Servei d'Estadística
Universitat Autònoma de Barcelona

Manual de Introducción a Jamovi: una interfaz gráfica para usuarios de R

Llorenç Badiella. Director del Servei d'Estadística Aplicada
Anabel Blasco. Asesora estadística del Servei d'Estadística Aplicada
Ester Boixadera. Asesora estadística del Servei d'Estadística Aplicada
Oliver Valero. Asesor estadístico del Servei d'Estadística Aplicada
Ana Vázquez. Asesora estadística del Servei d'Estadística Aplicada

Manual de Introducción a Jamovi

Servei d'Estadística Aplicada
Universitat Autònoma de Barcelona

Campus UAB - Edifici CM7
08193 Cerdanyola del Vallès
(Barcelona)
Tel. 93.581.13.47
s.estadistica@uab.cat
http://serveis.uab.cat/estadistica

Publicado por el Servei d'Estadística Aplicada de la UAB

1ª edición, Abril 2021

CONTENIDOS

1 PRESENTACIÓN

Este manual de introducción a **Jamovi** pretende ser una primera aproximación al uso del programa Jamovi para aquellas personas que deseen adquirir conocimientos de Estadística, y que deseen introducirse en el uso de este software para aplicarlo en su área de conocimiento y trabajo.

Jamovi es una Interfaz Gráfica de Usuario (GUI en inglés), creada por Jonathon Love, Damian Dropmann y Ravi Selker, que permite acceder a muchas capacidades del entorno estadístico **R** sin que el usuario tenga que conocer el lenguaje de comandos propio de este entorno.

Algunas ventajas de la interfaz **Jamovi** son:

- Es sencilla de usar.
- Permite el acceso a las funciones y gráficos estadísticos más comunes.
- Facilita la realización de tareas más complejas.
- Es multisistema y multiplataforma.
- Versión demo que permite trabajar directamente desde la nube.
- Es fácilmente extensible y personalizable.
- Ofrece la posibilidad de incluir código R.

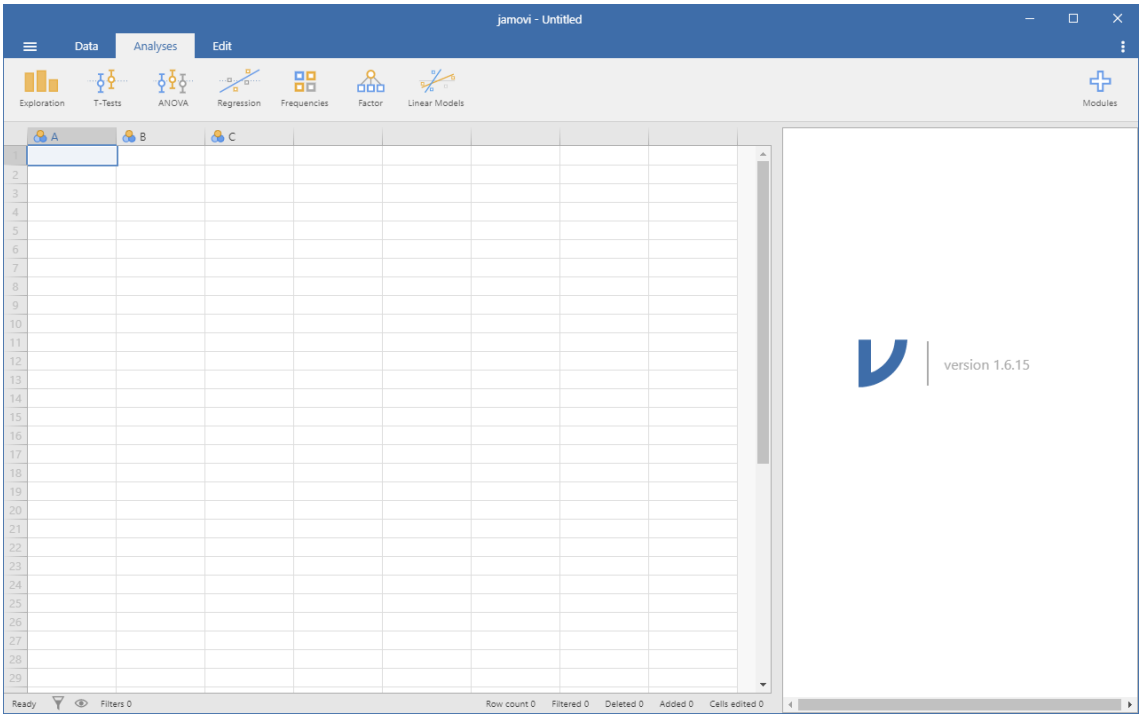
Para ver más detalles sobre el programa consultar la página web:

www.jamovi.org

2 INTRODUCCIÓN A JAMOVI

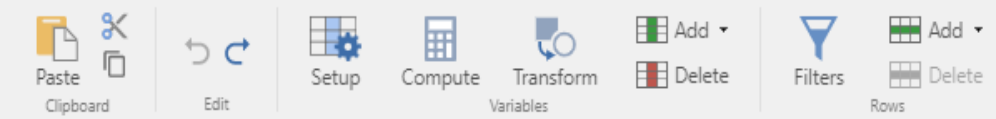
2.1 Las ventanas de Jamovi

Jamovi es una nueva hoja de cálculo estadístico de "tercera generación". En la cuadrícula de la izquierda podremos introducir datos manualmente o bien importar ficheros con extensión ".csv", ".txt" o ".xlsx" entre otros. En la ventana de la derecha se irán guardando los resultados generados, y estos se actualizarán automáticamente si se modifica la base de datos.



Desde el menú principal podremos abrir una nueva hoja de datos, abrir una base de datos ya creada en Jamovi (extensión ".omv") o importar una base de datos de texto, Excel, SPSS, R, Stata, o SAS. También podremos exportar una base de datos a CSV, R, SPSS, SAS o Stata, y los resultados a un documento pdf o html.

Desde el menú **Data** podremos realizar operaciones en la base de datos:



- **Clipboard**: Permite pegar, cortar o copiar observaciones.
- **Edit**: Permite deshacer/rehacer una acción.
- **Variables**: Permite definir las propiedades de las variables (Setup), crear nuevas variables (Compute), transformar o recodificar variables (Transform) o añadir o eliminar variables (Add/Delete).
- **Rows**: Permite aplicar filtros y añadir o eliminar observaciones.

En el menú **Analyses** podremos realizar operaciones en la base de datos:

- **Exploration**: Realiza tablas de resumen y gráficos.
- **T-Tests**: Incorpora pruebas paramétricas y no paramétricas para comparar medias/medianas de 1 población, 2 muestras independientes y 2 muestras relacionadas.
- **ANOVA**: Para realizar ANOVA de 1 factor, de varios factores, ANCOVA, MANCOVA, medidas repetidas y pruebas no paramétricas.
- **Regression**: Permite calcular correlaciones, ajustar modelos de regresión lineal y modelos de regresión logística (binaria, ordinal y multinomial).
- **Frequencies**: Tablas de contingencia, test de Chi-cuadrado y test de McNemar.
- **Factor**: Análisis de fiabilidad, análisis de componentes principales y análisis factorial.

Desde 

se pueden incorporar módulos adicionales que incorporan técnicas estadísticas más avanzadas. Algunos de estos módulos son:

- **surveymv**: Genera gráficos de resumen para múltiples variables.
- **jpower**: Análisis de potencia para los diseños de investigación más comunes.
- **deathwatch**: Análisis de supervivencia.
- **jsq**: Incorpora métodos estadísticos bayesianos.
- **gamlj**: Para la estimación de modelos lineales, como el modelo lineal general, el modelo lineal mixto, los modelos lineales generalizados y los modelos mixtos generalizados.
- **Rj**: Proporciona un editor que permite insertar código R y analizar los datos usando R dentro de Jamovi.

2.2 Crear y abrir ficheros

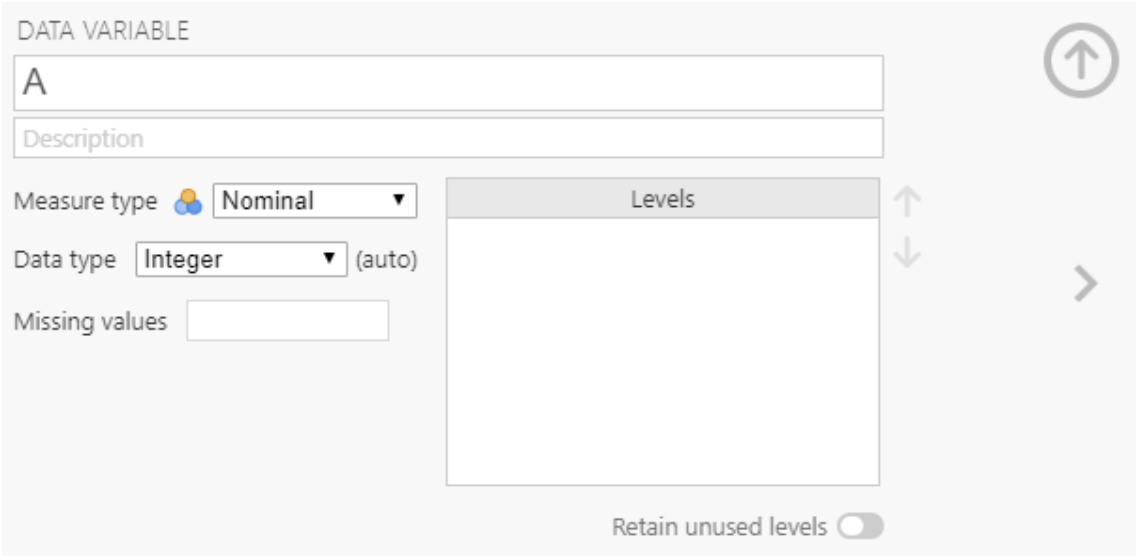
Para analizar datos lo primero es crear o abrir un archivo de trabajo. Se pueden introducir datos creando una nueva base de datos e introduciendo los datos manualmente, abriendo un fichero de **Jamovi** existente, o importando un fichero procedente de otra aplicación.

2.2.1 Crear una nueva base de datos

La base de datos está dividida en filas y columnas dando lugar a celdas o casillas donde se recogen los datos. Cada columna tiene asignado un nombre de variable, ya sea especificado por el usuario o bien por el propio programa. Las filas, a su vez, están numeradas de forma correlativa. A partir de las pestañas "**Add column**" y "**Add row**" se pueden añadir filas y columnas respectivamente.

Para introducir datos se pueden crear nuevas filas y columnas e introducir datos manualmente, o bien copiar datos de otras aplicaciones y pegarlos en la tabla.

En la primera fila podremos definir los nombres de las variables. Para acceder a las propiedades de las variables podemos utilizar el menú **Data → Setup** o bien hacer doble clic en una variable. Se abrirá el editor de variables:



Para minimizar esta ventana y volver a la base de datos podemos utilizar la flecha , y para acceder a las propiedades de otras variables utilizaremos las flechas laterales .

Observación: Los nombres de las variables pueden tener acentos y espacios, pero no es aconsejable si más adelante queremos incorporar código de R.

Tipos de variables

Las variables, tal y como hemos dicho, definen las columnas del fichero de datos y son características de los individuos. Pueden ser diferenciadas según:

- **Cualitativas** o Categóricas: etiquetas que representan el grupo o categoría a la cual pertenece un individuo. Se puede diferenciar entre nominales (por ejemplo, el sexo) y ordinales (nivel de estudios).

- **Cuantitativas**: valores numéricos para los que tiene sentido realizar aritmética. Se puede diferenciar entre continuas (índice de masa corporal) y discretas (número de hijos). Pueden ser valores enteros o con decimales.

Observación: En Jamovi el separador de decimales es el punto.

- **ID**: este tipo de variable es específico de Jamovi. Está destinado a variables que contienen identificadores que casi nunca se analizan, como los nombres o el número de historia clínica.

Valores faltantes (missings)

Los valores faltantes se indican dejando la casilla en blanco. Adicionalmente, desde la casilla "**Missing values**" se podrían indicar otros valores para ser tratados como missings, como por ejemplo el 99:

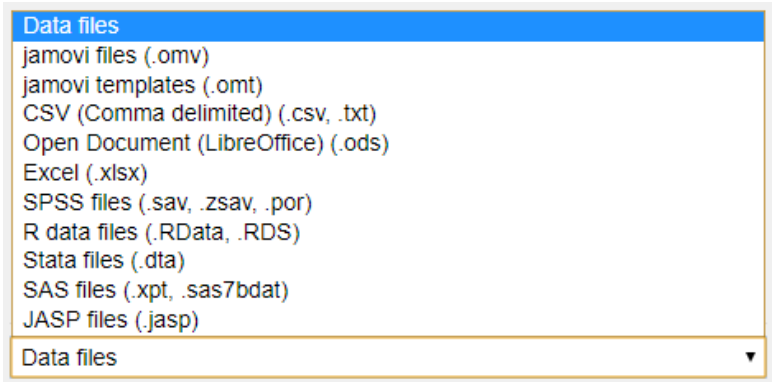
Ejercicio:

Crear una base de datos con la siguiente información:

	NHIST	Sexo	Edad	IMC	Valoración
1	39	Mujer	66	22.8	Muy mala
2	45	Hombre	57	24.4	Regular
3	52	Hombre	41	25.2	Buena
4	57	Hombre	59	22.4	Buena
5	56	Hombre	63	28.2	Regular

2.2.2 Importar bases de datos

Podemos abrir una base de datos utilizando el menú ☰ → **Open**. Con esta opción podemos abrir datos que se encuentren en formato de **Jamovi** (extensión "**.omv**"), en formato texto, u otros tipos de formato como por ejemplo Excel o SPSS. Estos son los formatos compatibles con Jamovi:

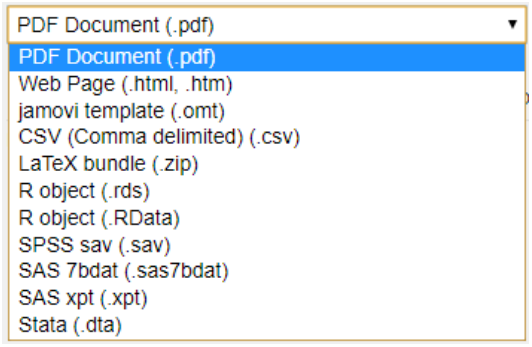


Observación: El menú ☰ → Import también permite abrir una base de datos externa, pero no se recomienda utilizarlo si ya tenemos una base de datos abierta.

Desde el menú ☰ → **Open** seleccionamos la opción Browse e indicamos dónde se encuentra la base de datos que deseamos abrir. Una vez abierta deberemos revisar que el tipo de variables se ha definido correctamente, así como el orden de los niveles de las variables cualitativas.

