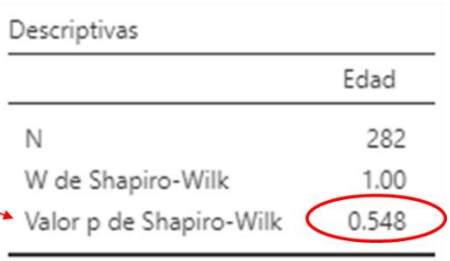
H1: la distribución NO se distribuye según la Normal

Existen diversas pruebas de normalidad, y cada una utiliza diferentes enfoques para evaluar si los datos provienen de una distribución normal. Muchos de los análisis que veremos a continuación se denominan pruebas paramétricas y requieren que las distribuciones de las muestras sean normales. Algunos análisis como las pruebas T, incluyen un apartado de comprobaciones de supuestos para realizar pruebas de normalidad:



En Jamovi también podemos realizar pruebas de normalidad de cualquier distribución en la opción “Análisis”, “Exploración”, “Descriptivas” y “Normalidad”:





En el ejemplo, como el valor de p=0.548, mantendríamos la H0 y podríamos considerar que la edad se distribuye según la distribución normal.

Referencias

The jamovi project (2022). jamovi. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2022-01-01).

TEMA 4

Asociación

Adrián García Mollá
Zaira Torres Romero
Sara Martínez Gregorio

Tema 4. Asociación

1.La inferencia estadística

A fin de introducir el tema sobre la asociación de variables, comenzaremos repasando el concepto de inferencia tratado en anteriores capítulos. Se trata de un tipo de razonamiento que procede de lo particular a lo general, así pues, permite extraer conclusiones sobre el comportamiento de una población en base a la información extraída de una muestra contenida en dicha población. El uso de herramientas o técnicas inferenciales permite establecer comparaciones y las relaciones entre variables, siempre teniendo presente la existencia de determinado riesgo de error establecido en términos de probabilidad.

En la inferencia estadística podemos hacer uso de dos estrategias diferenciadas: la **estimación de parámetros** y el **contraste de hipótesis**. Se trata de dos procedimientos equivalentes, puesto que permiten alcanzar las mismas conclusiones. Sin embargo, la información que resulta de ambos métodos no es exactamente igual.

Estimación de parámetros

Por medio de este procedimiento somos capaces de inferir valores poblacionales a partir de la información extraída de una muestra. Dichos valores poblacionales son los llamados parámetros, mientras que los valores extraídos de la muestra son los estimadores. Habitualmente, encontramos los estimadores denotados matemáticamente con letras del alfabeto latino, mientras que los parámetros siempre se denotan utilizando letras del alfabeto griego.

Podemos llevar a cabo la estimación de parámetros de varias maneras. La primera de ellas es la llamada **estimación puntual**, por medio de la cual atribuimos el valor muestral al parámetro de interés. Por ejemplo, en el supuesto de que queramos conocer la media de edad de un grupo, podríamos averiguar la media de edad de una muestra extraída del grupo y asignarla al conjunto poblacional. Este método entraña graves problemas dado que no es posible conocer el *error muestral* (*E*), es decir, la distancia entre el parámetro y el estimador:

$$E = |\hat{\theta} - \theta|$$

La segunda manera que trataremos es la **estimación por intervalos.** En este caso, en lugar de asignar un valor único al parámetro, se establece un rango de valores en el que es altamente probable que se encuentre el valor del parámetro (esta probabilidad siempre es conocida). Así pues, llamamos *Intervalo de Confianza (IC)* al rango de valores dentro del cual se encuentra el parámetro. Por otra parte, encontramos los *límites de confianza*, que son los valores máximo y mínimo que delimitan el *IC*. Estos se obtienen sumando o restando el Error Muestral al estimador puntual según hablemos del límite superior o del límite inferior.

$$L_S = \hat{\theta} + E$$

$$L_i = \hat{\theta} - E$$

Contraste de hipótesis

Se trata de un método útil a la hora de tomar decisiones acerca de si una proposición sobre una determinada población debe mantenerse o rechazarse. Dicha proposición o afirmación se denomina hipótesis científica y debe ser operativizada de manera que sea contrastable por medio de una evidencia empírica.

En los contrastes de hipótesis habitualmente formulamos dos hipótesis:

- La hipótesis **nula**, H_0, que es la que someteremos a juicio. Plantea una afirmación concreta sobre la distribución de probabilidad o el valor de alguno de los parámetros.
- La hipótesis **alternativa**, H_1, es la negación de la hipótesis nula e incluye todas las posibles afirmaciones que la H_0 no incluye.

En definitiva, ambas hipótesis se plantean enfrentadas. Por lo tanto, son mutuamente excluyentes y exhaustivas.

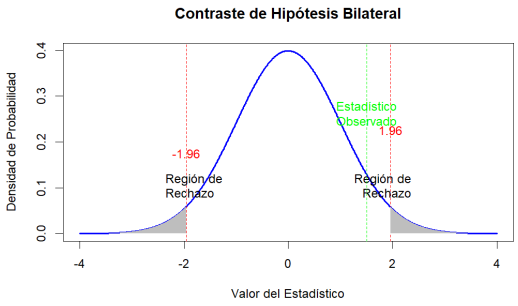
En los contrastes de hipótesis es necesario que la distribución muestral del estadístico de contraste se encuentre especificada. En caso contrario, nos resultará imposible tomar una decisión sobre la retención o el rechazo de la hipótesis nula.

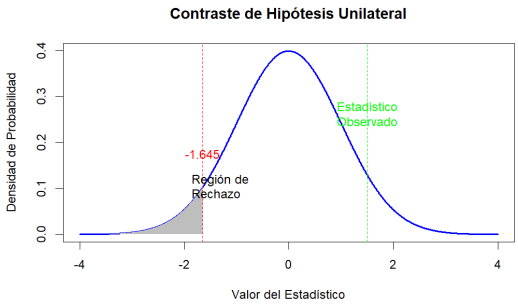
Para la toma de decisiones sobre la hipótesis nula resulta indispensable establecer un criterio fundamentado en la compatibilidad de la hipótesis nula con los datos. La determinación de esta compatibilidad se lleva a cabo estableciendo dos áreas

excluyentes y exhaustivas dentro de la distribución de probabilidad del estadístico de contraste: la zona de rechazo y la zona de aceptación.

- **Zona de rechazo**: se trata del área de la distribución muestral donde se ubican los valores del estadístico de contraste alejados de lo establecido en la H_0. La probabilidad asociada a esta área es conocida como nivel de significación y se denota matemáticamente con la letra α (habitualmente de .05).
- **Zona de aceptación:** se corresponde con el área de la distribución de probabilidad donde se sitúan los valores del estadístico que son compatibles con la H_0. La probabilidad asociada esta zona es el complementario de α (1- α).

A continuación, se presentan dos distribuciones de probabilidad correspondientes a los contrastes de hipótesis bilateral y unilateral, donde el área sombreada en gris se corresponde con las zonas de rechazo (α).





2. Medidas de asociación

El objetivo de las pruebas de asociación es estudiar la relación entre dos variables. Este análisis se basa en el cálculo de uno o más coeficientes que describen el nivel de asociación o correlación entre las variables. Asociación y correlación son sinónimos, pero se tiende a utilizar uno u otro según la escala de medida de las variables que se analizan, para variables cuantitativas o semicuantitativas (ordinales), se llaman medidas de correlación y para variables cualitativas medidas de asociación.

La relación entre variables presenta dos características: sentido e intensidad/magnitud.

- **Sentido**: Diremos que una relación entre variables es positiva (o directa) siempre que los valores superiores en una variable tienden a coincidir con los más altos de la otra variable. Diremos que es se negativa (o inversa) siempre que los valores superiores en una variable tiendan a coincidir con los inferiores de la otra variable.
- **Intensidad o magnitud**: es el grado en que las posiciones relativas de los datos son coincidentes (ya sea en sentido positivo o inverso).

Para conocer en profundidad los diferentes tipos de medidas de asociación, incidiremos sobre ellas en función de la naturaleza de las variables sobre las que estemos trabajando.

Variables cualitativas

Se trata de variables que reportan información nominal a cerca de la muestra u ordinal con pocos valores. Es habitual encontrarlas en ciencias de la salud y ciencias sociales, y permiten la clasificación de los individuos en categorías.

Es posible observar la asociación de dos o más variables categóricas haciendo uso de **tablas de contingencia,** aunque en este caso nos ceñiremos a los casos de dos variables. Para ello, debemos representar las frecuencias de las categorías de ambas variables y observar si existe una asociación entre ellas o no, es decir, debemos constatar que no haya diferencias entre las frecuencias de las categorías de una de las variables al alternar las categorías de la otra variable. De esta manera únicamente podremos observar las diferencias en la distribución de frecuencias en las diferentes categorías dentro de la muestra (**interpretación descriptiva**), sin embargo, no

podemos constatar si se trata de una asociación entre variables en la población o no (**inferencia**).