2.3 Guardar bases de datos

Las bases de datos pueden ser guardadas en formato **Jamovi** (extensión "**.omv**") desde el menú ☰ → **Save as**, o bien ser exportadas a otro formato desde ☰ → **Export**. Las opciones disponibles son:



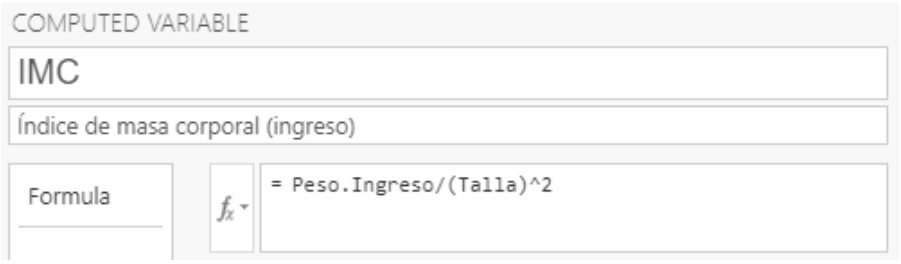
Ejercicio: Abrir los ficheros **ADL.xlsx**, **ADL.txt** y **ADL.sav**. Comparar las propiedades de las variables de las 3 bases de datos. Realizar las correcciones oportunas en el fichero importado desde Excel y guardar la base de datos en formato "**.jmv**".

3 GESTIÓN DE BASES DE DATOS

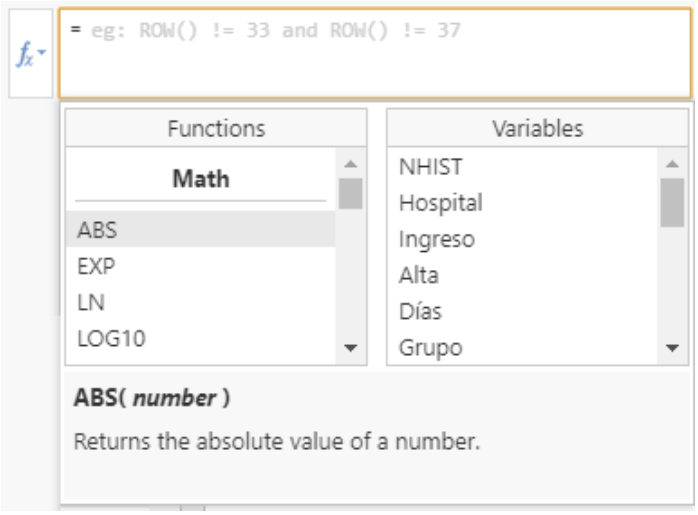
El menú "**Data**" permite gestionar las bases de datos. En particular permite calcular nuevas variables a partir de las ya existentes, aplicar alguna transformación, recodificar variables, editar los factores de las variables categóricas, o seleccionar un subconjunto de datos.

3.1 Crear nuevas variables

El menú **Data → Compute variable** permite crear nuevas variables:



También se pueden aplicar transformaciones a las variables utilizando funciones ya implementadas:



Algunas funciones de interés son:

- **LN(x)**: Devuelve el logaritmo neperiano (para valores mayores que 0).
- **ABS(x)**: Devuelve el valor absoluto.
- **Z**: Reescala las variables para que tengan media 0 y desviación estándar 1 (variable estandarizada).
- **RANK**: Reemplaza los valores por su rango.

Ejercicio: Calcular una nueva variable 'Factores.Riesgo' como la suma de los factores de riesgo (obesidad, fumador, diabetes y HTA).

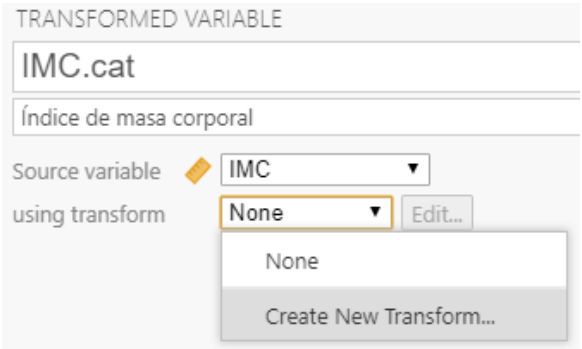
Observación: Las variables calculadas a partir de una fórmula pasan a ser automáticamente variables numéricas. En el caso de la variable 'Factores.Riesgo', si quisiéramos tratarla también como categórica, deberíamos crear una variable nueva, copiar los datos de la variable numérica, y definirla como ordinal.

3.2 Recodificar variables

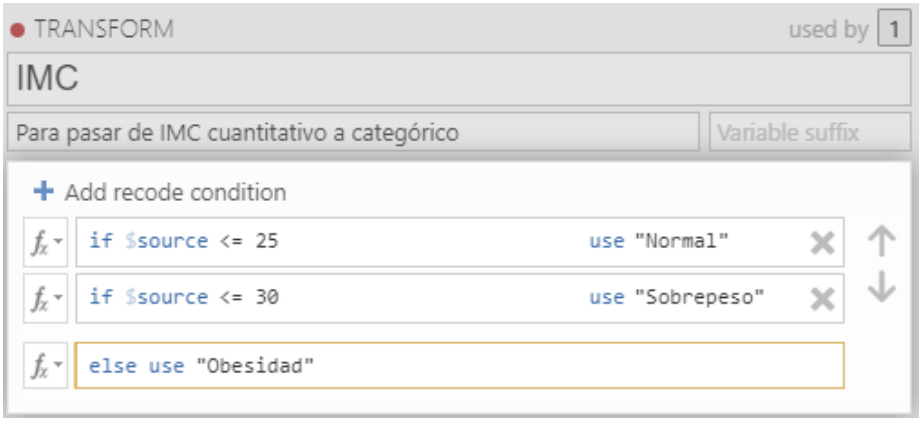
Recodificar una variable consiste en asignar una nueva codificación a sus valores originales, o agrupar rangos de valores existentes en nuevos valores, de manera que se modifica su codificación original.

Existen varias situaciones donde nos puede interesar recodificar los valores de una variable: cuando queremos pasar de una variable continua a una variable categórica, como por ejemplo el IMC o la edad, o cuando queremos agrupar categorías de una variable categórica como por ejemplo el número de factores de riesgo o la valoración de salud.

Para recodificar una variable, al hacer clic con el botón derecho seleccionaremos la opción "**Transform**". Una vez definido el nombre de la nueva variable debemos crear una nueva transformación:



Esta transformación podría ser empleada posteriormente en otras variables. Tras indicar el nombre de la transformación, podemos añadir condiciones desde ➕:

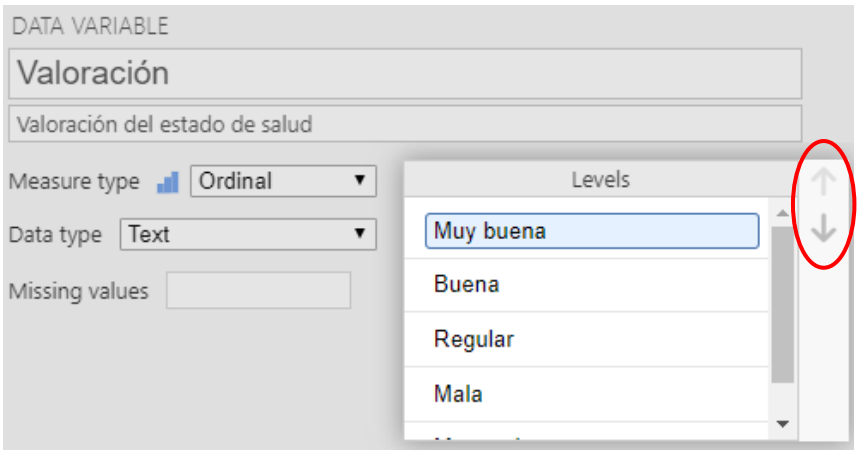


Al definir las condiciones, $source hace referencia a los valores de la variable original. Los valores de las categorías de la nueva variable deberán indicarse entre comillas (pueden ser dobles o simples).

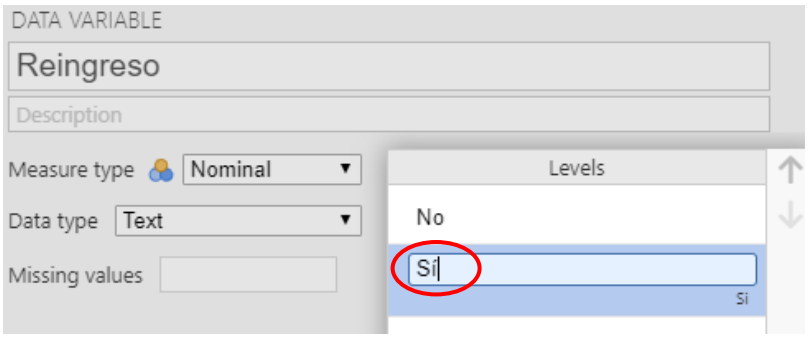
Ejercicio: Recodificar la variable 'Factores.Riesgo' en una nueva variable agrupando los valores de 2 a 4 en una misma categoría. A continuación, recodificar las variables 'Edad' y 'Valoración' en tres categorías cada una.

3.3 Editar factores

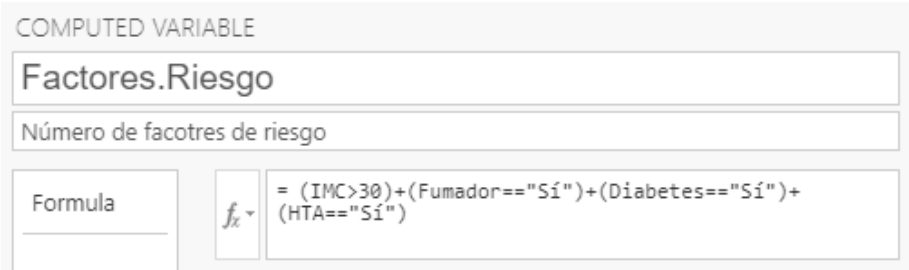
Los niveles de las variables categóricas (factores) se muestran en la pestaña "**Levels**" (se accede a través del menú **Data → Setup**, o bien haciendo doble clic en el nombre de una variable). Por defecto de ordenan alfabéticamente, pero se puede modificar el orden desplazando los niveles hacia arriba y hacia abajo:



También es posible modificar los nombres de los niveles, simplemente sobrescribiendo los valores:



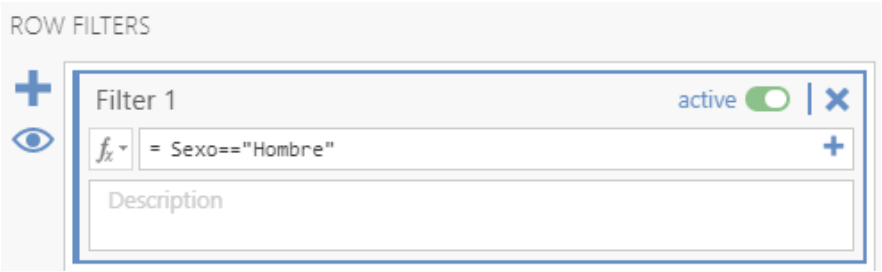
Observación: Al cambiar los valores de los factores de riesgo habrá que cambiar también los valores en la fórmula para calcular los factores de riesgo:



Ejercicio: Revisar los niveles de las variables categóricas de la base de datos. Ordenar los niveles que corresponda y modificar las etiquetas de las variables binarias (Sí/No). Guardar la base de datos resultante como "ADL_Final.omv".

3.4 Filtrar casos

En ocasiones podemos estar interesados en estudiar un subconjunto de registros de la base de datos. El menú **Data → Filters** permite aplicar filtros a la base de datos seleccionando únicamente los registros que cumplan una determinada condición. Para crear una base de datos con los pacientes de sexo masculino utilizaremos la siguiente instrucción:



Observación: Para las variables categóricas el valor de la categoría debe ir entre comillas. El programa distingue entre mayúsculas y minúsculas.

Ejercicio: Seleccionar los pacientes mayores de 60 años que tengan hipertensión.

4 VALIDACIÓN DE LA BASE DE DATOS

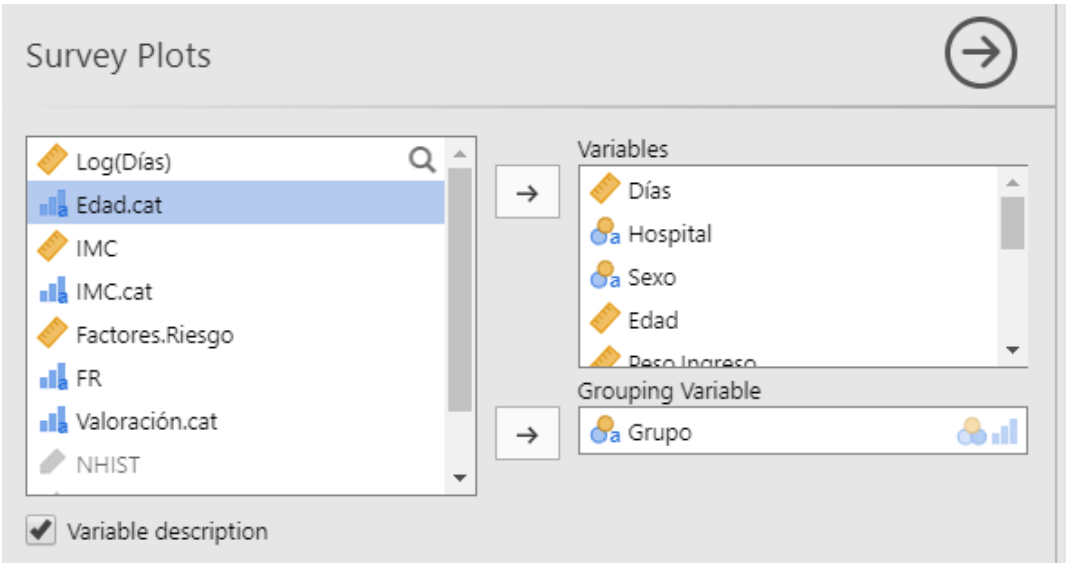
Antes de realizar cualquier análisis debe hacerse un ejercicio de **validación de la base de datos**.

- Detectar posibles errores en las variables, esto quiere decir encontrar **rangos de valores** y algunos estadísticos descriptivos para las variables cuantitativas, y **tablas de frecuencias** para las variables cualitativas.

- Validar la **consistencia interna** de los datos. Así, por ejemplo, en datos de encuesta es validar la congruencia de las respuestas en el sentido que, si un individuo responde una determinada opción en una pregunta, entonces sólo puede responder unas opciones concretas de otras.

Para poder llevar a cabo este proceso hace falta conocer bien de donde provienen los datos.

Una herramienta útil para validar las variables es a partir del módulo extra '**surveymv**'. Este módulo se instala desde la pestaña → **Jamovi library** → **install**. Tras su instalación aparecerá un submenú adicional dentro de **Analyses** → **Exploration**. Podríamos seleccionar todas las variables y graficarlas según el grupo:



5 ANÁLISIS DESCRIPTIVO

5.1 Introducción

Plantearse algunas preguntas preliminares puede ayudar a distinguir qué tiene sentido y qué no:

- Conocer la fuente de dónde provienen los datos nos puede informar de su calidad.
- Saber si la información de que disponemos es completa en el sentido que sea posible extraer conclusiones y no sólo impresiones.
- Plantear qué pueden ilustrar los datos.

La **ESTADÍSTICA DESCRIPTIVA** es un conjunto de métodos e ideas para organizar y describir los datos mediante gráficos y medidas de resumen numéricas.

5.2 Estadísticos resumen

Como hemos visto en los apartados previos, las variables pueden ser diferenciadas según:

- **CUALITATIVAS** o **CATEGÓRICAS**
- **CUANTITATIVAS**

Las variables también las clasificamos en función del rol que tienen en el análisis:

- Variable **RESPUESTA** (variable de interés, Y). Mide el resultado del estudio.
- Variables **EXPLICATIVAS** (X). Variables de control que contribuyen a explicar su comportamiento.

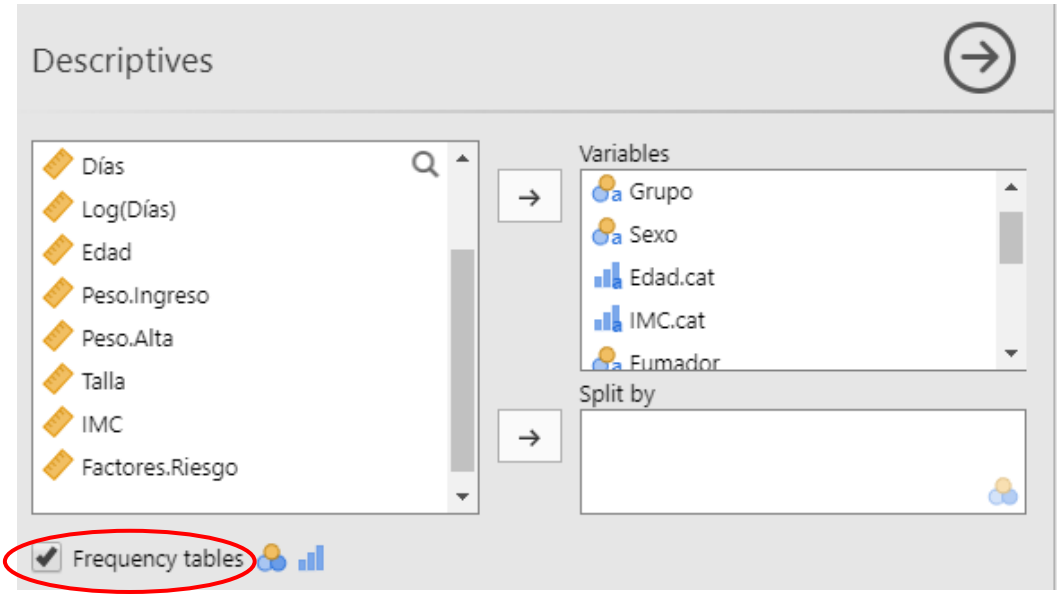
5.2.1 Variables cualitativas

Para resumir una variable cualitativa o cuantitativa de valores enteros utilizaremos las **Tablas de Frecuencias**.

- El número de veces que se repite un valor en una variable es la **frecuencia absoluta,** $\mathbf{f_a}$. Si n es el total de individuos, entonces $\mathbf{f_a/n}$ es su **frecuencia relativa**.
- La **frecuencia acumulada** es la suma de frecuencias absolutas hasta un determinado valor una vez ordenados de forma creciente los valores de la variable (ordinal o cuantitativa con valores enteros).

La **distribución de una variable** es el conjunto de valores juntamente con sus frecuencias (absolutas o relativas).

En **Jamovi** podemos obtener las frecuencias a través del menú **Analyses → Exploration → Descriptives**, y marcaremos la casilla "**Frequency tables**":



Para seleccionar más de una variable a la vez, utilizar la tecla 'Control'.

Tras seleccionar las variables automáticamente aparecerán las tablas de resultados en la ventana de la derecha. Para finalizar podemos hacer clic en la flecha:

Frequencies

Frequencies of Hospital

Levels	Counts	% of Total	Cumulative %
A	111	34.2 %	34.2 %
B	214	65.8 %	100.0 %

Frequencies of Sexo

Levels	Counts	% of Total	Cumulative %
Hombre	162	49.8 %	49.8 %
Mujer	163	50.2 %	100.0 %

Para cada variable seleccionada obtenemos la tabla de frecuencias con las frecuencias absolutas (**Counts**) y relativas (**% of Total**) de las distintas categorías.

Observación: Estos resultados se pueden copiar y pegar en un documento Word haciendo clic con el botón derecho. También se pueden añadir comentarios (**Add note**).

5.2.2 Variables cuantitativas

Para las variables cuantitativas, en las que puede haber un gran número de valores observados distintos, se debe optar por un método de análisis distinto, respondiendo a las siguientes preguntas:

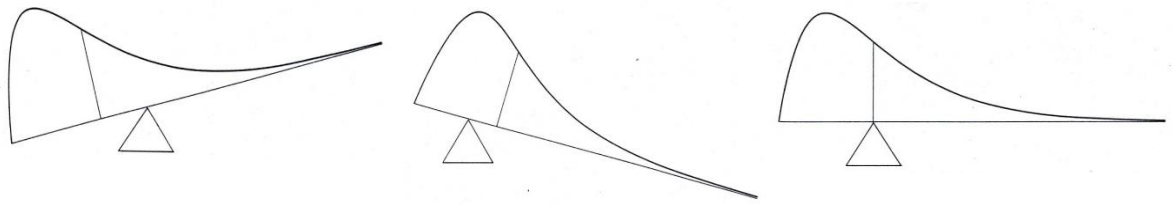
1. ¿Alrededor de qué valor se agrupan los datos?
2. Supuesto que se agrupan alrededor de un número, ¿cómo lo hacen? ¿muy concentrados? ¿muy dispersos?

5.2.2.1 Medidas de localización

Se utilizan para resumir las características más relevantes de los datos. Podemos utilizar:

- **Media ($\overline{X}$): centro de masas**
- **Mediana: punto medio**
- **Moda: el valor más repetido**

La media se sitúa en el punto de equilibrio del histograma de una variable cuantitativa:



La **Media** y la **Mediana** coinciden si la distribución es simétrica. Si no coinciden, es preferible la mediana (es menos sensible a datos extremos).

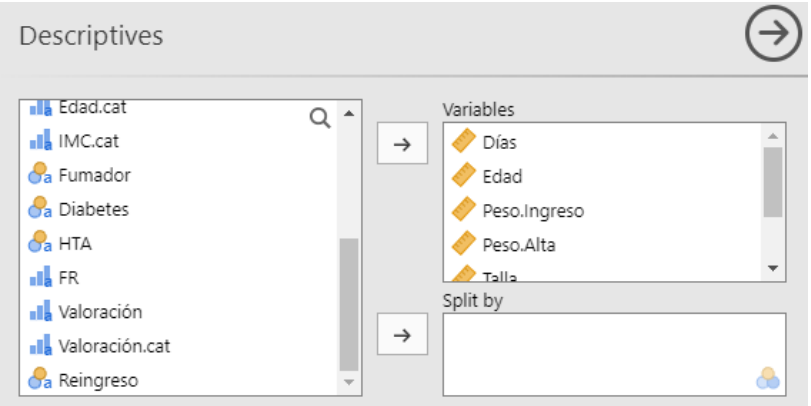
Otras medidas de resumen son los **Cuartiles** (Q1, Q2 y Q3), también llamados **Percentiles** (P25, P50, P75). Estos tres valores dividen la distribución en cuatro partes iguales.

5.2.2.2 Medidas de dispersión

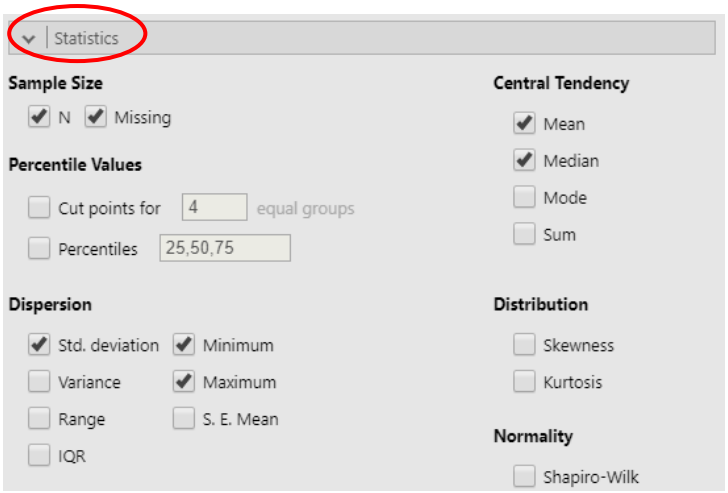
Sirven para resumir la dispersión. Las más habituales son:

- **Rango** = max – min
- **Rango Intercuartil (IQR)** = Q3 – Q1
- **Varianza** (S^2): una medida de la dispersión entorno de la media
- **Desviación estándar** (S)
- **Error estándar de la media** = $S/\sqrt{n}$

En **Jamovi** podemos obtener los estadísticos de resumen a través del menú **Analyses → Exploration → Descriptives**, y seleccionaremos variables cuantitativas:



El desplegable '**Statistics**' podemos seleccionar los estadísticos:



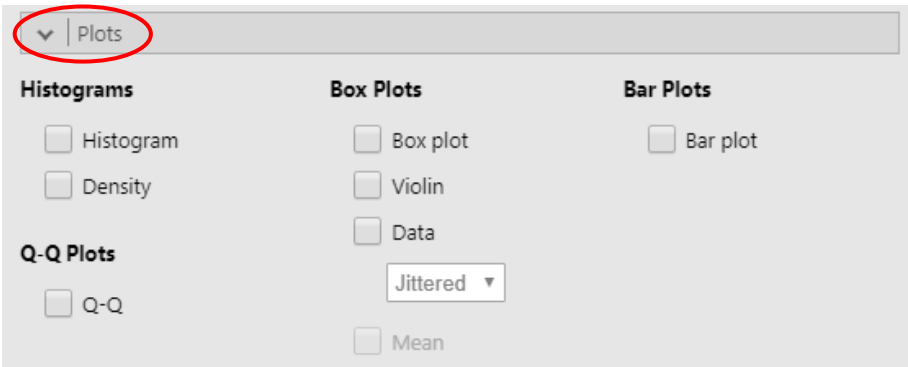
Descriptives

	Días	Edad	IMC	Factores.Riesgo
N	325	325	325	325
Missing	0	0	0	0
Mean	19.9	52.2	26.3	0.868
Median	20	52	26.0	1
Standard deviation	7.46	8.30	1.88	0.823
Minimum	2	34	20.8	0
Maximum	41	76	32.4	4

Observación: Haciendo doble clic en una tabla de resultados se vuelve a abrir el menú donde podemos seleccionar otros estadísticos.

5.3 La representación gráfica más adecuada

En el mismo menú de exploración encontramos un desplegable "**Plots**", donde podemos seleccionar los siguientes gráficos:

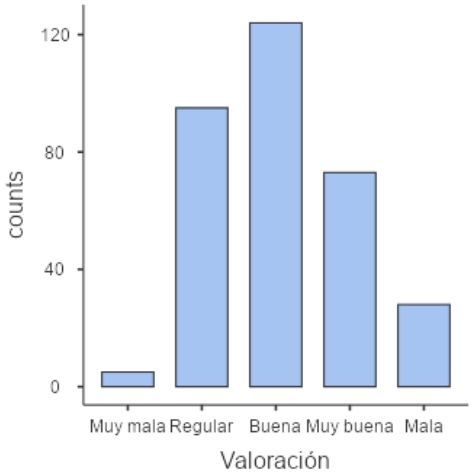


5.3.1 Variables cualitativas

Se representan las frecuencias o porcentajes de las diferentes categorías. Se pueden utilizar **diagramas de barras** o **gráficos de sectores**.

5.3.1.1 Diagrama de barras

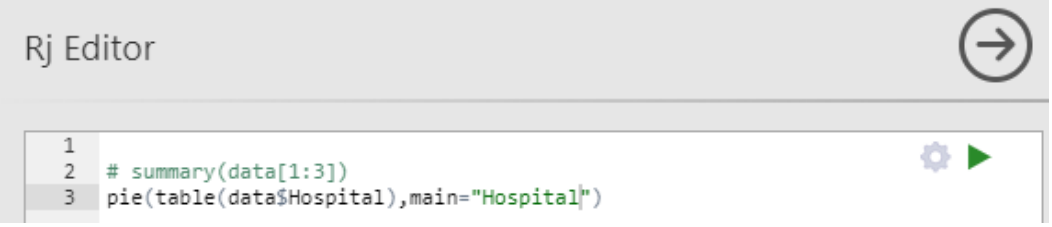
El gráfico de barras ("**Bar Plots**") es la única opción disponible por el momento en Jamovi para representar variables cualitativas. Al marcar la casilla automáticamente nos añade en la ventana de resultados los gráficos para las variables seleccionadas. El gráfico de barras para la variable 'Valoración' es el siguiente:



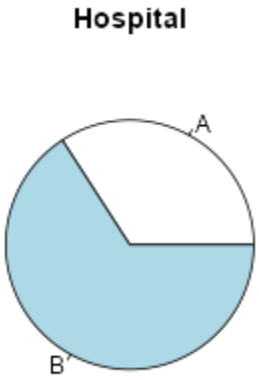
Observación: Los gráficos se pueden exportar a pdf, png o eps.

5.3.1.2 Gráficos de sectores

Si bien este tipo de gráfico no se podría obtener directamente desde los menús de Jamovi, se podría obtener mediante el módulo extra que permite introducir código de **R**:



El gráfico obtenido tras compilar el código ( ) es el siguiente:



Observación: El módulo Rj Editor se puede añadir desde la pestaña Jamovi library → install.

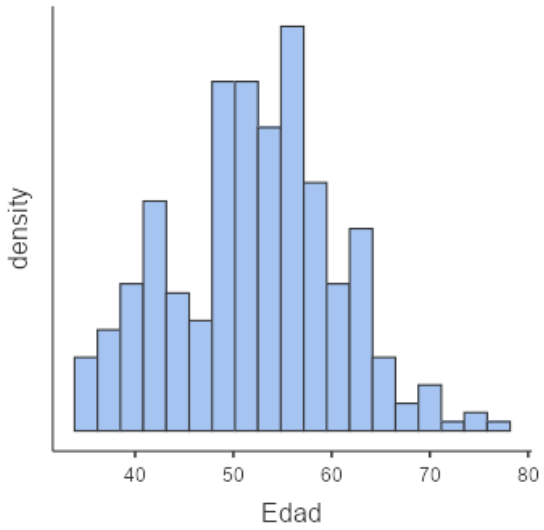
5.3.2 Variables cuantitativas

Para las variables cuantitativas se describe el patrón general de la distribución de las variables y permiten detectar ***outliers***.

5.3.2.1 Histograma

El histograma permite representar variables cuantitativas una vez agrupados los valores en clases. Representa las frecuencias y las clases de una variable cuantitativa. Las clases deben formar un sistema exhaustivo y excluyente.

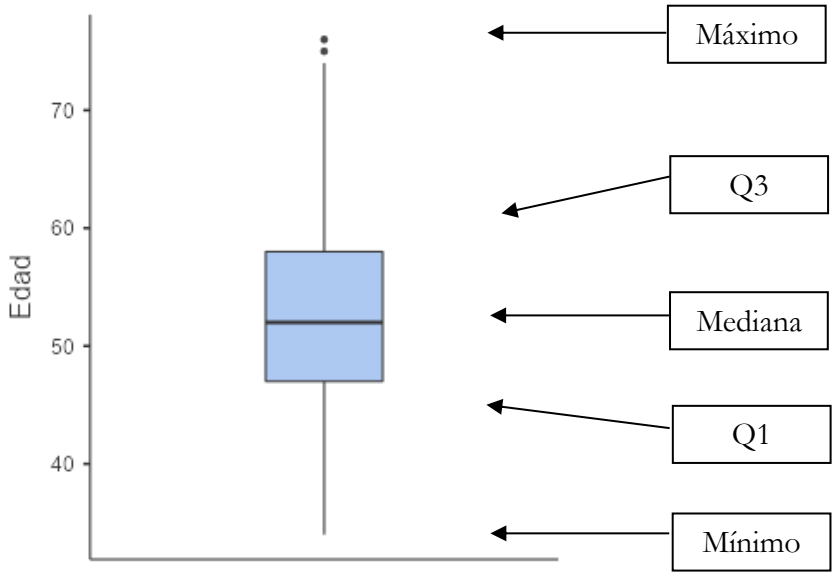
Al seleccionar la opción "**Histogram**" del menú "**Plots**" obtenemos la siguiente representación de la variable edad:



Al seleccionar la opción "Density" se añade al histograma la curva de densidad.