Por esta razón, para llevar a cabo el proceso de inferencia sobre el supuesto de independencia de las variables de estudio haremos uso de los estadísticos **ji-cuadrado y V de Cramer.**

El nivel crítico o significación (p) de la prueba ji-cuadrado nos indica si hay una asociación significativa entre las variables. En este caso, se plantean las siguientes hipótesis para el contraste:

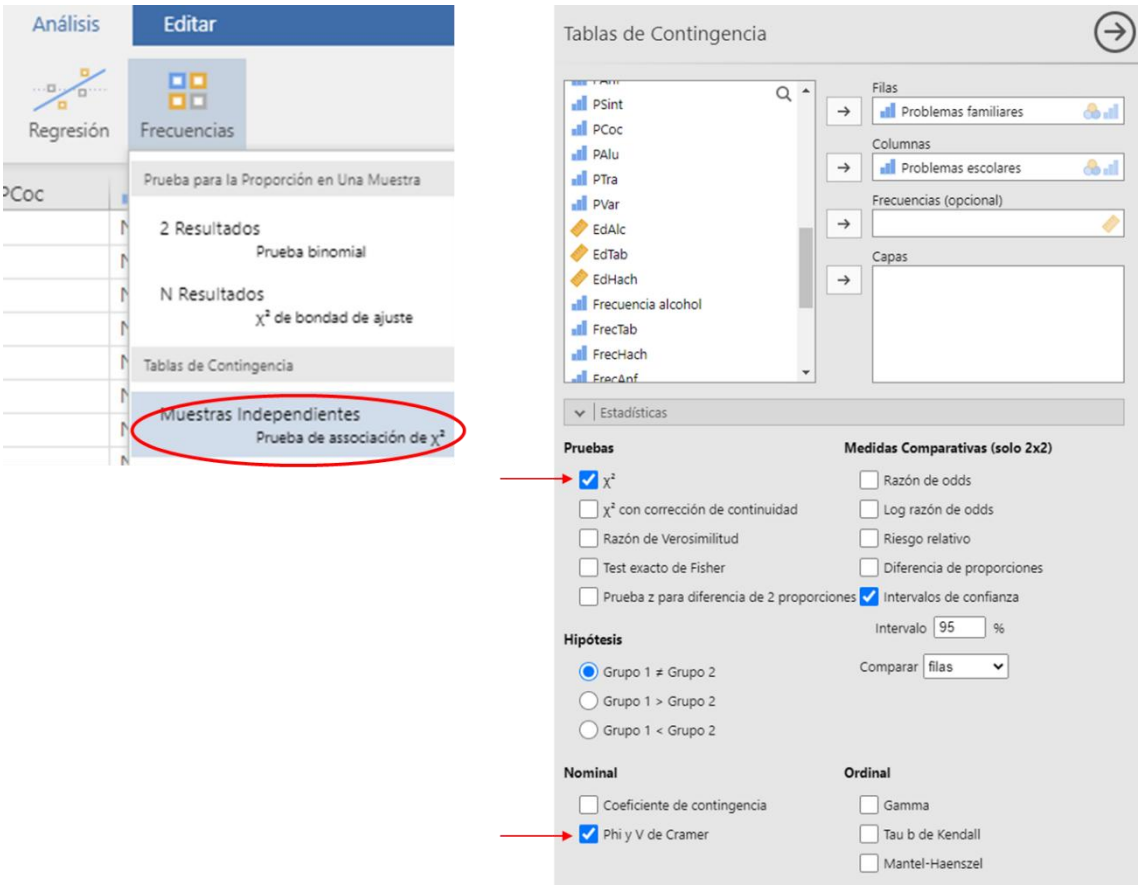
- H_0: No existe asociación, es decir, las variables son independientes.
- H_1: Existe una asociación entre las variables, es decir, no son independientes.

Si el valor de p es inferior a 0.05, rechazamos la H0 y podemos afirmar que existe una asociación entre las variables. Por el contrario, si p toma un valor mayor a 0.05, diremos que se retiene la H0 y que existe independencia entre las variables.

Para su interpretación, es relevante conocer previamente que:

- El estadístico ji-cuadrado tiene un rango de 0 hasta infinito.
- La V de Cramer es una medida del tamaño del efecto para la prueba ji-cuadrado. Sólo puede adoptar valores positivos, 0 indica ausencia de relación y 1 indica una relación perfecta.

Pasamos ahora a ver un ejemplo de estos análisis utilizando el paquete Jamovi. La ruta a seguir sería la siguiente: “Análisis”, “Frecuencias” y “Pruebas de asociación de X²”. En este caso estudiaremos la asociación entre los problemas familiares y los problemas escolares.



Los resultados obtenidos se contemplan en la siguiente imagen:

Pruebas de χ^2

	Valor	gl	p
χ^2	8.23	1	0.004
N	298		

Nominal

	Valor
Coeficiente Phi	0.17
V de Cramer	0.17

Como se puede observar, hay una asociación significativa entre los problemas familiares (medido como sí o no) y entre los problemas escolares (medidos como sí o no) $p<0.05$ y vemos la V de Cramer tiene un valor de 0.17.

Con respecto de las tablas de contingencia, podemos pararnos a observar qué categorías presentan mayor frecuencia de aparición en nuestra muestra.

Tablas de Contingencia

Tablas de Contingencia

	Problemas escolares		
Problemas familiares	No	Sí	Total
No	6	11	17
Sí	32	249	281
Total	38	260	298

Variables semicuantitativas

Son un tipo de variable que combina características de variables cuantitativas y cualitativas. En este tipo de variable, los valores se pueden ordenar o clasificar, pero no se pueden asignar de manera precisa o absoluta a una escala numérica. En otras palabras, los datos se representan mediante categorías o niveles que tienen un orden relativo, pero no necesariamente una distancia o magnitud numérica definida entre ellos.

Por ejemplo, en un estudio médico, una variable semicuantitativa podría ser la evaluación de la gravedad de una enfermedad en una escala de "leve", "moderada" o "grave". Aunque estas categorías tienen un orden relativo, no representan valores numéricos absolutos. Otro ejemplo habitual podría ser una escala de dolor, donde los pacientes clasifican su nivel de dolor como "ninguno", "leve", "moderado" o "severo".

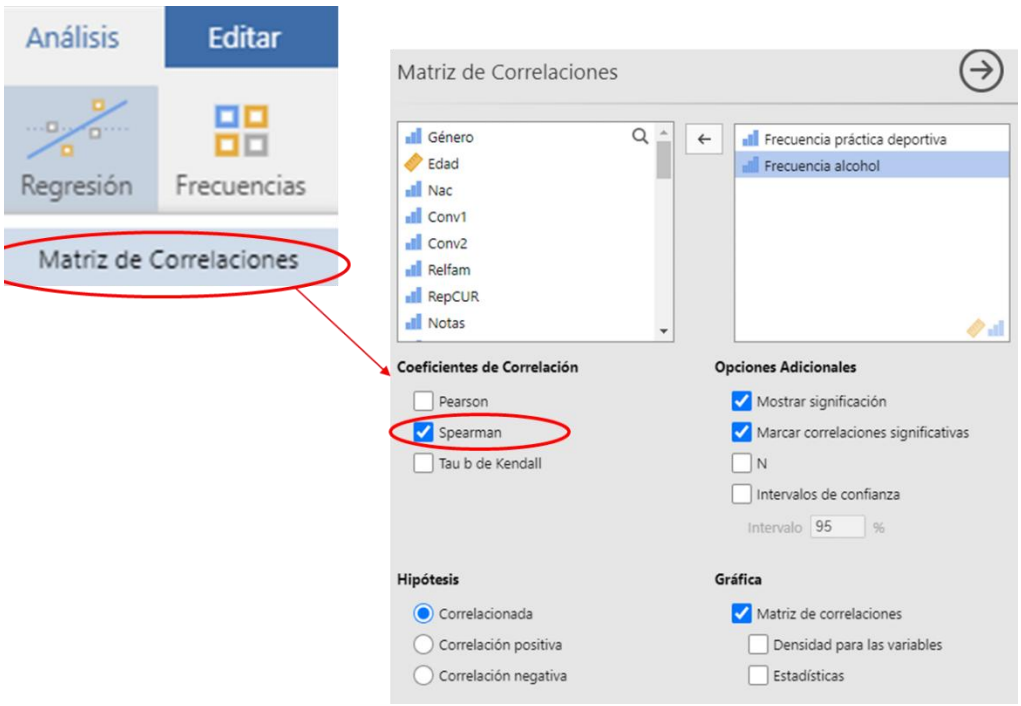
Las variables semicuantitativas son útiles cuando la medición exacta no es posible o práctica, pero se requiere una evaluación relativa o comparativa de los datos. Sin embargo, es importante tener en cuenta que la interpretación de los resultados puede ser subjetiva y depende en gran medida del contexto específico en el que se utilice la variable semicuantitativa.

Para estudiar la asociación de dos variables semicuantitativas, utilizaremos la correlación de **Spearman (rho)**. La ruta en Jamovi para llevar a cabo el análisis es: “Análisis”, “Regresión” y “Matriz de correlaciones”. En el apartado de coeficientes de correlación, debemos seleccionar Spearman.

La interpretación de los resultados se daría de manera similar a cualquier contraste, si el valor de p es inferior a 0.05, rechazamos la H_0 y podemos afirmar que existe una asociación entre las variables. Por el contrario, si p toma un valor mayor a 0.05, diremos que se retiene la H_0 y que existe independencia entre las variables.

Además, debemos interpretar el coeficiente de correlación rho que es un valor acotado entre -1 y 1, donde:

- Rho= -1 indica una relación lineal perfecta negativa/inversa
- Rho= 0 indica ausencia de relación lineal
- Rho= +1 indica una relación lineal positiva/directa perfecta



Como vemos en el ejemplo, no habría una correlación estadísticamente significativa entre la frecuencia de práctica deportiva y frecuencia de consumo de alcohol p> 0.05 y el valor del coeficiente de Spearman es muy bajo rho= 0.06.

Matriz de Correlaciones

		Frecuencia práctica deportiva	Frecuencia alcohol
Frecuencia práctica deportiva	Rho de Spearman	—	
	valor p	—	
Frecuencia alcohol	Rho de Spearman	0.06	—
	valor p	0.289	—

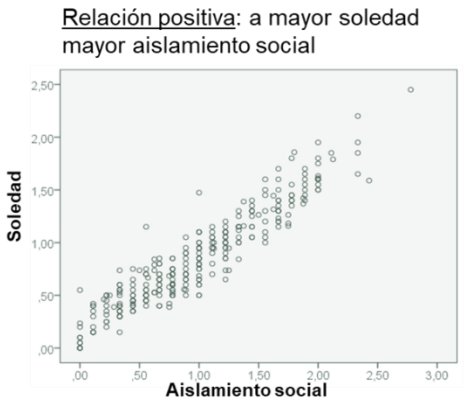
Nota. * p < .05, ** p < .01, *** p < .001

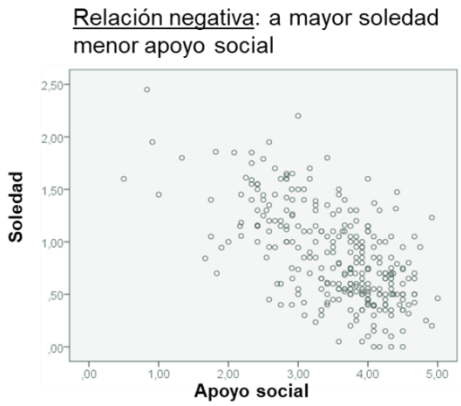
Variables cuantitativas

Se trata de variables que toman valores numéricos y se puede medir o contar. Representan cantidades numéricas que pueden ser discretas o continuas, es decir, pueden tomar valores enteros o bien pueden tomar cualquier valor dentro de un rango.