5.3.2.2 Diagrama de caja

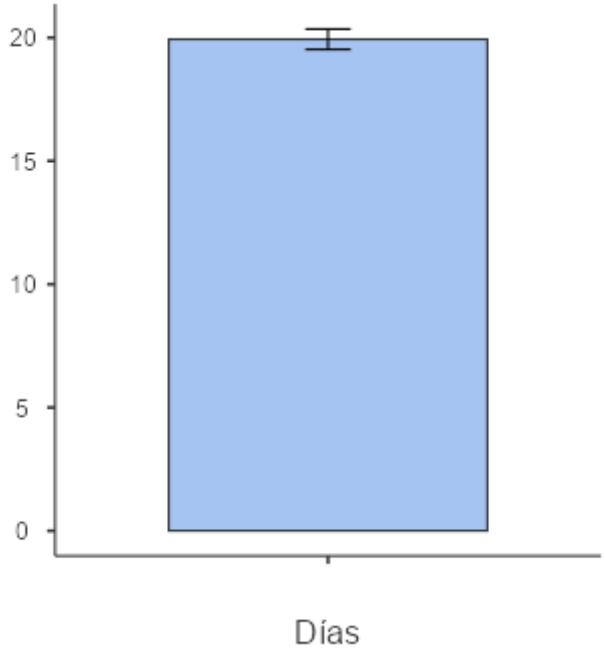
Un diagrama de caja es un gráfico ("**Box plot**") basado en los valores **mínimo**, **máximo** y los **cuartiles** (Q1, Q2 o mediana y Q3). Informa sobre la existencia de valores atípicos y la simetría de la distribución:



Al seleccionar la opción "Mean" se añade al gráfico el valor de la media.

5.3.2.3 Gráfico de barras

Al seleccionar la opción "**Bar Plot**" obtenemos la representación de un gráfico de barras que corresponde a la media, y un intervalo de confianza que corresponde a una desviación estándar de la media:



Ejercicio: Realizar un resumen descriptivo de cada una de las variables de la base de datos.

5.4 Medidas de asociación

El principal objetivo cuando se tienen dos o más variables está en medir la posible asociación entre ellas.

La relación Causa-Efecto

Muchas veces es fuente de interpretaciones erróneas de los resultados. En estadística, generalmente, se busca analizar si ciertos factores presentan un **efecto** sobre una determinada variable respuesta. No siempre se puede asegurar que la **causa** de este efecto sea el factor.

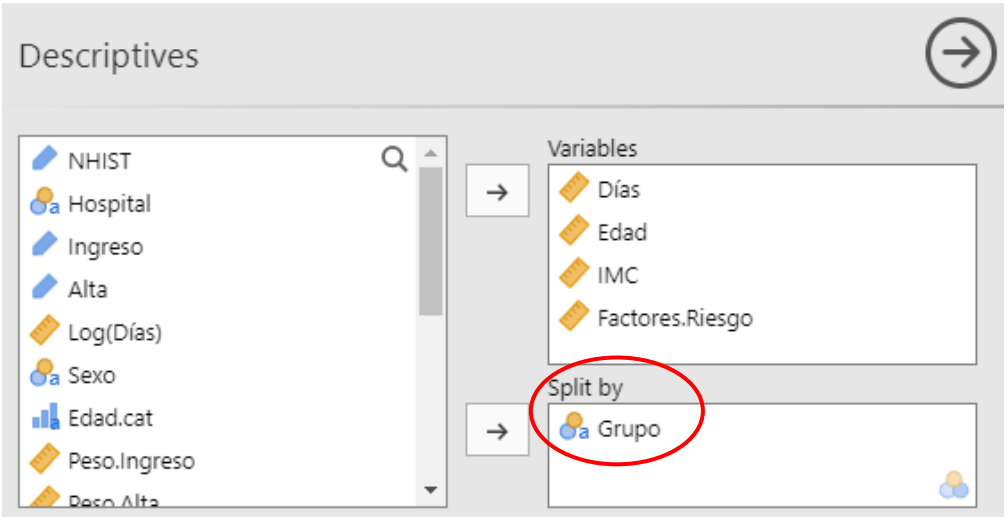
Establecer una relación causal no es nada simple. Raramente A es la causa de B. Fumar, por ejemplo, es sólo una **causa que contribuye** a desarrollar cáncer de pulmón; es una de las causas que aumenta la probabilidad de cáncer.

5.4.1 Una variable cuantitativa y una cualitativa

Para estudiar la relación entre una variable CUALITATIVA y una CUALITATIVA podemos calcular los estadísticos de resumen para cada una de las categorías de la variable cualitativa.

Ejemplo: Relación entre las variables cuantitativas 'Días', 'Edad', 'IMC', 'FR' y el 'Grupo'.

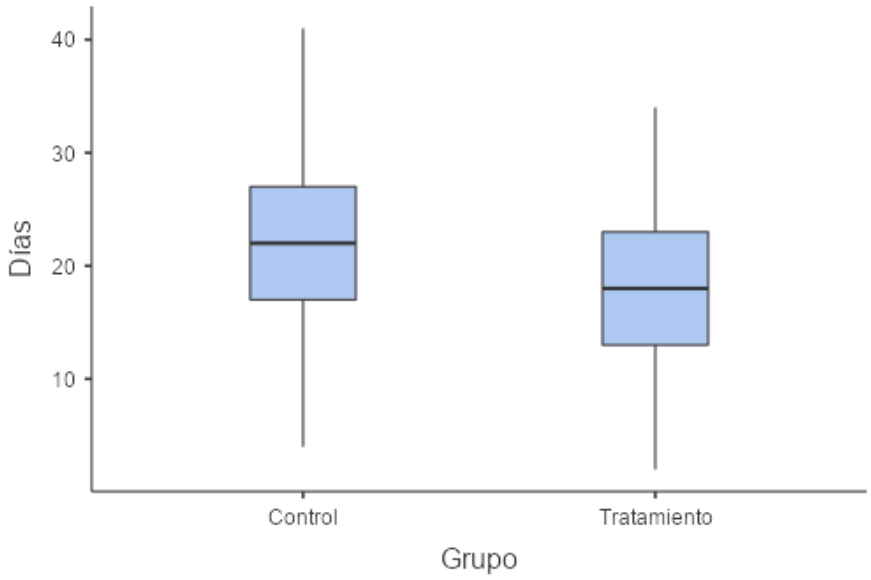
En el recuadro "**Split by**" del menú **Analyses → Exploration → Descriptives** podemos indicar una variable categórica para obtener los estadísticos de resumen para cada una de las categorías de esta variable:



Descriptives

	Grupo	Días	Edad	IMC	Factores.Riesgo
N	Control	162	162	162	162
	Tratamiento	163	163	163	163
Mean	Control	21.9	52.8	26.5	0.907
	Tratamiento	17.9	51.6	26.0	0.828
Median	Control	22.0	54.0	26.2	1.00
	Tratamiento	18	52	25.8	1
Standard deviation	Control	7.62	8.57	1.80	0.876
	Tratamiento	6.74	8.01	1.94	0.767
Minimum	Control	4	34	21.5	0
	Tratamiento	2	36	20.8	0
Maximum	Control	41	76	31.9	3
	Tratamiento	34	74	32.4	4

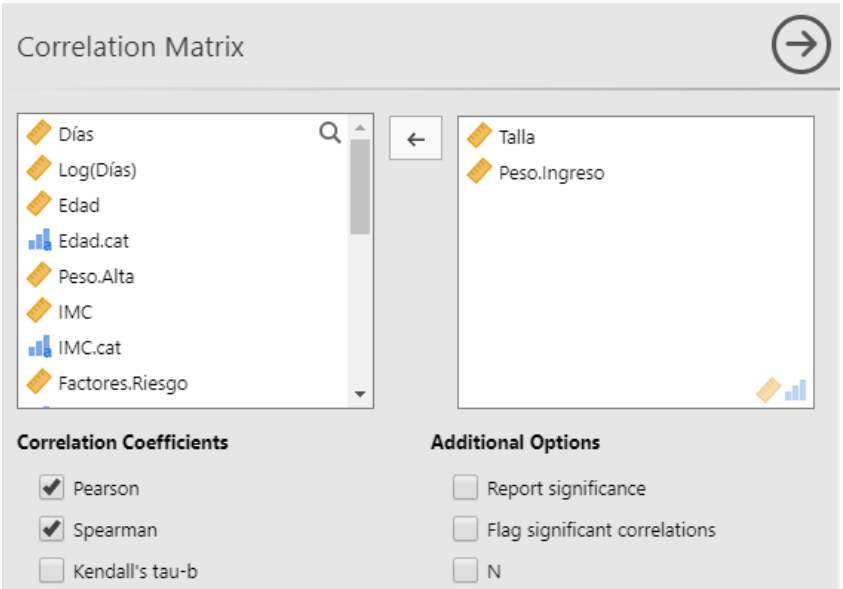
Representación gráfica: Diagrama de caja agrupado.



5.4.2 Dos variables cuantitativas

Para variables CUALITATIVAS la asociación entre ellas se analiza a partir del **Coeficiente de correlación** (menú **Regression → Correlation Matrix**).

Ejemplo: Relación entre la 'Talla' y el 'Peso.Ingreso'.



El estadístico que aparece seleccionado por defecto es el **coeficiente de correlación de Pearson** (**ρ**), que sirve para medir la asociación **lineal** entre las variables:

$$\rho = \frac{S_{XY}}{S_X S_X}$$

donde S_{xy} es la covarianza entre las variables.

Cuando las variables son discretas ordinales, podemos utilizar el coeficiente de correlación de **Spearman**. Este coeficiente se basa en los rangos, por lo que es un estadístico no paramétrico. Permite medir la relación monótona entre dos variables, y no se ve afectado por *outliers*. Se calcula como:

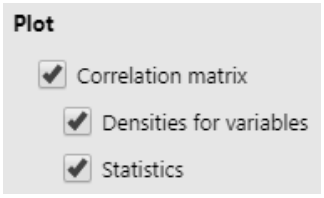
$$\rho = 1 - \frac{6\sum D_i^2}{n(n^2 - 1)}$$

donde D es la diferencia de rangos de los valores de X e Y, y n el tamaño muestral.

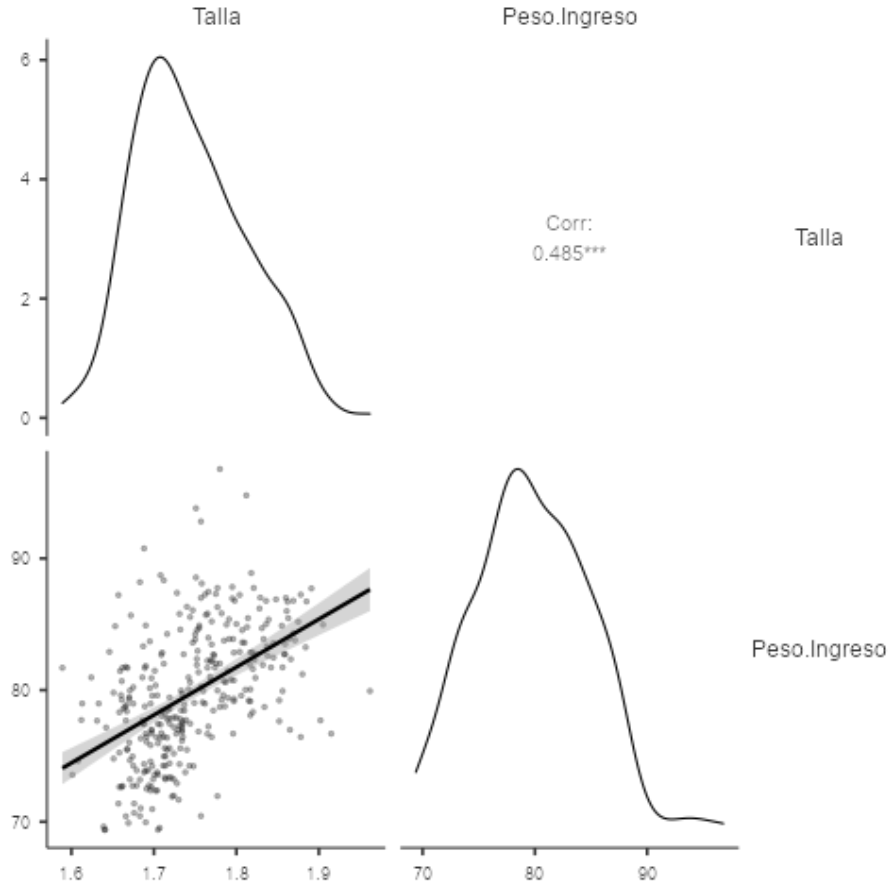
Correlation Matrix

		Talla
Peso.Ingreso	Pearson's r	0.485
	Spearman's rho	0.528

Gráficamente se pueden representar mediante el **Diagrama de dispersión**, que se obtiene seleccionando la casilla "**Correlation Matrix**" en "**Plot**":



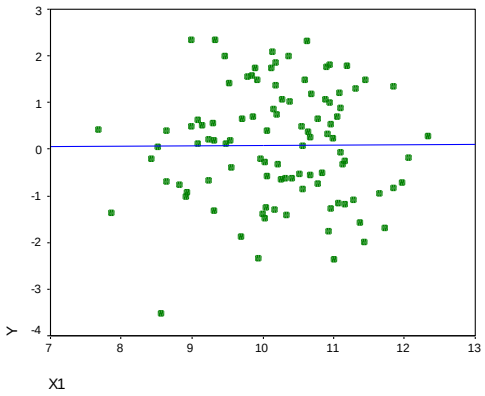
Observación: En el gráfico de dispersión la variable respuesta debe ir en el eje vertical (Y) y la variable explicativa en el eje horizontal (X).



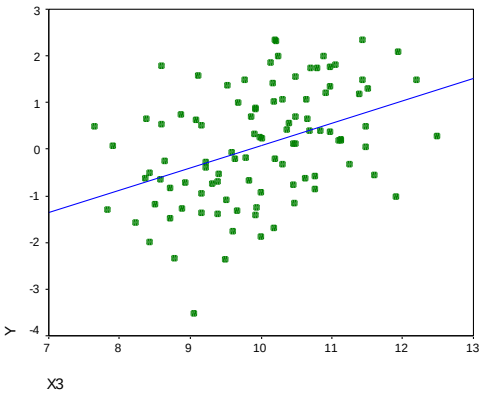
Al seleccionar las opciones "Density" y "Statistics" se añaden los gráficos de densidad de cada variable y los estadísticos.

Relación entre los valores del coeficiente de correlación y el gráfico de dispersión de las variables:

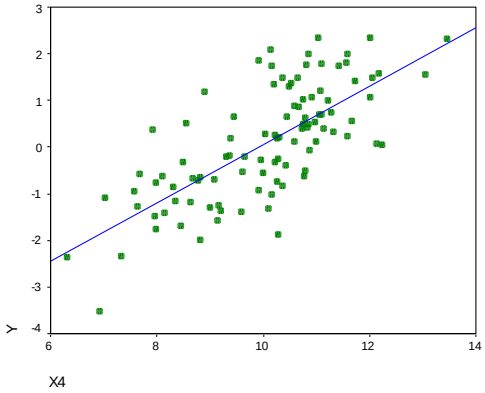
R = 0



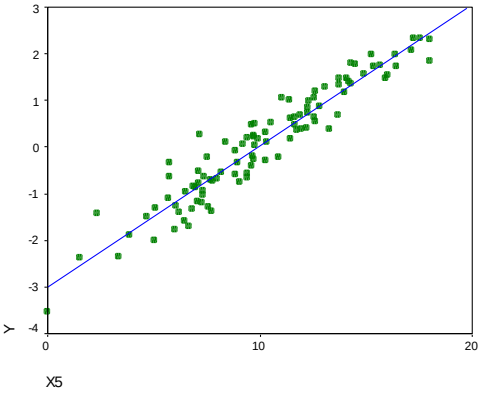
R = 0,20



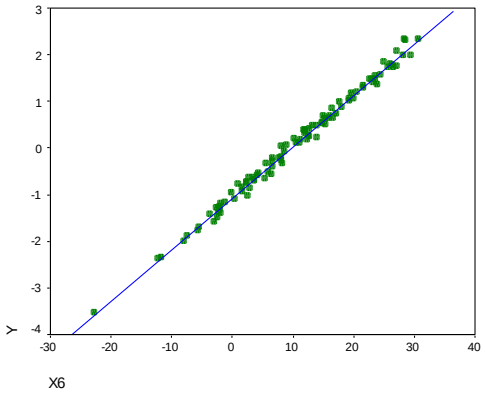
R = 0,50



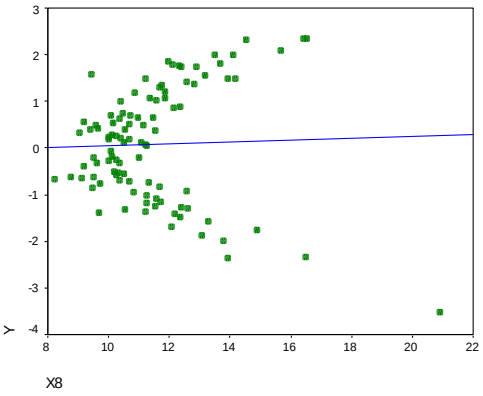
R = 0,90



R = 0,99



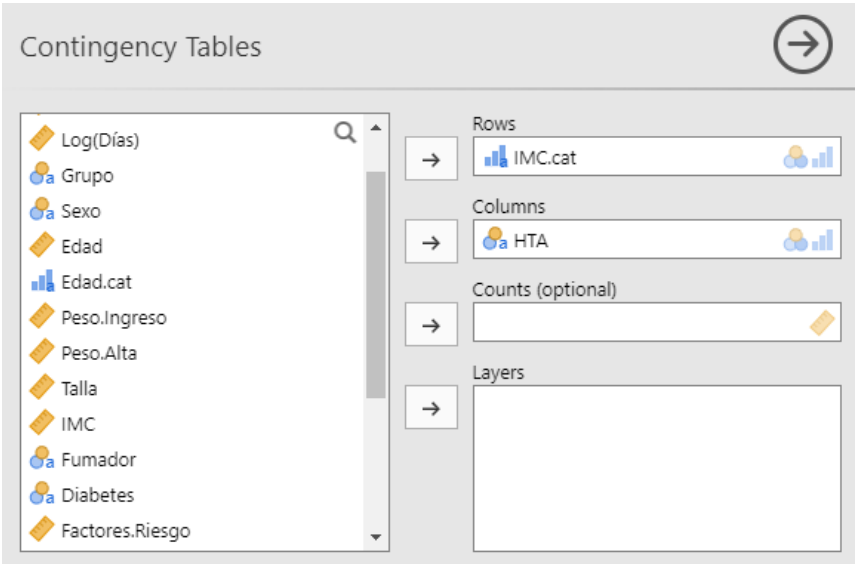
R = 0,01



5.4.3 Dos variables cualitativas

Para variables CUALITATIVAS la asociación entre ellas se analiza a partir de la **Tabla de Contingencia** (menú **Frequencies → Independent samples**).

Ejemplo: Relación entre las variables IMC (categórica) y la hipertensión arterial.



Observación: La opción "**Counts**" permitiría obtener la tabla de contingencia a partir de datos agregados. La opción "**Layers**" permitiría segmentar los resultados en función de una tercera variable, por ejemplo grupo de tratamiento.

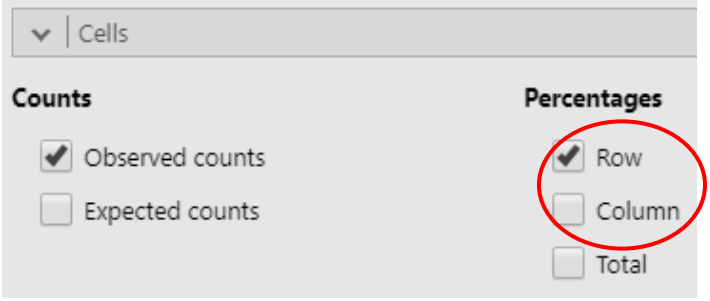
Contingency Tables

	HTA		
IMC.cat	**No**	**Sí**	**Total**
Normal	71	11	82
Sobrepeso	111	118	229
Obesidad	0	14	14
Total	182	143	325

A partir de las frecuencias observadas se definen los perfiles fila y columna:

- Frecuencia relativa conjunta = n_{ij} / n
- Perfil fila i = $\{n_{ij} / n_{i.}$ para j=1,..J$\}$
- Perfil columna j = $\{n_{ij} / n_{.j}$ para i=1,..I$\}$

En la pestaña "**Cells**" podemos seleccionar los perfiles fila ("**Row**") o columna ("**Column**"):

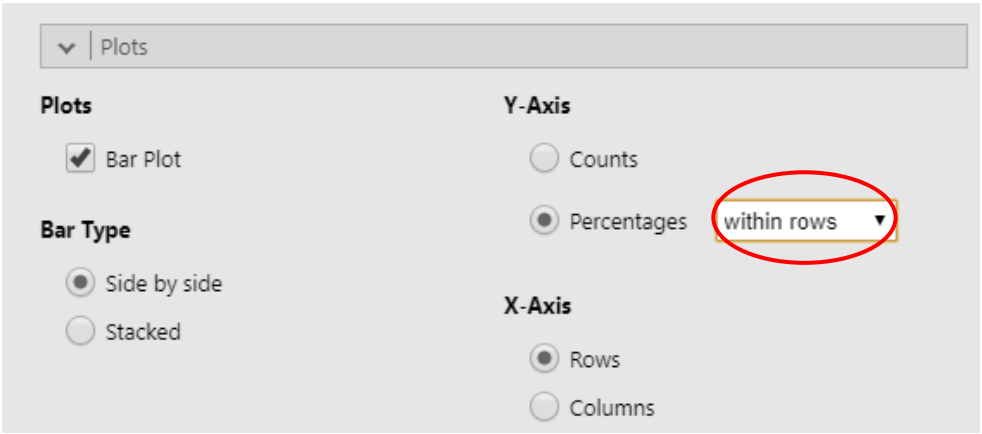


En este caso seleccionaremos los perfiles fila, para poder comparar el porcentaje de hipertensos según las categorías de IMC:

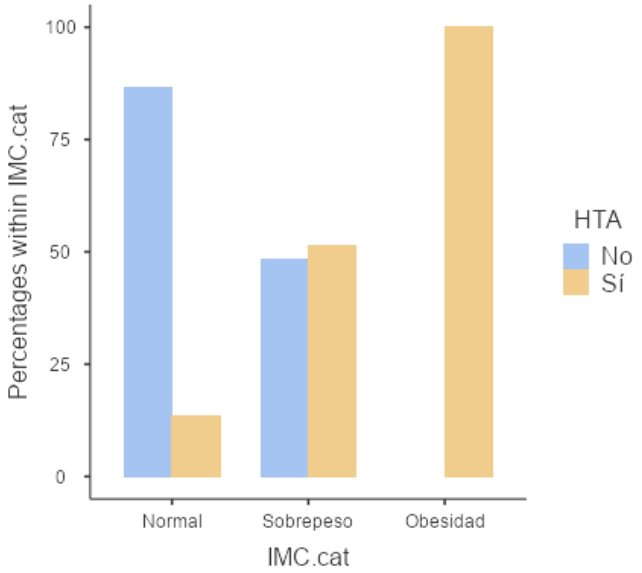
Contingency Tables

IMC.cat		HTA No	HTA Sí	Total
Normal	Observed	71	11	82
	% within row	86.6 %	13.4 %	100.0 %
Sobrepeso	Observed	111	118	229
	% within row	48.5 %	51.5 %	100.0 %
Obesidad	Observed	0	14	14
	% within row	0.0 %	100.0 %	100.0 %
Total	Observed	182	143	325
	% within row	56.0 %	44.0 %	100.0 %

Representación gráfica: gráfico de barras agrupado ("**Bar Plot**", seleccionando la variable de agrupación en la pestaña "**Plots**"):



En el eje Y se recomienda utilizar los mismos porcentajes seleccionados en la tabla de contingencia:



Ejercicio: Realizar un análisis descriptivo entre la variable 'IMC.cat' y el resto de factores de riesgo.

6 INFERENCIA PARA UNA POBLACIÓN

6.1 Introducción

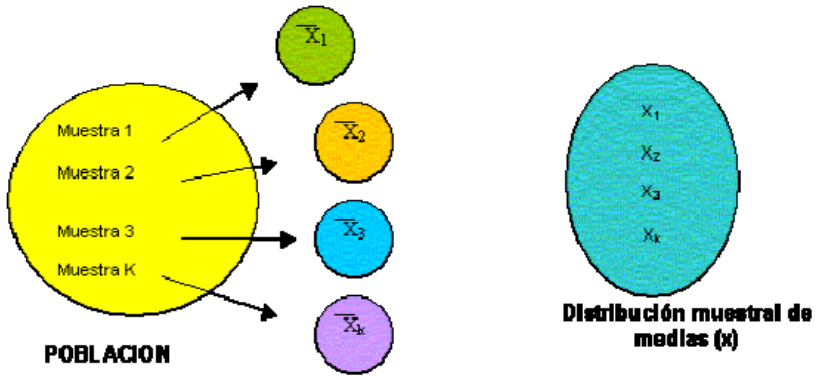
Después de llevar a cabo un análisis descriptivo de los datos el objetivo es poder generalizar los resultados para conjuntos más grandes de individuos, así como poder sacar conclusiones a partir de los datos.

La PROBABILIDAD permite calibrar el poder de nuestras conclusiones.

Población: Conjunto completo de individuos para los cuales se desea obtener información.

Muestra: Subconjunto de individuos de la población para los cuales realmente se obtiene la información de interés.

De una misma población se pueden obtener varias muestras diferentes, de tamaño *n*:



OBSERVACIÓN: La población está definida a partir de nuestro deseo de conocimiento.

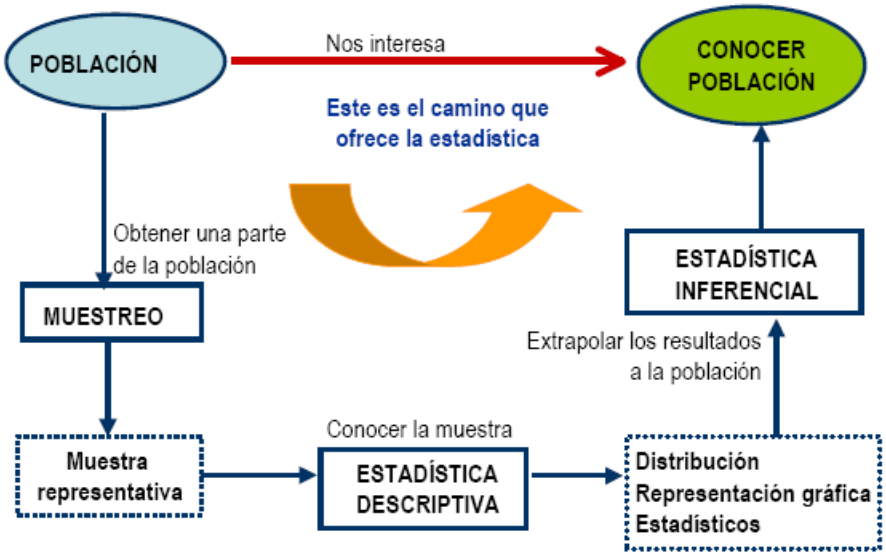
Los resultados obtenidos en una muestra serán **extrapolables** a la población de referencia si la muestra cumple dos características fundamentales:

- **Fiabilidad** (precisión): La fiabilidad de una muestra está vinculada a la precisión de sus resultados, es decir, al tamaño de muestra.
- **Validez** (sesgo): La validez de una muestra se refiere a que la muestra no presente sesgos, es decir errores de medida sistemáticos atribuibles a otra causa distinta del azar.

Un buen diseño del experimento permitirá controlar las posibles fuentes de sesgo y asegurar la validez del estudio.

- Una muestra representativa debe ser fiable y válida.
- No confundir muestra significativa con muestra representativa.
- Una muestra de 20.000 individuos no tiene por qué ser representativa de nada a no ser que se compruebe su validez, aunque seguramente sea suficientemente fiable.
- En una muestra de 10 individuos los resultados serán poco fiables, aunque seguramente la muestra sea suficientemente válida.

La **Estadística** es una herramienta que permite describir y cuantificar las evidencias observadas en una muestra intentando diferenciar entre lo que podría haber sucedido por azar y lo que podría atribuirse a otras causas (de interés).



Inferir significa sacar conclusiones de los datos teniendo en cuenta la variación debida al azar.

6.2 Variables aleatorias

Los datos que habitualmente se analizan provienen de un experimento aleatorio:

- Un **experimento aleatorio** o **estocástico** es aquel que bajo las mismas condiciones puede producir resultados diferentes, pero con una distribución regular de resultados para un número grande de repeticiones. Un ejemplo de experimento aleatorio es el lanzamiento de un dado.
- Un experimento es **no aleatorio** o **determinista** si bajo las mismas condiciones siempre conduce a un mismo resultado. Un ejemplo son las fórmulas físicas: Fuerza = Masa * Aceleración.

Las **variables aleatorias** son aplicaciones que transforman los resultados de un experimento aleatorio en números con el fin de poder realizar las operaciones más usuales, luego todos los resultados de un experimento aleatorio quedan recogidos en una variable aleatoria.

Antes de realizar cualquier inferencia estadística es necesario identificar la distribución de probabilidad (la forma) de la variable aleatoria que se pretende analizar.

Algunos instrumentos para ello son:

- Histograma, diagrama de caja, QQ Plot.
- Pruebas de ajuste a una distribución (**Test de Shapiro Wilk**).