Para este tipo de variables se recomienda utilizar el coeficiente de **correlación de Pearson (r),** mide el grado de relación lineal entre dos variables.

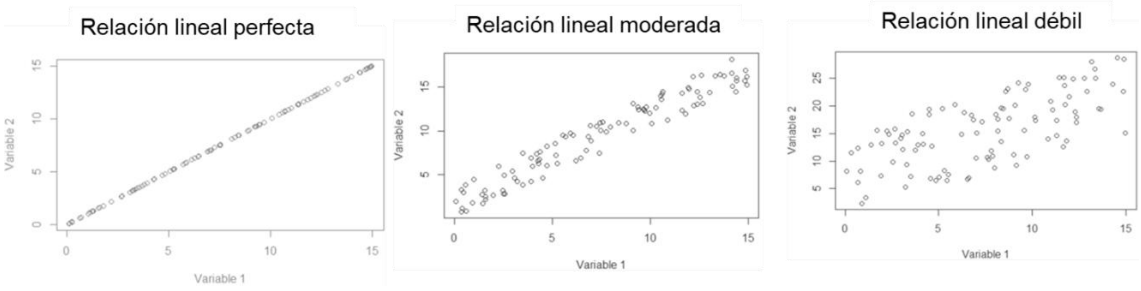
De nuevo, igual que el coeficiente de correlación de Spearman (rho), el coeficiente de correlación de Pearson (r) tiene un rango desde -1 hasta 1. De tal forma que, los valores próximos al 0 indican un grado de asociación bajo, mientras que los valores próximos a -1 y 1 indican una asociación alta (inversa o directa, respectivamente). Veamos un ejemplo.



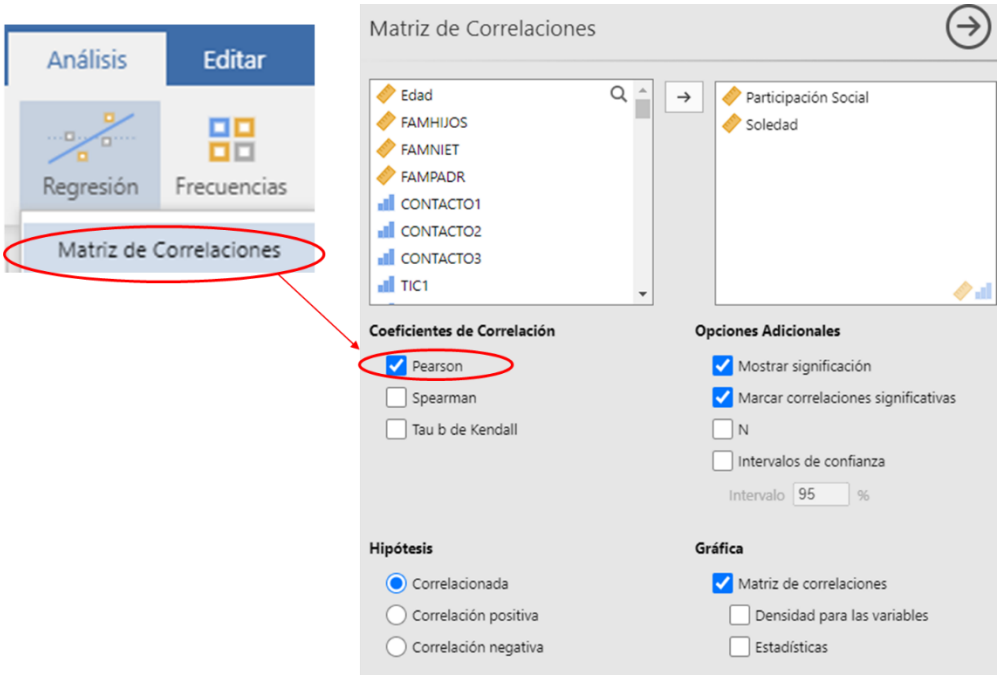


Según la interpretación de Cohen (1992) si:

- r= 0.10 a 0.30 (-0.10 a -0.30) la relación es pequeña.
- r= 0.30 a 0.50 (-0.30 a -0.50) la relación es media.
- r= 0.50 a 1.0 (-0.50 a -1.0) la relación es fuerte.



En Jamovi se puede llevar a cabo el análisis de correlación de Pearson en la opción “Análisis”, “Regresión” y “Matriz de correlaciones”. Tal y como se ve a continuación:

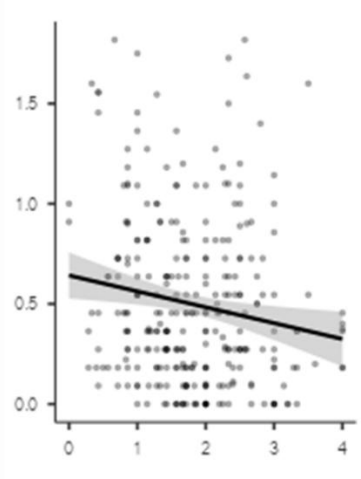


Además, si marcamos la opción “Marcar correlaciones significativas” y en gráfica “Matriz de correlaciones” aparecerá en la tabla de correlaciones marcados los análisis que sean estadísticamente significativos y la gráfica donde se ve la relación lineal entre variables.

Matriz de Correlaciones

		Participación Social	Soledad
Participación Social	R de Pearson	—	
	valor p	—	
Soledad	R de Pearson	−0.15 **	—
	valor p	0.008	—

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



En el ejemplo, podemos ver que las variables soledad y participación social mantienen una relación estadísticamente significativa p<0.05, negativa r= -0.15, y que se podía considerar pequeña.

Referencias

Cohen, J. (1992). Statistical Power analysis. *Current Directions in Psychological Science, 1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Díaz, M. Á. R., Merino, A. P., Ruiz, M. Á., Martín, R. S., & Castellanos, R. S. M. (2015). *Análisis de datos en ciencias sociales y de la salud*. Síntesis.

The jamovi project (2022). jamovi. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2022-01-01).

TEMA 5

Comparación de medias

Adrián García Mollá
Zaira Torres Romero
Sara Martínez Gregorio

Tema 5. Comparación de medidas

1.Introducción

En muchas ocasiones al llevar a cabo una investigación estamos interesados en estudiar la relación entre una variable de tipo categórico y una variable de tipo continuo. Generalmente, la variable de tipo categórico diferencia entre varios grupos entre los que queremos comparar los resultados en una determinada variable continua. Por ejemplo, podríamos estar interesados en conocer si existen diferencias en la satisfacción vital de las personas mayores en función de su viven en el medio rural o en el medio urbano. En este ejemplo tenemos dos variables implicadas: el tipo de medio de residencia (variable independiente categórica) y el nivel de satisfacción con la vida (variable dependiente continua).

Ahora bien, esta variable de tipo categórico no siempre hace referencia a diferentes grupos. En ocasiones, hace referencia a diferentes mediciones de un mismo grupo. Por ejemplo, imaginemos que estamos aplicando una intervención para la reducción del consumo de tabaco en fumadores. Hemos evaluado el número de cigarrillos que consumen al día antes de empezar con la intervención y una vez acabada la intervención. En este caso, podríamos entender que tenemos una variable dependiente continua (el número de cigarrillos consumidos al día) y una variable categórica independiente (el tiempo).

Ambos ejemplos se podrían resolver comparando de manera estadística la puntuación promedio de la variable continua en los diferentes grupos de la variable categórica. Es decir, podríamos obtener la media de la satisfacción vital de las personas mayores que viven en el medio rural y compararla frente a la media de la satisfacción vital de las personas mayores que viven en el medio urbano. De la misma manera, podríamos obtener el consumo de cigarrillos diario promedio antes de empezar la intervención y compararlo con el consumo diario promedio después de la intervención.

Es decir, la comparación de medias hace referencia a la evaluación de las diferencias entre los promedios de dos o más grupos. Como se ha comentado en temas anteriores, diferencias observadas a nivel descriptivo entre dos muestras pueden deberse al error muestral y no reflejar verdaderamente diferencias existentes en sus poblaciones de referencia. Por ello, nuevamente, debemos dar el salto a la estadística inferencial para poder responder a nuestras preguntas de investigación con un determinado nivel de confianza.

Existen diversas pruebas estadísticas diseñadas para llevar a cabo la comparación de medias, las más usadas son las pruebas T (en caso de querer comparar dos grupos) y las pruebas ANOVA (en caso de comprar más de dos grupos).

Las pruebas T y pruebas ANOVA se basan en las distribuciones normales de la variable estudiada para conocer la probabilidad de que haya diferencias o no entre los grupos comparados. Lo hacen mediante dos parámetros: la media y la desviación estándar. Por eso para poder aplicarlas correctamente se deben de cumplir una serie de supuestos:

- **Observaciones independientes**: Los errores que se presenten deben de ser independientes. Se cumple si se emplea muestreo aleatorio
- **Normalidad**: El análisis y observaciones que se obtienen de las muestras deben considerarse normales. Comprobación mediante gráficos o Pruebas de Normalidad.
- **Homocedasticidad** (Igualdad de varianzas): Los grupos deben presentar variables uniformes, es decir, que sean homogéneas. En el caso del ANOVA de medidas repetidas este supuesto es de **Esfericidad** donde se comprueba si las variaciones de las diferencias entre todos los pares posibles de condiciones intra-sujeto (es decir, niveles de la variable independiente) son iguales.

En general las pruebas T y ANOVA son bastante robustas al incumplimiento de los supuestos, sobre todo con muestras grandes. Además, incluyen correcciones en caso de incumplimiento del supuesto de varianzas homogéneas o de esfericidad. No obstante, cuando estas correcciones no son suficiente, o en caso de incumplir normalidad en muestras pequeñas, se pueden emplear pruebas NO paramétricas.