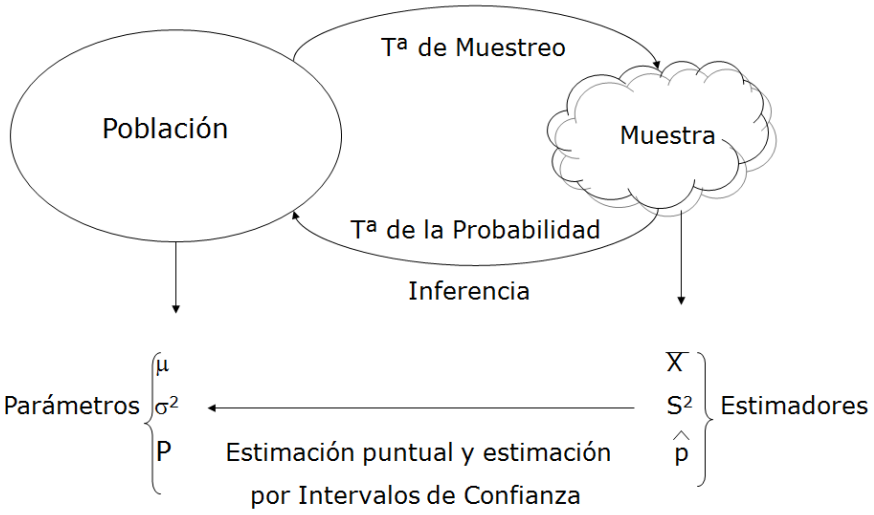
6.3 Estimación de parámetros

Un **parámetro** es un número que describe una característica de la población. En la práctica los valores de los parámetros son desconocidos.

Un **estadístico** es un número que se calcula a partir de los datos de una muestra de la población. En la práctica se utilizan los estadísticos para **estimar** los parámetros de la población.

Un **estimador** es cualquier función de una muestra, esto es, un estadístico es un estimador puntual.

Debemos observar que un estimador es una variable aleatoria mientras que una **estimación** es un valor concreto del estimador.



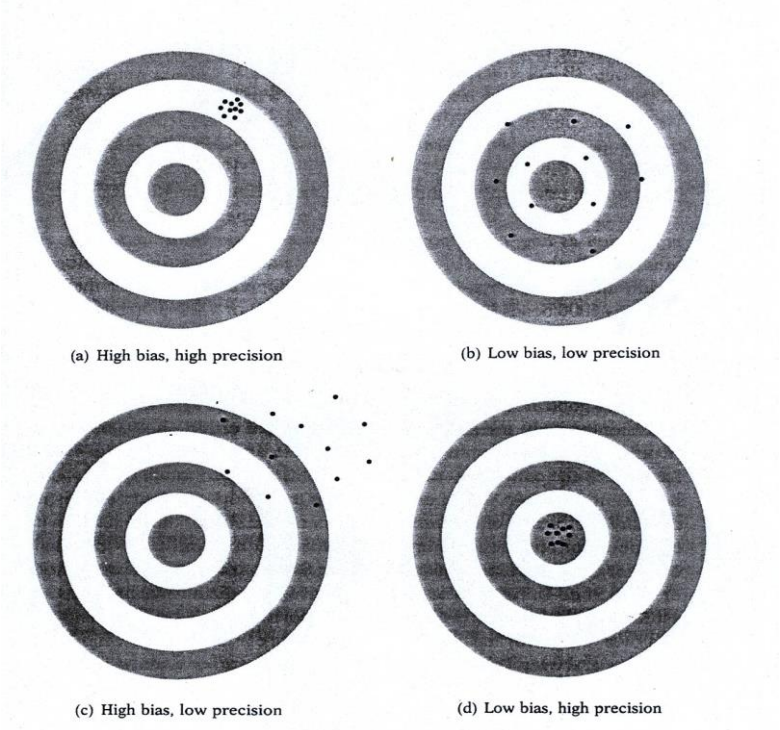
6.3.1 Estimación puntual

Una estimación puntual es el valor del estimador dada una muestra concreta. Los estimadores puntuales más frecuentemente utilizados son:

- Media muestral: $\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n}$
- Variancia muestral: $S^2 \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$
- Proporción: $\hat{p}$

A los estimadores básicamente se les requiere dos propiedades:

- **sin sesgo**, es decir que no se encuentren muy alejados del valor real del parámetro que estiman, y
- de **mínima varianza** posible, es decir que las distintas estimaciones estén próximas entre sí.



6.3.2 Intervalos de confianza

En inferencia estadística uno de los instrumentos más comunes para estimar el valor de un parámetro de la población son los **intervalos de confianza**.

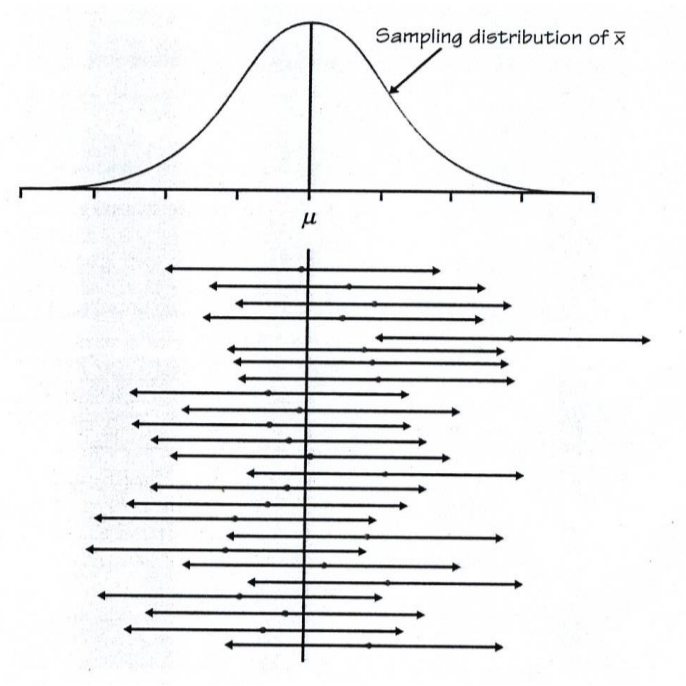
Un **intervalo de confianza del C%** para un parámetro es un intervalo de valores calculado a partir de los datos de la muestra utilizando un método que tiene una probabilidad **C** de que dicho intervalo contenga el verdadero valor del parámetro.

El parámetro poblacional pertenece al intervalo calculado con una confianza del C%.

La media muestral y la desviación estándar son buenos estimadores puntuales de la media y la desviación estándar de la población.

Dado que los datos son las observaciones de una variable aleatoria, estos estimadores son a la vez variables aleatorias. Por lo tanto, tienen una determinada distribución que en el caso de la media es la distribución Normal.

Para realizar inferencia estadística debemos interpretar los intervalos de confianza para un parámetro a partir del siguiente gráfico:



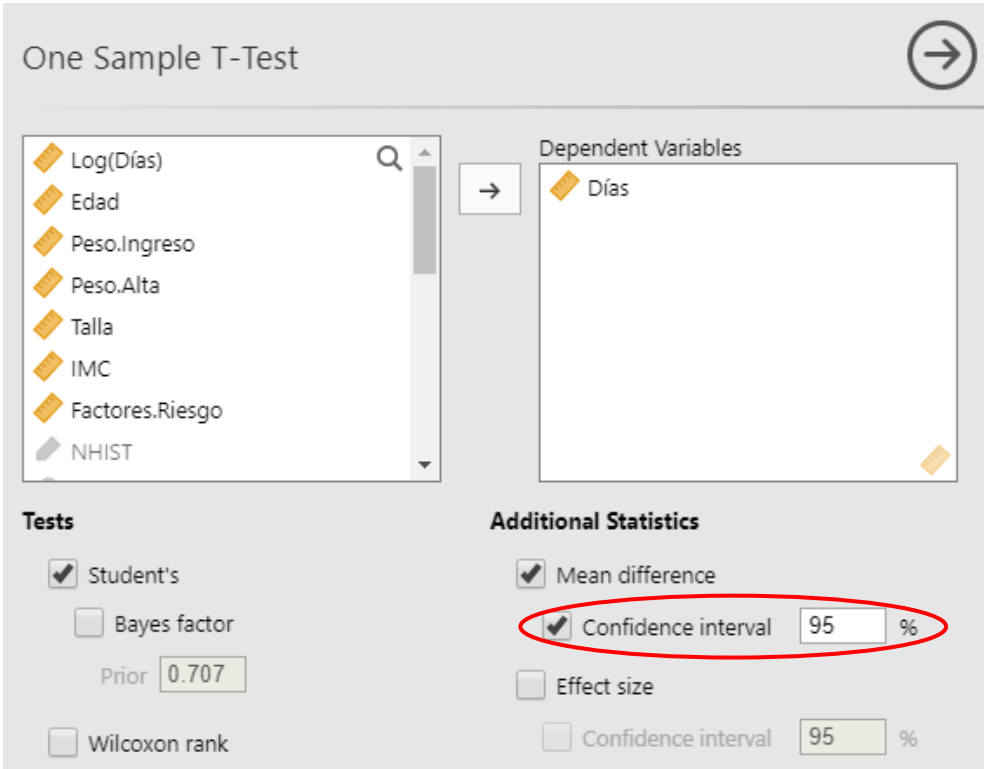
Si repetimos el experimento 100 veces o tomamos 100 muestras, en 95 ocasiones el parámetro pertenecerá al intervalo de confianza del 95% y en 5 ocasiones caerá fuera del intervalo.

Para una mejor comprensión de estos conceptos de la inferencia estadística básica se recomienda consultar los siguientes *statistical applets*, basados en el texto de Moore (2010):

https://www.macmillanlearning.com/studentresources/highschool/hsbridgepage/tps5e.html#anchor5

6.3.2.1 Intervalos de confianza para una media

Para obtener intervalos de confianza en **Jamovi** debemos seleccionar el menú **Analyses → T-Tests → One Sample T-Test** y seleccionar la opción "**Confidence interval**":



One Sample T-Test

			95% Confidence Interval	
		Mean difference	Lower	Upper
Días	Student's t	19.9	19.1	20.7

Desde el recuadro "**Confidence Level**" podemos indicar el nivel de confianza deseado.

6.3.2.1 Intervalos de confianza para una mediana

Para variables que no sigan una distribución normal (ver apartado 6.4.5) podemos obtener un intervalo de confianza para la mediana seleccionando el test **Wilcoxon rank**:

One Sample T-Test

			95% Confidence Interval	
		Mean difference	Lower	Upper
Días	Wilcoxon W	20.0	19.0	20.5

Observación: En la tabla aparece "Mean difference" pero corresponde a la mediana.

6.3.2.1 Intervalos de confianza para una proporción

Para obtener intervalos de confianza para una proporción seleccionaremos el menú **Analyses → Frequencies → 2 Outcomes**.

Ejercicio: Calcular los intervalos de confianza para las variables 'IMC.cat', 'Fumador', 'Diabetes', 'HTA' y 'Reingreso'.

6.4 Pruebas de hipótesis

Un segundo bloque de instrumentos para la inferencia estadística son las pruebas de hipótesis. Estas evalúan la evidencia de una afirmación sobre la población.

En estadística una afirmación sobre la población se plantea en forma de hipótesis de trabajo. Las dos hipótesis complementarias se llaman:

- **Hipótesis nula (H_0)**
- **Hipótesis alternativa o de investigación (H_1)**

La hipótesis nula corresponde a la hipótesis que creemos cierta por defecto y la alternativa corresponde a la hipótesis que se desea probar.

Las hipótesis hacen siempre referencia a los parámetros de la población.

Una prueba de hipótesis es un procedimiento que especifica:

1. Para qué valores muestrales la decisión será no rechazar la hipótesis nula.
2. Para qué valores muestrales la hipótesis nula será rechazada a favor de la alternativa.

P-valor: Probabilidad que, bajo H_0, el estadístico de contraste tome un valor al menos tan alejado como el realmente obtenido.

- Cuanto más pequeño sea el p-valor mayor es la evidencia en contra de H_0.
- Se rechazará la hipótesis nula si el p-valor es menor que el nivel de significación adoptado (en general 0,05).
- En un contraste de hipótesis, debemos rechazar o no la hipótesis nula a favor de la alternativa.

Deseamos que nuestra decisión sea correcta, pero a veces no lo será. Hay dos tipos de decisiones incorrectas:

Rechazar H_0 cuando de hecho es cierta: **error de tipo I**.
NO rechazar H_0 cuando realmente es cierta H_1: **error de tipo II**.

Observación: El error de tipo I = **nivel de significación** = α.

En siguiente cuadro resume los tipos de errores que se pueden cometer en un contraste de hipótesis:

Decisión basada en la muestra	Verdad acerca de la población		
		H_0 cierta	**H_1 cierta**
	Rechazo de H_0	Error de tipo I	Decisión correcta
	No rechazo de H_0	Decisión correcta	Error de tipo II

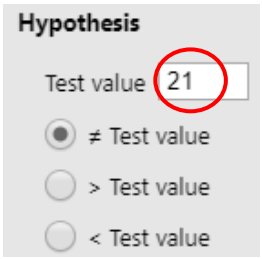
El error de Tipo I es más grave que el error de Tipo II.

6.4.1 Contraste de hipótesis para una media

La hipótesis que se contrasta es:

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$$

Para llevar a cabo un contraste de hipótesis para la media debemos volver al menú anterior y definir como valor de prueba el valor que deseamos contrastar:



One Sample T-Test

		Statistic	df	p
Días	Student's t	-2.57	324	0.011

Note. H_a population mean ≠ 21

Seleccionando la opción "**Effect size**" obtenemos el tamaño del efecto, una medida sobre la diferencia estandarizada entre el promedio de cada grupo:

$$\delta = \frac{\mu - \mu_0}{\sigma}$$

One Sample T-Test

		Effect Size
Días	Cohen's d	-0.143

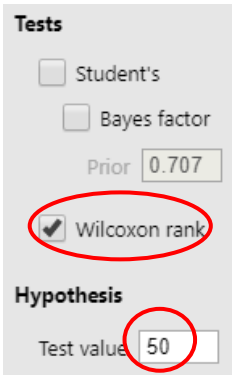
Observación: Se considera un efecto pequeño a partir de 0,1, mediano a partir de 0,3 y grande a partir de 0,5 (en valor absoluto).

6.4.2 Contraste de hipótesis para una mediana

La hipótesis que se contrasta es:

$$\begin{cases} H_0: \textit{mediana} = mediana_0 \\ H_1: \textit{mediana} \neq mediana_0 \end{cases}$$

Para llevar a cabo un contraste de hipótesis para la mediana debemos seleccionar la opción **Wilcoxon** y definir como valor de prueba el valor que deseamos contrastar en la pestaña "**Test value**":



One Sample T-Test

		Statistic	p
Edad	Wilcoxon W	31658	<0.001

Note. H_a population mean ≠ 50

Observación: En la tabla aparece "One Sample T-Test" pero corresponde al test no paramétrico de Wilcoxon. Podemos concluir que la mediana no es igual a 50 años

6.4.3 Contraste de hipótesis para una proporción

La hipótesis que se contrasta es:

$$\begin{cases} H_0: p = p_0 \\ H_1: p \neq p_0 \end{cases}$$

Ejercicio: Realizar un contraste para determinar si la proporción de hipertensos es del 40%.