Para cada prueba paramétrica existe un equivalente no paramétrico:

Prueba t de muestras independientes	→	U de Mann–Whitney
Prueba t de muestras dependientes	→	Prueba de Wilcoxon
ANOVA entre-sujetos	→	H de Kruskal-Wallis
ANOVA de medidas repetidas	→	Prueba de Friedman

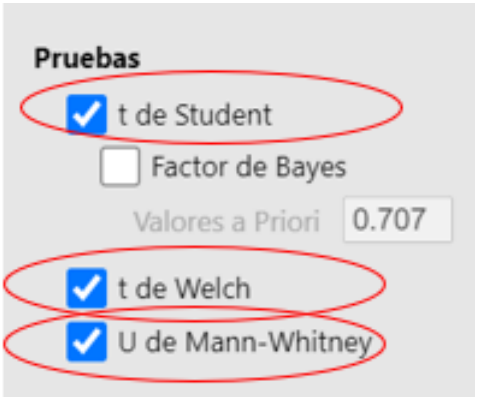
2. Comparación de dos medias

En el caso de dos muestras a comparar, es importante que primero distingamos si las muestras son independientes o dependientes. Diremos que dos grupos o muestras son independientes si los valores de una muestra no revelan información sobre los valores de la otra. Por ejemplo, si queremos comparar los niveles de colesterol entre dos grupos de personas, unas tratadas con una nueva medicación y otras con el tratamiento estándar, las puntuaciones de un grupo no influyen en las del otro y aplicaremos pruebas independientes.

2.1. Muestras independientes

Dentro de las comparaciones de dos medias independientes, encontramos la **Prueba T de Student** y la prueba **U de Mann-Whitney.**

Para elegir la mejor prueba para nuestros datos, comprobaremos dos supuestos, el de normalidad y el de homogeneidad de varianzas. Marcaremos la opción "**t de Student"** cuando las muestras sean aproximadamente normales y las varianzas sean homogéneas. En caso de que no se cumpla el supuesto de homogeneidad, tenemos la alternativa de escoger la **t de Welch**, que aplica una corrección para la falta de homogeneidad de las varianzas. Si los datos no cumplen con los supuestos de normalidad marcaremos la opción en pruebas "**U de Mann-Whitney"**.

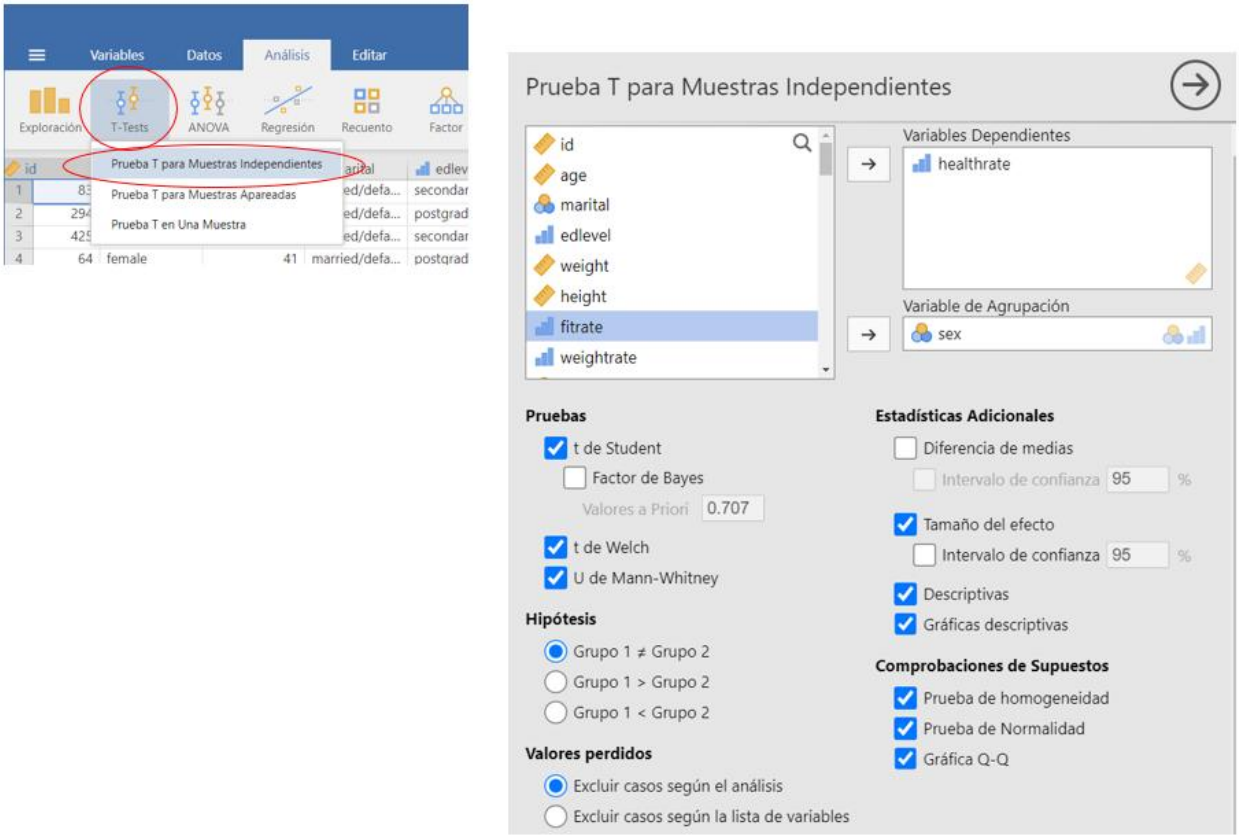


Pongamos que queremos comprobar en una muestra de personas mayores si hay diferencias en las medias de la salud autopercibida entre hombres y mujeres. Las hipótesis a contrastar son:

H0: μ1 = μ2 las medias son iguales

H1: μ 1 ≠ μ2 las medias son diferentes

En Jamovi, podremos lanzar el análisis de comparación de dos medias independientes en “Análisis”, “Pruebas T” y “Pruebas T para Muestras Independientes”:



Podemos dejar la opción para contraste bilateral que aparece por defecto en el apartado de “Hipótesis”, o si quisiéramos poner a prueba que un grupo tiene una media mayor o menor que el otro podemos cambiar la opción a alguna de las otras disponibles.

En el apartado “Comprobación de supuestos” marcaremos pruebas de normalidad y de homogeneidad para ver si se cumplen estos supuestos o no. En este ejemplo, los análisis nos muestran que se incumple el supuesto de normalidad de las muestras y sabemos que deberíamos emplear la prueba U de Mann-Whitney. Esto lo podemos comprobar observando el resultado del valor de p en la Prueba de Normal de Shapiro-Wilk, donde podemos comprobar que es inferior a 0.05. Adicionalmente, podemos explorar visualmente el Gráfico Q-Q. En caso de que la variable siguiese una distribución normal los puntos se encontrarían sobre la línea, alejamientos pronunciados indican falta de normalidad, como en este caso.

Supuestos

Prueba de Normalidad (Shapiro-Wilk)

	W	p
healthrate	0.909	< .001

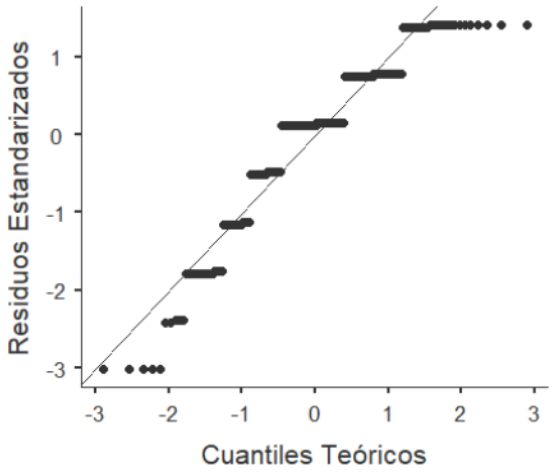
Nota. Un valor p bajo sugiere una violación del supuesto de normalidad

Prueba de Levene para homogeneidad de varianzas

	F	gl	gl2	p
healthrate	1.45	1	265	0.229

Nota. Un valor p bajo sugiere una violación del supuesto de varianzas iguales

[4]



La opción “Pruebas”, nos permite marcar la prueba que consideremos que debemos emplear, en este caso deberíamos interpretar la prueba U de Mann-Whitney. Ambas pruebas nos ofrecen la misma información, p= 0.849, lo que nos indica que debemos mantener la H0 y no hay diferencias estadísticamente significativas entre los grupos.

Al no salir diferencias estadísticamente significativas, a priori, no deberíamos interpretar el tamaño del efecto. Ahora bien, como pauta general recordamos que valores de d de Cohen próximos a 0.20 se consideran pequeños, entorno a 0.50 mediano y próximos a 0.80 grandes.