6.4.4 Relación entre IC y Test de hipótesis

Cuando en una prueba estadística se pretende comparar dos medias o una media frente a un valor de referencia, el IC proporciona información paralela a la proporcionada por el test de hipótesis correspondiente.

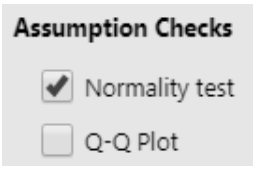
Es necesario que el nivel de confianza sea $1 - \alpha$, siendo α el nivel de significación del test aplicado.

- Si el IC no contiene el valor 21, se rechaza H_0: $\mu=21$.

Observación: Esta similitud es aplicable para pruebas T, o basadas en la distribución Normal.

6.4.5 Pruebas de normalidad

Para llevar a cabo un contraste de normalidad debemos seleccionar la opción **Normality test**.



Normality Test (Shapiro-Wilk)

	W	p
Días	0.994	0.268

Note. A low p-value suggests a violation of the assumption of normality

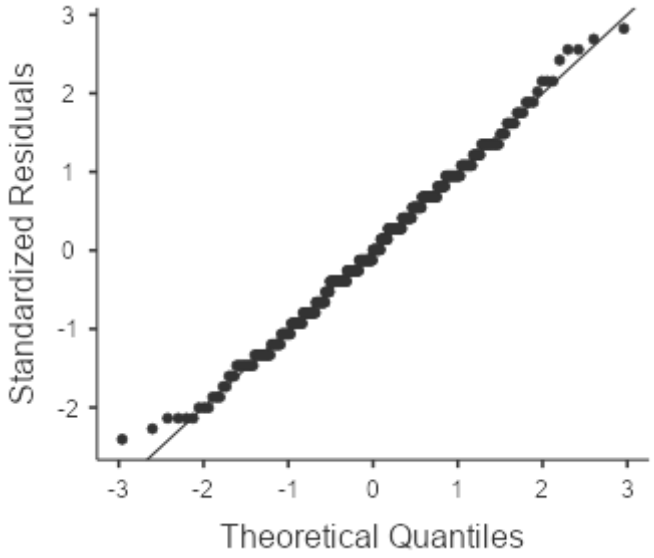
El contraste de hipótesis que realiza esta prueba es el siguiente:

$$\begin{cases} \mathbf{H_0: \text{la distribución es Normal}} \\ \mathbf{H_1: \text{la distribución NO es Normal}} \end{cases}$$

En este ejemplo hemos obtenido un nivel de significación (p-valor) de 0,268. Si fijamos el límite en 0,05 no rechazaríamos la hipótesis nula (podríamos considerar que la distribución de la variable "**Días**" es Normal).

Otra opción disponible es la de representar los datos mediante un **Q-Q Plot**. Este es un método gráfico para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la cual se ha obtenido la muestra y una distribución teórica. Se puede utilizar para contrastar la Normalidad de una variable: si al distribución de la variable es la misma que la distribución de comparación se obtendrá aproximadamente una línea recta.

En el caso de que se den desviaciones sustanciales de la linealidad, los estadísticos rechazan la hipótesis nula de similitud.



6.4.6 La sumisión de los investigadores al p-valor

La utilización sistemática del p-valor puede llevar a resultados engañosos.

EJEMPLO: Se quiere analizar la estancia en días de los turistas en Catalunya. En concreto se desea comparar las estancias de los europeos y los procedentes de países asiáticos. Un contraste en términos de las diferencias se plantea como:

$$\begin{cases} H_0: d = 0 \text{ (no hay diferencia)} \\ H_1: d \neq 0 \end{cases}$$

El p-valor del test estadístico resulta ser p=0,02, con lo que se concluye que hay diferencias. ¿Es suficiente?

Necesitamos medir el tamaño del efecto realizando un intervalo de confianza para la diferencia ya que podría ser, por ejemplo, que la diferencia se situara en el intervalo (0,5 - 1) o bien en el intervalo (10 - 15).

¿QUE ES UNA DIFERENCIA ESTADÍSTICAMENTE SIGNIFICATIVA?
(en un contraste de diferencias)

- Si se obtiene un p-valor inferior al nivel de significación al realizar el contraste, la diferencia es estadísticamente significativa.

- Si se obtiene un p-valor <0,05 al realizar el contraste, la diferencia no tiene por qué ser significativa.

- Si en un contraste se obtiene por ejemplo un p-valor=0,03 y en otro se obtiene un p-valor=0,42, no tiene por qué haber mayores diferencias entre grupos en el primer caso que en el segundo.

- Las diferencias pueden ser estadísticamente significativas, pero NO estadísticamente "muy" significativas, "ligeramente" significativas o "prácticamente" significativas.

- Recordar que una diferencia estadísticamente significativa implica "simplemente" que la diferencia no es nula.

- Para que una diferencia sea significativa, ésta debe ser relevante.

- En los resultados de un contraste SIEMPRE hay que presentar el p-valor y el Intervalo de Confianza de la diferencia para valorar su relevancia.

7 INFERENCIA PARA DOS POBLACIONES

7.1 Introducción

La Inferencia Estadística para dos poblaciones pretende generalizar los resultados y comparar los datos de una o diversas variables respuesta medidas en **dos muestras**, sin tener en cuenta otras variables (factores de riesgo).

Dos **muestras independientes** son aquellas para las cuales no existe ningún vínculo entre ellas. Provienen de poblaciones independientes.

Dos **muestras relacionadas** son aquellas que se refieren a la misma población y han medido la misma variable respuesta.

PLANTEAMIENTO DEL PROBLEMA

En primer lugar, el investigador debe identificar la naturaleza de las variables que desea estudiar. Es decir:

- **Variable Respuesta**: Distribución (continua, ordinal, categórica).
- **Variable Explicativa**: Número de grupos o niveles.

Así cómo la idoneidad del **tipo de prueba**: Homogeneidad Basal, grupos balanceados.

7.2 Muestras independientes

7.2.1 Comparar medias

Para comparar una variable respuesta entre dos muestras independientes cuando dicha variable sigue una **distribución normal** se utiliza la prueba **T de Student (T-Test) para muestras independientes**.

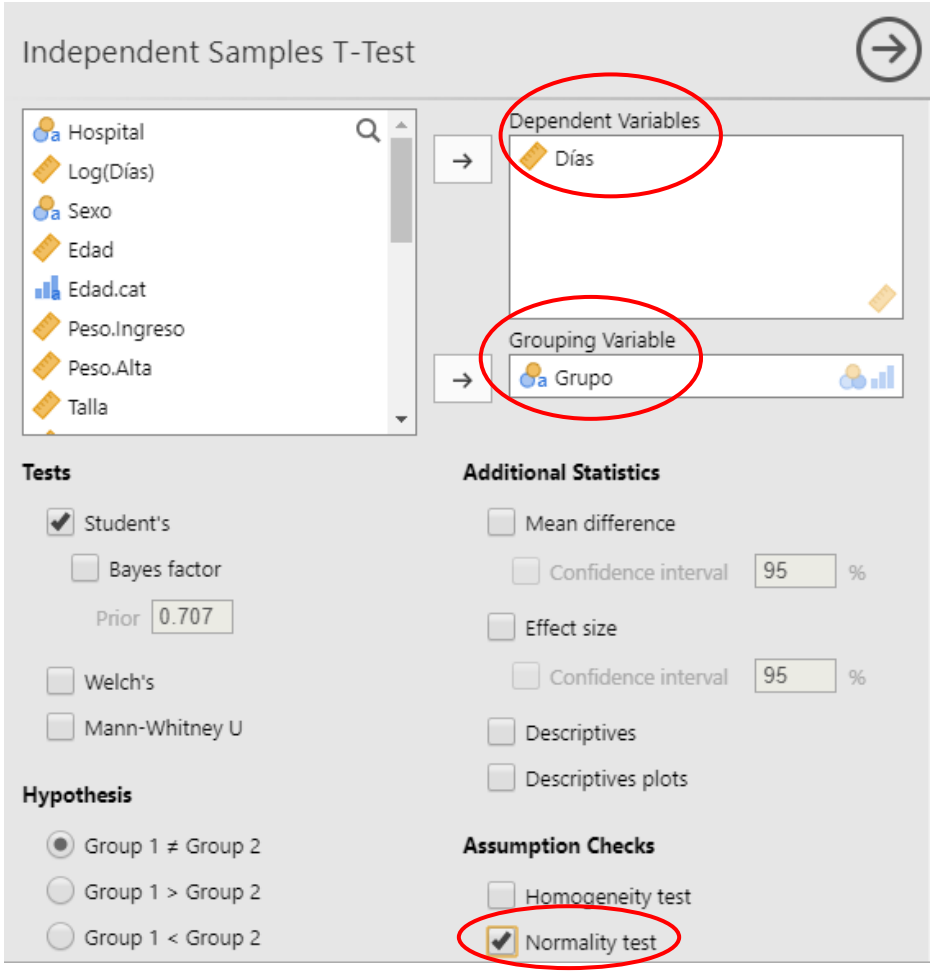
La hipótesis que contrasta es:

$$\begin{cases} H_0: \mu_1 = \mu_2 & \text{las medias son iguales} \\ H_1: \mu_1 \neq \mu_2 & \text{las medias son diferentes} \end{cases}$$

Ejemplo: Deseamos estudiar si hay diferencias entre los días de hospitalización en los pacientes del grupo control y tratamiento.

En primer lugar, debemos contrastar si podemos asumir que la distribución de la variable "**Días**" es Normal.

Para llevar a cabo estos contrates debemos ir al menú **Analyses → T-Tests → Independent Samples T-Test**:



Normality Test (Shapiro-Wilk)

	W	p
Días	0.994	0.221

Note. A low p-value suggests a violation of the assumption of normality

No se rechaza la hipótesis nula (p-valor > 0,05) por lo tanto podemos aceptar que la variable '**Días**' sigue una distribución Normal.

Para contrastar si hay diferencias entre los días de hospitalización entre los dos grupos utilizaremos el test "**Student's t-test**" (es el que viene seleccionado por defecto):

Independent Samples T-Test

							95% Confidence Interval	
		Statistic	df	p	Mean difference	SE difference	Lower	Upper
Días	Student's t	5.02	323	< .001	4.01	0.798	2.44	5.58

Observación: Al seleccionar la opción "**Confidence interval**" se obtiene el intervalo de confianza para la diferencia de medias.

Se observan diferencias estadísticamente significativas entre los días de hospitalización (en promedio) según el grupo de tratamiento (p-valor<0,001). La estancia es más larga en pacientes del grupo control.

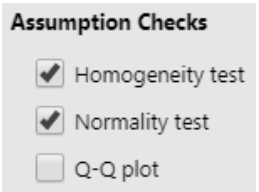
Observación: La prueba realizada considera que **las varianzas iguales** en los dos grupos. En caso de que las varianzas de los dos grupos fueran muy distintas se debería aplicar la corrección de "**Welch**".

7.2.2 Prueba de igualdad de varianzas

Para determinar si las varianzas son iguales podemos realizar el siguiente contraste de hipótesis:

$$\begin{cases} H_0: \sigma_1 = \sigma_2 & \text{Las variancias son iguales} \\ H_1: \sigma_1 \neq \sigma_2 & \text{Las variancias no son iguales} \end{cases}$$

Para llevar a cabo este contrate debemos seleccionar el test "**Homogeneity test**" en el apartado "**Assumption Checks**":



Homogeneity of Variances Test (Levene's)

	F	df	df2	p
Días	1.51	1	323	0.220

Note. A low p-value suggests a violation of the assumption of equal variances

Observación: Las pruebas de igualdad de varianzas son sensibles a distribuciones NO Normales, incluso para muestras grandes. Es por este motivo que **se puede utilizar siempre el test que considera varianzas distintas para comparar dos medias** (los resultados de este test son válidos tanto si las varianzas son iguales como si no).

7.2.3 Comparar medianas

A la práctica, muchas veces no podemos aceptar la hipótesis de normalidad en los datos. En esta situación se puede hacer uso de métodos **no paramétricos** que no suponen ninguna hipótesis sobre la distribución de los datos.

Para comparar una variable respuesta entre dos muestras independientes cuando dicha variable es continua (no normal) o bien ordinal se utiliza la prueba de **suma de rangos Wilcoxon** (también llamada prueba U de Mann-Whitney o prueba de Mann-Whitney-Wilcoxon).

La hipótesis que contrasta es:

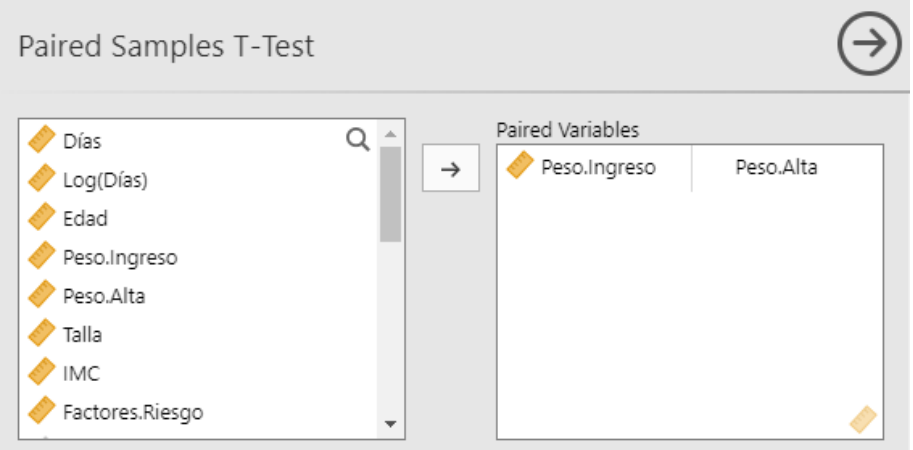
$\begin{cases} H_0: \text{La mediana del grupo 1 es igual a la mediana del grupo 2} \\ H_1: \text{La mediana del grupo 1 NO es igual a la mediana del grupo 2} \end{cases}$

Ejercicio: Estudiar si hay diferencias entre la edad de los pacientes según el sexo.

7.3 Muestras relacionadas

Para comparar una variable respuesta entre dos muestras relacionadas cuando dicha variable sigue una distribución normal se utiliza la prueba "**Paired t-test**", y para realizar una prueba no paramétrica "**Paired-samples Wilcoxon test**". Ambas se encuentran en el menú **Analyses → T-Tests → Paired Samples T-Test**.

Ejemplo: Deseamos contrastar si hay diferencias entre el peso al ingreso y el peso en el momento del alta.



En primer lugar, deberemos realizar un test de normalidad para ver si la distribución es Normal (seleccionar "**Normality test**"):

Normality Test (Shapiro-Wilk)

			W	p
Peso.Ingreso	-	Peso.Alta	0.983	< .001

Note. A low p-value suggests a violation of the assumption of normality

Al no seguir una distribución Normal la comparación entre el peso al ingreso y el peso al alta debería hacerse mediante un test no paramétrico ("**Wilcoxon Rank**"):

Paired Samples T-Test

							95% Confidence Interval	
			Statistic	p	Mean difference	SE difference	Lower	Upper
Peso.Ingreso	Peso.Alta	Wilcoxon W	50199	< .001	3.47	0.181	3.11	3.84

Se observan diferencias estadísticamente significativas (p-valor<0,01). El peso al ingreso y el peso al alta son distintos. En promedio han reducido 3,47 kg.