Prueba T para Muestras Independientes

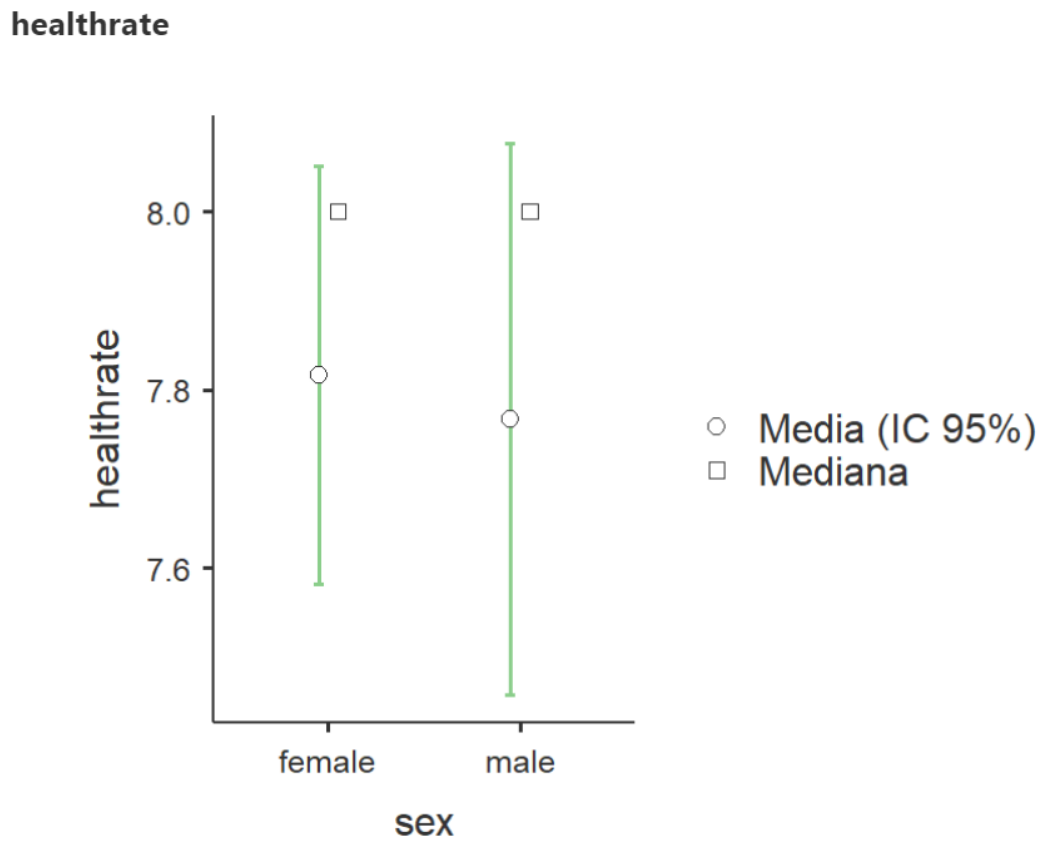
		Estadístico	gl	p		Tamaño del Efecto
healthrate	T de Student	0.255	265	0.799	La d de Cohen	0.0314
	T de Welch	0.251	233	0.802	La d de Cohen	0.0311
	U de Mann-Whitney	8704		0.849	Correlación biseriada de rangos	0.0132

Nota. H_a $\mu_{female} \neq \mu_{male}$

Además, en estadísticas adicionales, podemos marcar las opciones de “tamaño del efecto” para hacernos una idea de cómo de grandes son estas diferencias en la realidad y “descriptivas” dónde podemos ver los valores medios de cada grupo y saber cuál ha presentado una mayor media. Por último, si pedimos gráficas descriptivas, nos aparecerá una representación gráfica de cada media según el grupo.

Descriptivas de Grupo

	Grupo	N	Media	Mediana	DE	EE
healthrate	female	147	7.82	8.00	1.45	0.120
	male	120	7.77	8.00	1.73	0.158



En este ejemplo, vemos que las diferencias entre medias han sido de 0.05 unidades, con un tamaño del efecto de 0.01, donde las medias de las mujeres han sido

más altas que las de los hombres, aunque esta diferencia no es estadísticamente significativa. La manera de reportar esta información adecuadamente en formato APA sería: U = 8704, *p* = .85, *d* = .01.

2.2. Muestras dependientes

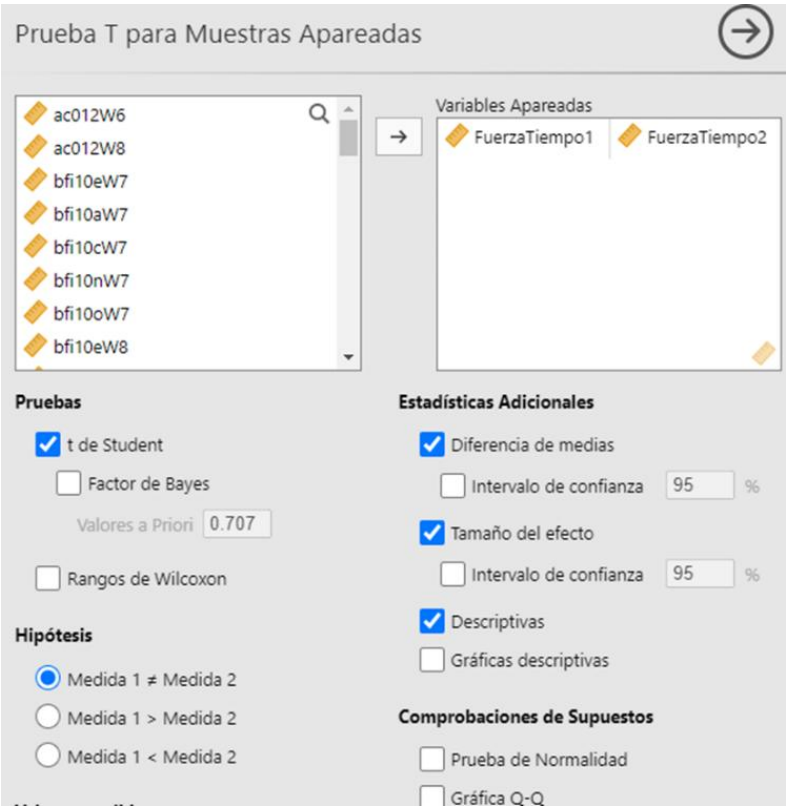
Si los valores de una muestra afectan a los valores de la otra muestra, entonces las muestras son dependientes. Generalmente, siempre que tengamos dos medidas que pertenecen a una misma persona aplicaremos medidas dependientes o relacionadas. Por ejemplo, si queremos analizar si hay una mejora significativa en los niveles de glucosa en sangre después de un programa de dieta y ejercicio y tenemos dos medidas de la glucosa de una misma persona (una antes y otra después del programa) aplicaremos prueba de muestra dependientes.

Para seleccionar la prueba más adecuada para nuestros datos, primero deberemos comprobar el supuesto de normalidad. Utilizaremos la **Prueba t de Student** cuando las muestras sean aproximadamente normales. En caso de trabajar con muestras pequeñas o con datos no normalmente distribuidos utilizaremos la **Prueba de Wilcoxon**.

En Jamovi, podremos lanzar el análisis de comparación de dos medias dependientes en “Análisis”, “Pruebas T” y “Pruebas T para Muestras Apareadas”. Este análisis presenta casi las mismas opciones que el de pruebas independientes. También deberemos en primer lugar comprobar el supuesto de normalidad y en función de si se cumple o no elegiremos la prueba T o la prueba de rangos de Wilcoxon. En el caso de muestras apareadas, solo contamos con un único grupo por lo que el supuesto de homogeneidad no aplica a estas circunstancias.

Pongamos un ejemplo, en el que queremos comprobar si hay cambios en la fuerza de agarre en el mismo grupo de personas pasados dos años. La fuerza de agarre se midió utilizando un dinamómetro manual en cada mano y calculando la fuerza máxima en kg. Ya hemos comprobado previamente que se cumple el supuesto de normalidad y disponemos de una muestra grande de 25719 personas.

En este caso como las variables son dependientes o apareadas, no tenemos una variable de agrupación, sino que tenemos dos variables medidas en distintos momentos temporales que introduciremos en el recuadro de “variables apareadas”:



Prueba T para Muestras Apareadas

Prueba T para Muestras Apareadas

			estadístico	gl	p	Diferencia de medias	EE de la diferencia		Tamaño del Efecto
FuerzaTiempo1	FuerzaTiempo2	T de Student	38.22	25718.00	< .001	1.25	0.03	d de Cohen	0.24

Descriptivas

	N	Media	Mediana	DE	EE
FuerzaTiempo1	25719	33.56	32.00	11.28	0.07
FuerzaTiempo2	25719	32.31	30.00	11.19	0.07

En este ejemplo, vemos que sí hay diferencias estadísticamente significativas entre la fuerza de agarre de las personas en el tiempo 1 y tiempo 2. Estas diferencias son de 1.25 kg y tienen un tamaño del efecto igual a 0.24. Los descriptivos de las medias nos muestran una media más baja para el segundo tiempo, es decir, la gente ha perdido fuerza. En este caso podríamos reportarlo como: t(27718) = 38.2, $p < .001$, $d = .24$.

3. Comparación de más de dos medias

Si queremos comparar medias de más de dos grupos, deberemos emplear pruebas ANOVA o las pruebas correspondientes no paramétricas (Kurskal-Wallis y Friedman). En las pruebas ANOVA las hipótesis a contrastar serán:

H0: $\mu 1 = \mu 2 = \ldots = \mu k$ las medias son iguales

H1: Al menos una de las medias no es igual al resto

Las pruebas ANOVA son lo que se denomina una prueba OMNIBUS. En caso de que la hipótesis nula se mantenga, todas las medias poblacionales se consideran iguales. Pero si la hipótesis nula no se mantiene, sabremos que al menos una media es más alta que las otras. Para saber qué media o que medias son diferentes deberemos hacer pruebas adicionales de comparaciones múltiples (también conocidas como pruebas post-hoc o pruebas a posteriori).

Además, igual que en el caso de las comparaciones entre dos medias, en las comparaciones de más de dos medias también deberemos tener en cuenta el tipo de muestra:

- Independientes (o entre-sujetos)
- Dependientes o medidas repetidas (intra-sujetos)
- Mixto: combinas variables entre e intra

Sabremos si las muestras son independientes o no aplicando la misma lógica que para las pruebas de comparación de dos medias, en caso de que los grupos no estén relacionados diremos que son independientes y en caso de que las puntuaciones de un grupo tengan algún tipo de relación con las de los otros grupos, sabremos que son muestras dependientes o relacionadas.

Además, en las pruebas de comparación de más de dos medias deberemos tener en cuenta el número de factores:

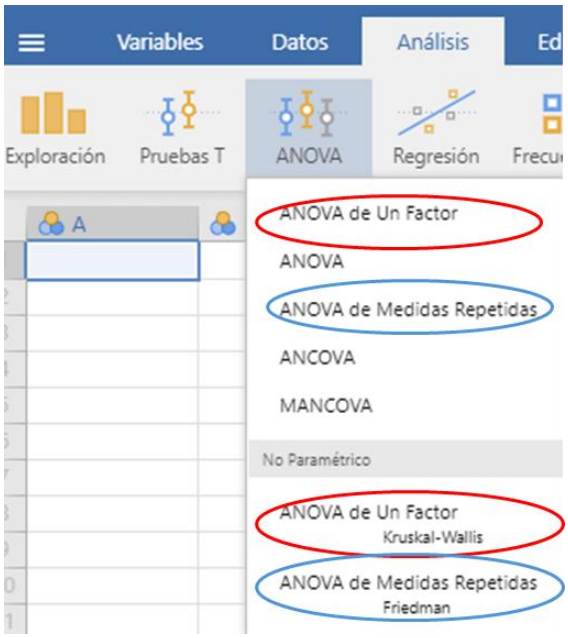
- Un solo factor (unifactorial o de una vía)
- Factorial (2 o más factores)

Factor hace referencia a la variable independiente que genera los grupos. Por ejemplo, si estamos llevando a cabo un estudio para analizar la eficacia de tres tratamientos diferentes (A, B y C) en el control de la presión arterial, y además queremos investigar si hay diferencias significativas en los resultados dependiendo del nivel de actividad física de la persona (baja, moderada, nula). Tendremos un diseño factorial de

dos factores, Factor 1 sería el tipo de tratamiento y el Factor 2 sería el nivel de actividad física.

En este tema nos centraremos en los análisis unifactoriales, donde comparemos las medias de varios grupos que pertenecen a un mismo factor.

En Jamovi, este tipo de pruebas las encontraremos en la opción “Análisis” y “ANOVA”. En rojo, las opciones para muestras independientes y en azul las opciones para muestras dependientes.

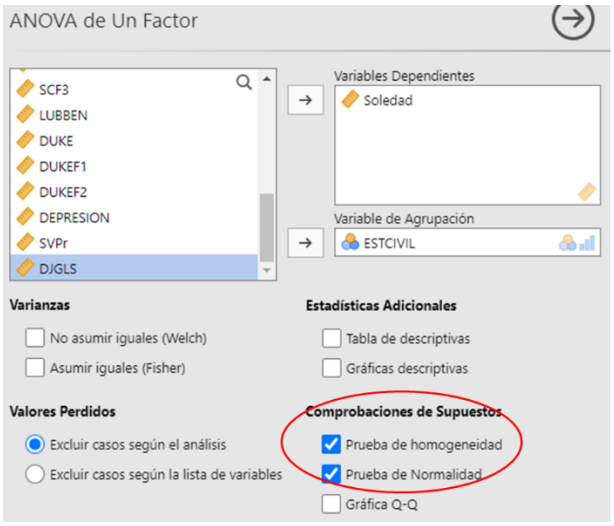


3.1. Muestras independientes

ANOVA de un factor

La prueba **ANOVA de Un Factor** será la recomendada si se cumple el supuesto de normalidad (y elegiremos opciones adicionales dentro de este análisis en función de si se cumple el supuesto de homogeneidad). En caso de que se incumpla el supuesto de normalidad podremos aplicar la prueba no paramétrica equivalente de **Kruskal-Wallis**.

Por ejemplo, imaginemos que queremos comparar el nivel de soledad en una muestra de 250 personas en función de su estado civil (casado, divorciado, soltero o viudo). Un primer paso sería introducir las variables y analizar si se cumplen los supuestos:



Comprobaciones de Supuestos

Prueba de Normalidad (Shapiro-Wilk)

	W	p
Soledad	0.98	0.059

Nota. Un valor p bajo sugiere una violación del supuesto de normalidad

Prueba de Levene para homogeneidad de varianzas

	F	gl1	gl2	p
Soledad	1.85	3	230	0.139

[3]

En este caso como se cumple el supuesto de normalidad no hará falta cambiar a la prueba Kruskal Wallis y como se cumple el supuesto de homogeneidad de varianzas podremos clicar en la opción “Asumir varianzas iguales (Fisher)” para los análisis posteriores.

ANOVA de Un Factor

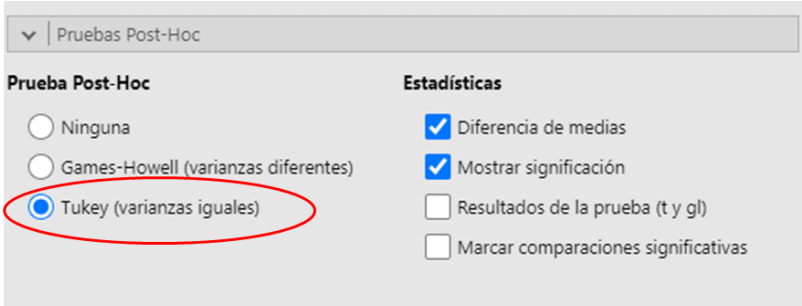
ANOVA de Un Factor (Fisher)

	F	gl1	gl2	p
Soledad	3.24	3	230	0.023

Descriptivas de Grupo

	ESTCIVIL	N	Media	DE	EE
Soledad	Casado	170	0.82	0.43	0.03
	Soltero	21	0.86	0.44	0.10
	Divorciado/separado	20	1.07	0.57	0.13
	Viudo	23	1.05	0.48	0.10

La prueba es estadísticamente significativa y nos indica que sí hay diferencias entre las medias de los grupos y en la Tabla de descriptivas del grupo podemos ver cuál es la media de cada uno para hacernos una idea de qué medias pueden diferir. No obstante, si queremos saber entre qué medias en concreto hay diferencias estadísticamente significativas deberemos analizar las pruebas post hoc. Las pruebas post hoc consisten en comparaciones por pares para comparar todos los grupos. Es casi como coger cada par de grupos y aplicar una prueba t sobre cada par, pero en ANOVA se mantiene la tasa de error tipo I a través de todas las comparaciones.



Pruebas Post Hoc

Tukey Post-Hoc Test – Soledad

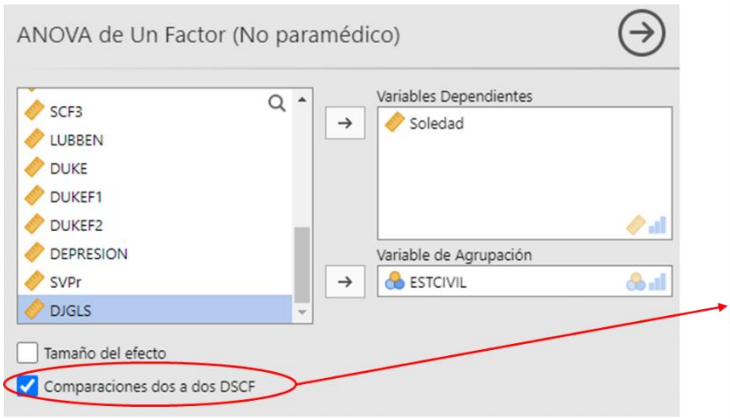
		Casado	Soltero	Divorciado/separado	Viudo
Casado	Diferencia de medias	—	−0.04	−0.25	−0.23
	valor p	—	0.978	< .001	< .001
Soltero	Diferencia de medias		—	−0.21	−0.18
	valor p		—	< .001	< .001
Divorciado/separado	Diferencia de medias			—	0.03
	valor p			—	0.997
Viudo	Diferencia de medias				—
	valor p				—

En este análisis hay dos tipos disponibles, Tukey si se cumple el supuesto de homogeneidad entre los grupos y Games-Howell si no se cumple. En este caso, como se cumple el supuesto de igualdad de varianzas pediremos la opción “Tuckey (varianzas iguales)”.

Podemos ver que las diferencias se encuentran en concreto entre las medias de los grupos de casados y divorciados, casado y viudos, teniendo los casados menor soledad en ambos casos. Y también entre solteros y divorciados y solteros y viudos teniendo los solteros menor media en soledad. Para ayudarnos a la interpretación, en el menú de la prueba podemos seleccionar la opción de “Tablas descriptivas” para ver el valor de la media para cada uno de los grupos.

Kruskal-Wallis

En la práctica muchas veces se incumple el supuesto de normalidad y deberemos recurrir a la prueba Kruskal Wallis:



ANOVA de Un Factor (No paramédico)

Kruskal-Wallis

	χ^2	gl	p
DEPRESION	7.50	3	0.035

Comparaciones dos a dos Dwass-Steel-Critchlow-Fligner

Comparaciones entre parejas - DEPRESION

		W	p
Casado	Soltero	0.90	0.920
Casado	Divorciado/separado	−0.41	0.991
Casado	Viudo	3.81	0.035
Soltero	Divorciado/separado	−0.83	0.936
Soltero	Viudo	2.21	0.399
Divorciado/separado	Viudo	2.51	0.285

Si la prueba de Kruskal Walis es estadísticamente significativa podemos también pedir pruebas post hoc en la opción “Comparaciones dos a dos DSCF”. El procedimiento de interpretación es equivalente.

3.2. Muestras dependientes

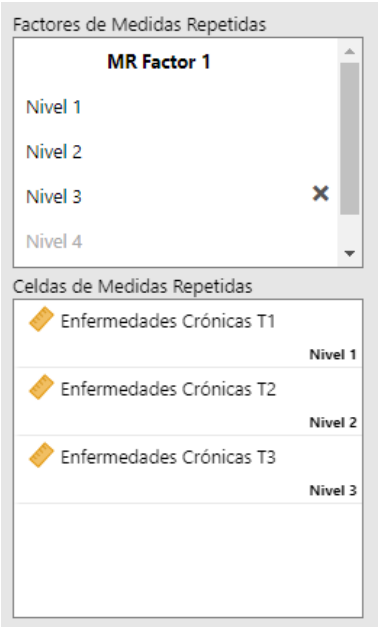
ANOVA de medidas repetidas

En el ANOVA de medidas dependientes o repetidas hay más de una medida por sujeto, y por tanto una nueva fuente de variación, que se extrae del error y al final el error es menor. Esta prueba no incluye en el apartado de supuestos la opción de comprobación del supuesto de normalidad, pero este se puede comprobar mediante la ruta de Jamovi que vimos en el punto 3 dedicado a la prueba de normalidad.

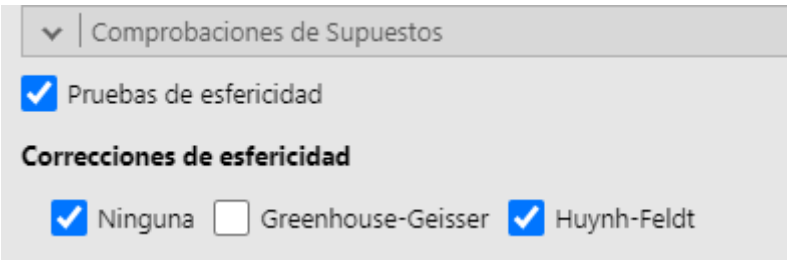
En el propio ANOVA de medidas repetidas sí podemos comprobar el supuesto de Esfericidad, un supuesto específico para este análisis, equivalente al supuesto de homogeneidad de varianzas en las otras pruebas.

Pongamos de ejemplo que nos interesa conocer si ha habido cambios en el número de enfermedades crónicas en una muestra de 250 personas en tres momentos temporales. Primero deberemos especificar en el ANOVA de Medidas Repetidas el

número de niveles que tenemos en el factor (en nuestro caso tenemos tres ya que la medida se repite tres veces). Luego hay que introducir la variable que tenemos repetida en la opción “Celdas de Medidas Repetidas”:



Comprobamos el supuesto de esfericidad, en caso de que se cumpla, marcaremos la opción “Ninguna” en correcciones de esfericidad. Si no se cumple el supuesto de Esfericidad marcaremos la opción “Huynh-Feldt”.



Supuestos

Pruebas de Esfericidad

	W de Mauchly	p	ε de Greenhouse-Geisser	ε de Huynh-Feldt
MR Factor 1	0.99	< .001	0.99	0.99

Como en este caso la prueba de Esfericidad es significativa (p<.001), no se cumple este supuesto y marcaremos la corrección de Huynh-Feldt.

ANOVA de Medidas Repetidas

Efectos Dentro de los Sujetos

	Corrección de Esfericidad	Suma de Cuadrados	gl	Media Cuadrática	F	p	η^2_p
MR Factor 1	Huynh-Feldt	1087.27	1.97	550.96	591.63	< .001	0.02
Residual	Huynh-Feldt	53997.39	57983.43	0.93			

Nota. Suma de Cuadrados Tipo 3

[3]

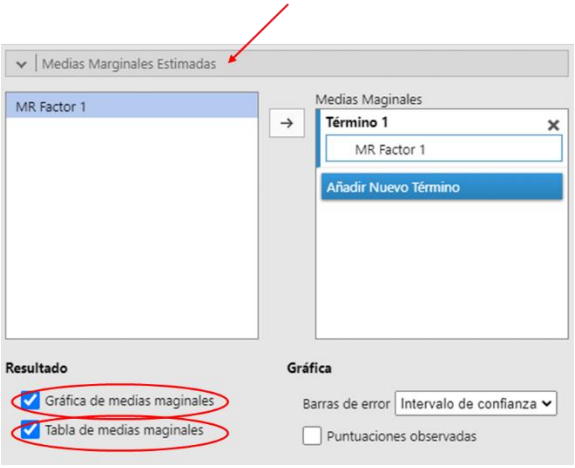
Comprobamos que la prueba para los efectos intra-sujetos es significativa ($p<.001$) con un tamaño del efecto de $\eta^2= 0.02$, por lo que sabemos que hay diferencias entre las medias de enfermedades crónicas que presentan las mismas personas a lo largo de los tres momentos temporales. Hacemos notar que la prueba de tamaño del efecto en esta ocasión es eta al cuadrado parcial, y así debemos indicarlo al solicitar el análisis. De nuevo para poder conocer entre que medias existen estas diferencias deberemos aplicar pruebas Post Hoc.

Pruebas Post Hoc

Comparaciones Post Hoc - MR Factor 1

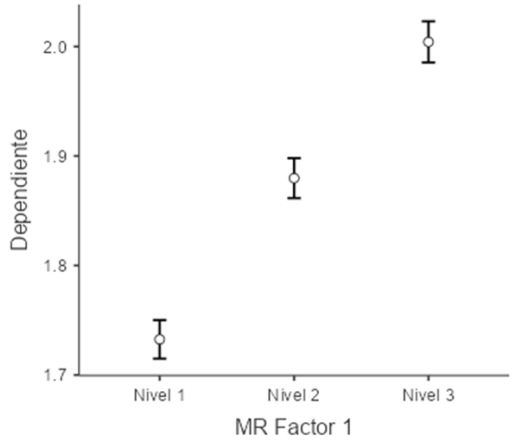
Comparación			Diferencia de Medias	EE	gl	t	p_{tukey}
MR Factor 1		MR Factor 1					
Nivel 1	-	Nivel 2	−0.15	0.01	29382.00	−19.70	< .001
	-	Nivel 3	−0.27	0.01	29382.00	−32.84	< .001
Nivel 2	-	Nivel 3	−0.12	0.01	29382.00	−15.64	< .001

Esta vez vemos que hay diferencias estadísticamente significativas entre todos los niveles (es decir entre todos los momentos temporales). En la Tabla de las pruebas Post Hoc las diferencias de medias ya nos da la información de en qué comparación de momentos temporales las personas han tenido una mayor media (por ejemplo, la diferencia entre el nivel 1 y nivel 2 es -0.15, por lo que sabemos que en el tiempo 2 la media es mayor que en el tiempo 1). No obstante, si queremos ver los valores de las medias en cada momento temporal, podemos pedir en la opción “Medias Marginales Estimadas”, “Resultado”, las gráficas de medias marginales y la tabla de medias marginales.



Medias Marginales Estimadas

MR Factor 1



Medias Marginales Estimadas - MR Factor 1

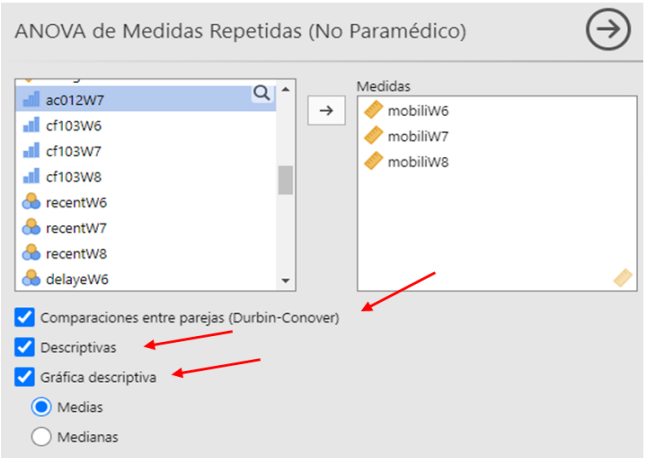
			Intervalo de Confianza al 95%	
MR Factor 1	Media	EE	Inferior	Superior
Nivel 1	1.73	0.01	1.71	1.75
Nivel 2	1.88	0.01	1.86	1.90
Nivel 3	2.00	0.01	1.99	2.02

En el gráfico y la tabla también se muestra que la media de enfermedades crónicas ha ido aumentando.

Prueba de Friedman

La Prueba de Friedman es la equivalente no paramétrica a la ANOVA de medidas repetidas. Será la prueba que utilicemos en caso de muestras pequeñas en las que las muestras no sigan una distribución normal. Este análisis también permite que

observemos las comparaciones entre pares, las medias o medianas de cada grupo y permite hacer una gráfica de las medias o medianas.



ANOVA de Medidas Repetidas (No Paramédico)

Friedman

χ^2	gl	p
1192.66	2	< .001

Comparaciones Entre Parejas (Durbin-Conover)

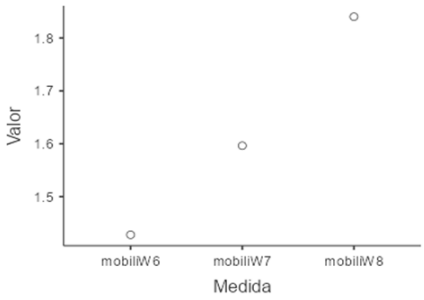
			Estadístico	p
mobiliW6	-	mobiliW7	14.71	< .001
mobiliW6	-	mobiliW8	34.75	< .001
mobiliW7	-	mobiliW8	20.05	< .001

[5]

Descriptivas

	Media	Mediana
mobiliW6	1.43	0.00
mobiliW7	1.60	1.00
mobiliW8	1.84	1.00

Gráfica Descriptiva



Referencias

Cohen, J. (1992). Statistical Power analysis. *Current Directions in Psychological Science, 1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Díaz, M. Á. R., Merino, A. P., Ruiz, M. Á., Martín, R. S., & Castellanos, R. S. M. (2015). *Análisis de datos en ciencias sociales y de la salud*. Síntesis.

The jamovi project (2022). jamovi. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2022-01-01).

TEMA 6

Regresión lineal

Adrián García Mollá
Zaira Torres Romero
Sara Martínez Gregorio

Tema 6. Regresión lineal

1. Introducción al modelo

En capítulos anteriores se explicó la inferencia estadística y hemos aprendido las diferentes técnicas para la comparación y la asociación entre variables de diferentes tipos. Sin embargo, la regresión lineal permite el estudio de la asociación entre variables orientado a la explicación o predicción. Lo cual significa que encontraremos una variable dependiente (cuantitativa) que es la variable que pretendemos explicar, y una o más variables independientes (de tipo cuantitativo o semicuantitativo), cuyos valores determinarán el valor de la variable dependiente.

Para contextualizar esta técnica estadística, comenzaremos definiendo los modelos estadísticos. A diferencia de la correlación, el análisis de regresión lineal requiere del establecimiento de un modelo, es decir, la forma que toman los datos que disponemos.

El análisis de regresión lineal se incluye dentro del modelo lineal general. En esencia, un modelo de regresión se plantea como una correspondencia de valores que sigue una pauta lineal donde debemos hallar los parámetros que mejor ajustan a los datos que disponemos.

2. Regresión lineal simple

Como se explicaba anteriormente, este modelo se emplea para estudiar el comportamiento una variable dependiente en función de la variable predictora o independiente. En el caso de la regresión simple dispondremos únicamente de dos parámetros, el modelo se representa matemáticamente como sigue:

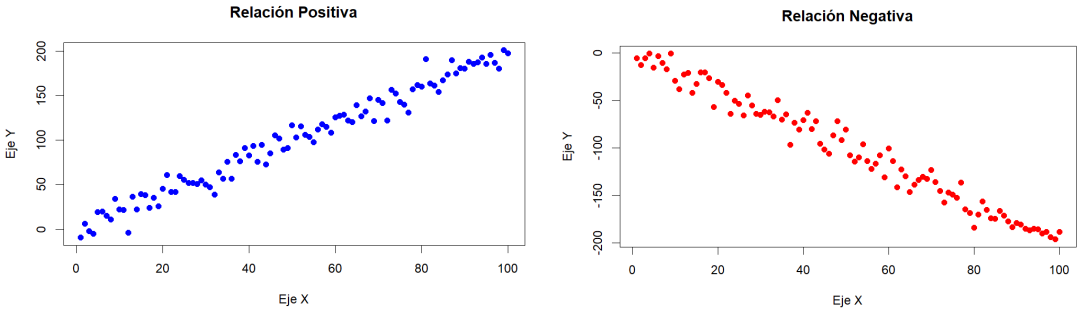
$$Y = A + BX + \varepsilon$$

Donde **Y** se corresponde con la variable dependiente y **X** con la variable independiente o predictora. **A** representa la intercepta, es decir, el valor de la variable dependiente cuando la variable independiente equivale a 0. **B** es la pendiente, es decir, indica el cambio producido en la variable dependiente (Y) por cada cambio de unidad de la

variable independiente (X). Por último, **ε** representa el error de medida, es decir, la diferencia entre la puntuación predicha **Ŷ** y la puntuación verdadera de **Y**.

$$\varepsilon = Y - \hat{Y}$$

En última instancia, el modelo de regresión trata de cuantificar la relación lineal que podemos observar en un diagrama de dispersión de puntos de manera similar a como ocurre en los análisis de correlación. Aquí se muestran dos ejemplos de diagramas de dispersión que muestran tanto una relación positiva como una relación negativa.



Por lo tanto, el fin de este tipo de análisis es hallar la recta que mejor se ajuste a los datos.

2.1. Mínimos cuadrados

En una situación ideal donde el predictor fuese capaz de explicar por completo el comportamiento de la variable dependiente, no existiría la necesidad de hallar la recta que mejor describa el diagrama. Sin embargo, en investigación jamás tendremos un ajuste perfecto de nuestro modelo a los datos.

El criterio de selección de modelos mayormente utilizado es el de **mínimos cuadrados**, y se basa en las distancias entre las puntuaciones verdaderas y las puntuaciones predichas (error de predicción). Se trata de elevar al cuadrado todos los errores de predicción y sumarlos. Nuestro modelo será mejor cuanto menor sea la suma total de cuadrados, y esto se traduce en escoger el modelo, es decir, los coeficientes **A** y **B** que minimizan el error de predicción.

2.2. Coeficientes de regresión

Una vez seleccionado el modelo, deberemos realizar la interpretación de los coeficientes para poder explicar la manera en que los predictores afectan a la variable dependiente. De esta manera, el coeficiente **A** no nos proporcionará información demasiado valiosa, simplemente indica el valor de Y cuando X = 0.

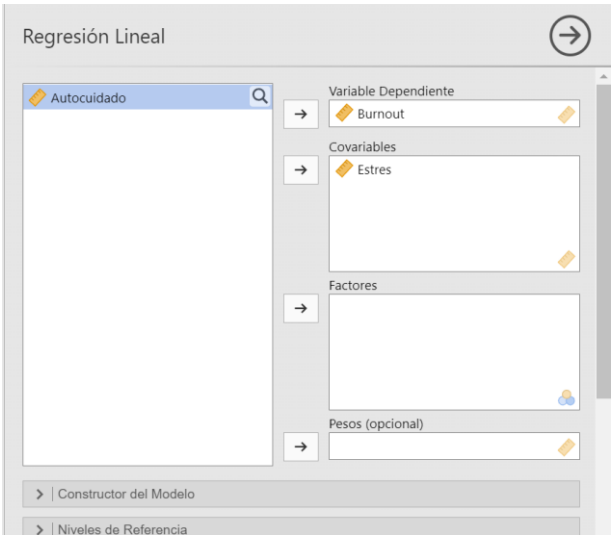
Sin embargo, el coeficiente **B** permite conocer si la relación es positiva o negativa, tal y como se muestra en los diagramas expuestos anteriormente. En caso de que B < 0, la relación será **negativa** y, en caso de que B > 0, la relación será **positiva**. En caso de que la relación lineal entre las variables no exista, el coeficiente será 0, es decir, la pendiente será nula y la recta será paralela al eje x.

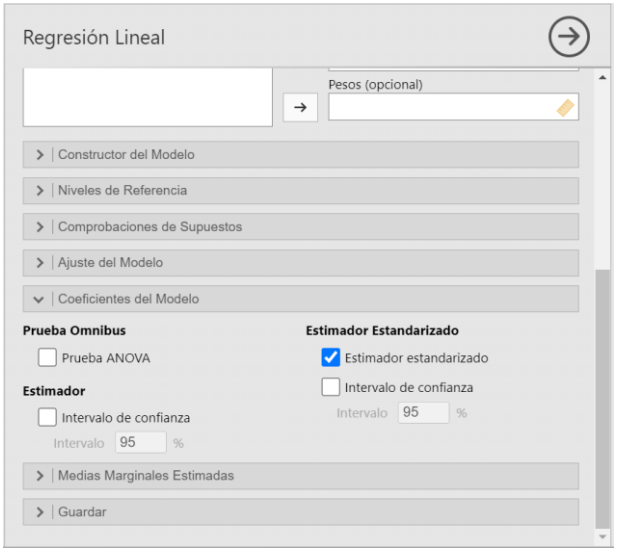
Con respecto del valor del coeficiente **B**, indicará el cambio en la variable dependiente por cada unidad de cambio en la variable independiente. Veámoslo en un ejemplo con Jamovi.

2.3. Ejemplo de regresión lineal simple

La ruta en Jamovi para llevar a cabo un análisis de regresión lineal es la siguiente: “Análisis”, “Regresión” y “Regresión lineal”. A continuación, introduciremos la variable dependiente (Y) y la variable predictora (X). También es conveniente seleccionar la opción de “Estimador estandarizado” en la pestaña de “Coeficientes del Modelo”.

A continuación, se muestra un ejemplo con un modelo de regresión lineal simple con una puntuación de una escala de Burnout como variable dependiente y una puntuación de una escala de estrés como variable independiente.





A continuación, se muestran los resultados del modelo en dos tablas en el visualizador de Jamovi. En la primera tabla, observamos los ajustes del modelo, nos reporta información sobre R (el coeficiente de relación) y R^2 (coeficiente de determinación). R en el análisis de regresión simple coincide con la correlación entre las variables (β).

Regresión Lineal

Medidas de Ajuste del Modelo

Modelo	R	R^2
1	0.731	0.534

Coeficientes del Modelo - Burnout

Predictor	Estimador	EE	t	p	Estimador Estándar
Constante	60.211	7.101	8.48	< .001	
Estres	0.676	0.132	5.14	< .001	0.731

El coeficiente de determinación (R^2) muestra la proporción de varianza de la variable Y explicada por la variable X. Tiene valores mínimos de 0 y máximos de 1. Si la proporción es 0, la variable predictora no tiene ninguna capacidad predictiva. Si es 1, la variable predictora explicaría toda la variación de las predicciones y las predicciones no tendrían error.

En la segunda tabla, encontramos los coeficientes del modelo sin estandarizar y estandarizados. Respecto a los coeficientes estandarizados encontramos:

A = es el valor de la constante, indica el valor de la ordenada en su origen, Y cuando X=0.

B = es la pendiente, indica las unidades de incremento de Y por cada unidad que incrementa X, indica la tasa de cambio en la escala original de las variables.

En la columna de los coeficientes estandarizados encontramos:

α = es el valor de la constante estandarizado, siempre es 0 en caso de recta de regresión estandarizada

β = es la pendiente estandarizada, continúa indicando la tasa de cambio, pero en lugar de indicar unidades de incremento de Y por cada unidad de incremento de X, indica el aumento de desviaciones típicas en Y por cada desviación típica que aumenta X. Los coeficientes estandarizados tienen valores que van de -1 a 1 y se interpretan de manera parecida al coeficiente de Pearson. Valores más cercanos a 0, en valor absoluto, indicarán una menor asociación entre la variable predictora (X) y dependiente (Y). Es decir, se considerará un peor predictor de la variable dependiente. Mientras que los valores cercanos a 1, en valor absoluto, indicarán que hay mayor relación y que la variable es un mejor predictor.

3. Regresión lineal múltiple

En la mayoría de las ocasiones, las variables de estudio presentan una naturaleza compleja y pueden deberse a la presencia de diferentes circunstancias. Por lo tanto, será necesario incluir en los modelos de regresión diferentes variables predictoras (X) a la vez, en este caso la recta de regresión introducirá un nuevo valor de X y B por cada variable que se incluya en el modelo:

$$Y = A + B_1X_1 + \ B_2X_2 + \cdots + B_kX_k + \varepsilon$$

La interpretación de la regresión múltiple es muy parecida a la de la regresión simple, pero tiene algunas peculiaridades:

- El coeficiente de correlación múltiple (R) medirá el grado de relación entre la variable dependiente (Y) y todas las variables independientes.
- Mediante los valores estandarizados (β) podemos observar qué variable es un predictor más potente de la variable dependiente.

3.1. Ejemplo de regresión lineal múltiple

En Jamovi, la ruta para la regresión lineal múltiple es igual que para la regresión lineal simple. Sin embargo, deberemos incluir más de una variable independiente. Un ejemplo con varios predictores sería el siguiente:

Regresión Lineal

Medidas de Ajuste del Modelo

Modelo	R	R^2
1	0.777	0.603

Coeficientes del Modelo - Burnout

Predictor	Estimador	EE	t	p	Estimador Estándar
Constante	125.320	34.023	3.68	0.001	
Estres	0.401	0.188	2.13	0.045	0.433
Autocuidado	−0.524	0.268	−1.95	0.064	−0.397

Como se observa, al modelo anterior se le añade el autocuidado como predictor. Este predictor no presenta un efecto significativo en el modelo, sin embargo, es interesante tener en cuenta que su relación con el Burnout es negativa, es decir, cuanto mayor es la puntuación de autocuidado menos probable es sufrir Burnout. Por el contrario, el estrés aumenta la probabilidad de sufrir Burnout.

Referencias

The jamovi project (2022). jamovi. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2022-01-01