8 INFERENCIA PARA *K* POBLACIONES

8.1 Introducción

La Inferencia Estadística para *k* poblaciones generaliza los métodos estadísticos vistos en el apartado anterior.

Se dispone de una variable respuesta (continua, categórica, ordinal) y una variable explicativa que define *k* grupos o categorías.

8.2 Variables cuantitativas: comparar medias

8.2.1 Muestras independientes: prueba ANOVA

El análisis de la varianza (ANOVA: **An**alysis **o**f **Va**riance) es un procedimiento estadístico que tiene como objetivo descomponer la variabilidad observada en un ensayo experimental en función de los posibles factores que han podido influir en el resultado.

Esta técnica se utiliza en las situaciones en las que se desea analizar una **variable cuantitativa** medida bajo ciertas condiciones experimentales identificadas por uno o más **factores cualitativos**. Cada factor identifica 2 o más situaciones experimentales complementarias, y por lo tanto distingue grupos o niveles.

Cuando hay un único factor estudiado, el análisis recibe el nombre de **ANOVA de un factor**.

La prueba ANOVA de un factor generaliza la prueba T para dos muestras independientes.

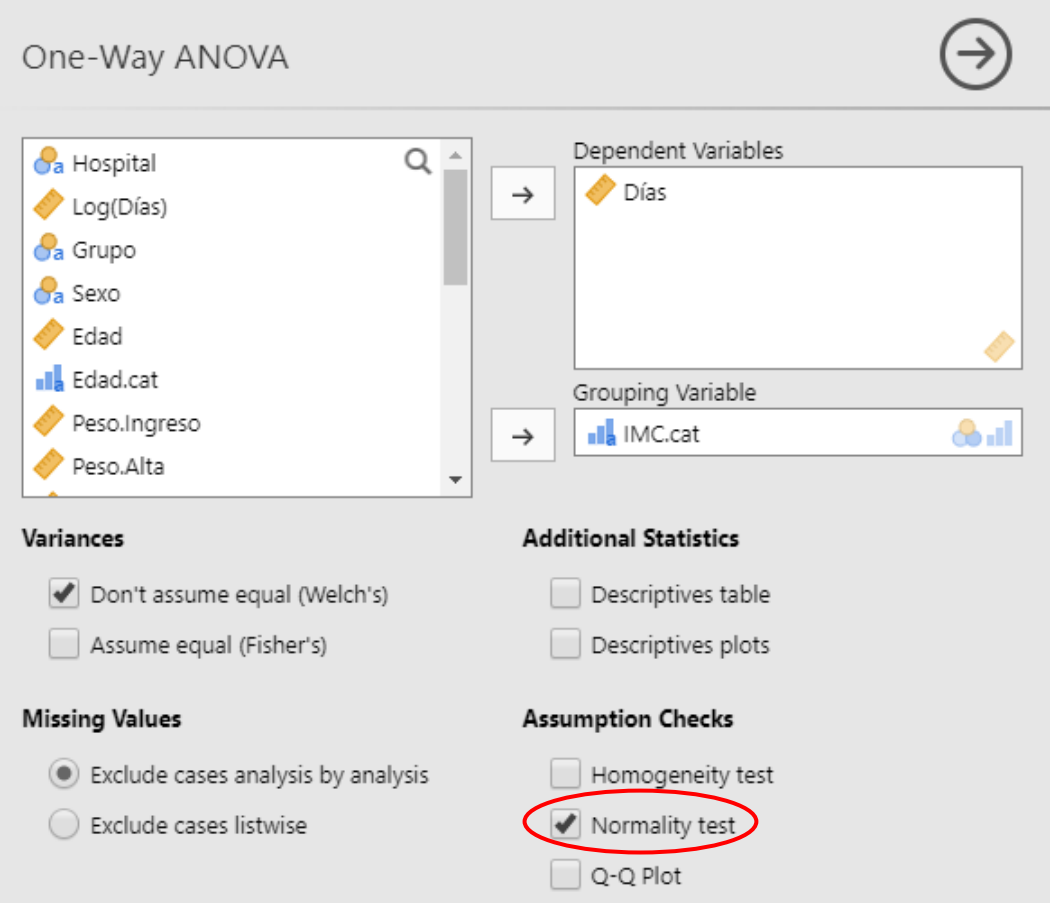
La hipótesis que contrasta es:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \ldots = \mu_k & \text{las medias son iguales} \\ H_1: \text{Al menos una de las medias no es igual al resto} \end{cases}$$

La prueba ANOVA se sustenta en los supuestos de **normalidad**, **igualdad de variancias**, independencia y aleatoriedad.

Ejemplo: Deseamos estudiar si existen diferencias entre la estancia media de los pacientes según el IMC (3 categorías).

Como en el caso de comparar dos medias, en primer lugar debemos contrastar si podemos asumir que la distribución de la variable "**Días**" es Normal dentro de cada categoría de "**IMC.cat**". Para ello, seleccionamos la prueba de normalidad de Shapiro-Wilk en el menú **Analyses → One-Way ANOVA**.



Normality Test (Shapiro-Wilk)

	W	p
Días	0.994	0.234

Note. A low p-value suggests a violation of the assumption of normality

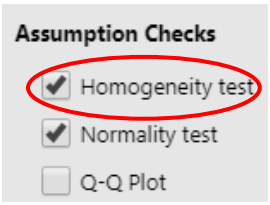
No se rechaza la hipótesis de normalidad (p-valor > 0,05).

Prueba de homogeneidad de varianzas

Para determinar si las varianzas son iguales podemos realizar el siguiente contraste de hipótesis:

$$\begin{cases} H_0: \text{Las variancias son iguales en todos los grupos} \\ H_1: \text{Al menos un grupo presenta una variabilidad diferente al resto} \end{cases}$$

Para realizar este contraste debemos seleccionar la opción "**Homogeneity test**" en el apartado "**Assumption Checks**":



Este contraste se realiza mediante el test de **Levene**:

Homogeneity of Variances Test (Levene's)

	F	df1	df2	p
Días	1.11	2	322	0.330

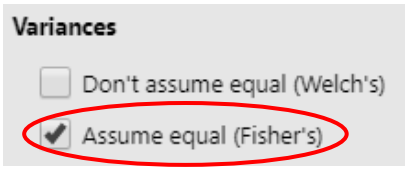
No se rechaza la igualdad de variancias (p-valor > 0,05). Luego, existe homogeneidad de varianzas en los grupos.

Ejemplo (continuación):

La hipótesis que deseamos contrastar en la prueba ANOVA es:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \text{Al menos una de las medias no es igual al resto} \end{cases}$$

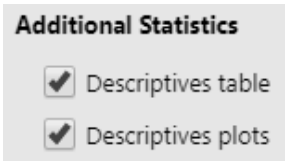
La prueba que viene seleccionada por defecto es la prueba ANOVA con la corrección de Welch (varianzas no iguales). Como en nuestro ejemplo hemos visto que las varianzas se pueden considerar iguales, podemos seleccionar el test ANOVA para varianzas iguales:



One-Way ANOVA (Fisher's)

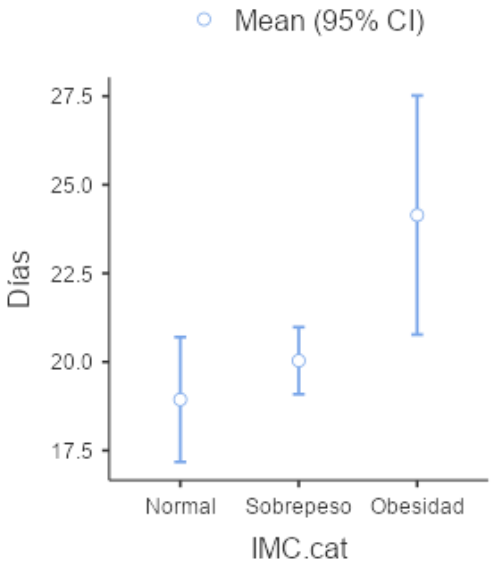
	F	df1	df2	p
Días	3.02	2	322	0.050

Dado el p-valor obtenido no es superior al nivel de significación (0,05), por lo que podemos rechazar la hipótesis nula. Seleccionando los estadísticos adicionales podemos visualizar cuáles son estas diferencias:



Group Descriptives

	IMC.cat	N	Mean	SD	SE
Días	Normal	82	18.9	8.02	0.885
	Sobrepeso	229	20.0	7.27	0.480
	Obesidad	14	24.1	5.84	1.561



8.2.2 Comparaciones múltiples 2 a 2

Hemos visto que el procedimiento ANOVA permite determinar si existen diferencias entre más de dos grupos, pero no informa sobre qué grupo o grupos son los que difieren. Por ello, tras la realización de la prueba ANOVA es interesante realizar las llamadas comparaciones múltiples a posteriori o 2 a 2.

Las comparaciones múltiples consisten en contrastar simultáneamente todas las parejas dos a dos que se puedan dar.

Las hipótesis que se contrastan son:

$$\begin{cases} H_0: \mu_1 = \mu_2 & \text{las medias son iguales} \\ H_1: \mu_1 \neq \mu_2 & \text{las medias no son iguales} \end{cases}$$

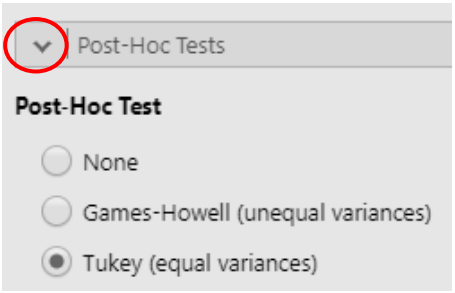
$$\begin{cases} H_0: \mu_1 = \mu_k & \text{las medias son iguales} \\ H_1: \mu_1 \neq \mu_k & \text{las medias no son iguales} \end{cases}$$

$$\vdots$$

$$\begin{cases} H_0: \mu_{k-1} = \mu_k & \text{las medias son iguales} \\ H_1: \mu_{k-1} \neq \mu_k & \text{las medias no son iguales} \end{cases}$$

La realización de todas las comparaciones 2 a 2 conduce habitualmente a un elevado número de comparaciones. Dichas comparaciones no son independientes las unas de las otras y por ello es necesario aplicar **correcciones por multiplicidad de contrastes** para garantizar que el nivel de significación conjunto no sea superior al 5%.

Para obtener los contrastes múltiples hay que activar la opción "**Post-Hoc Test**". Seleccionaremos "Tukey" o bien "Games-Howell" según las varianzas sean iguales o no:



Tukey Post-Hoc Test – Días

		Normal	Sobrepeso	Obesidad
Normal	Mean difference	—	-1.10	**-5.20***
	p-value	—	0.485	**0.042**
Sobrepeso	Mean difference		—	-4.11
	p-value		—	0.111
Obesidad	Mean difference			—
	p-value			—

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

Las comparaciones múltiples (Tukey Post-Hoc) indican que las diferencias entre los grupos 'Peso normal' y 'Obesidad' son estadísticamente significativas.

8.2.3 Inferencia no paramétrica: Prueba de Kruskal-Wallis

A la práctica, muchas veces no podemos aceptar la hipótesis de normalidad en los datos. En estas situaciones se puede hacer uso de métodos no paramétricos, que no suponen ninguna hipótesis sobre la distribución de los datos.

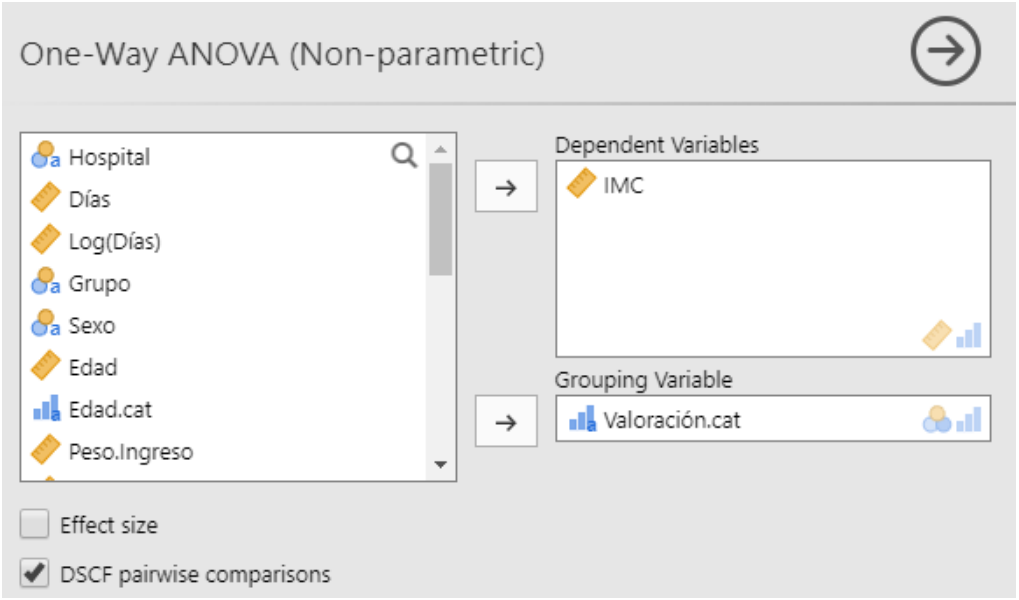
Para comparar una variable respuesta entre *k* muestras independientes cuando dicha variable es continua (no-normal) o bien ordinal se utiliza la prueba de **Kruskal-Wallis**.

La hipótesis que contrastan es:

$$\begin{cases} H_0: \text{La } \underline{\text{mediana}} \text{ de todos los grupos es igual} \\ H_1: \text{Al menos una de las } \underline{\text{medianas}} \text{ no es igual al resto} \end{cases}$$

Este test se encuentra en el menú **Analyses → Non-Parametric → One-Way ANOVA (Kruskal-Wallis)**.

Ejemplo: Deseamos estudiar si existen diferencias entre el índice de masa corporal según la valoración de salud (3 categorías).



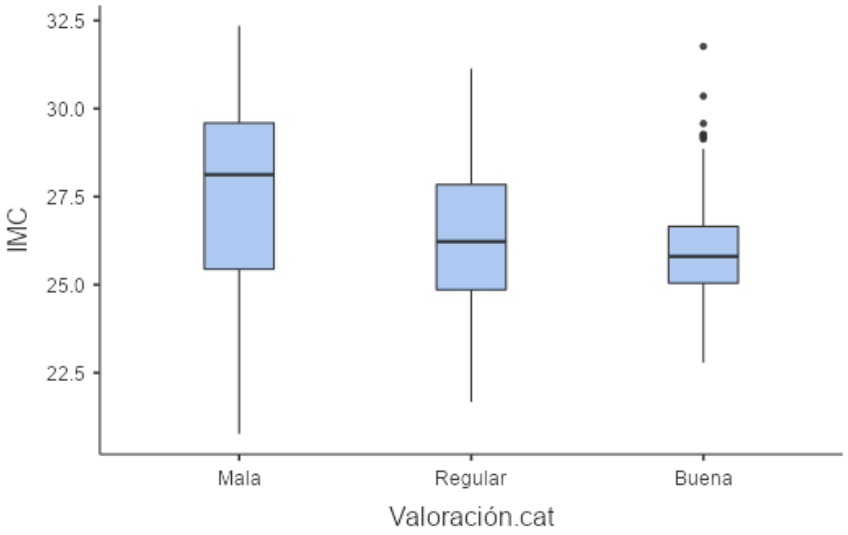
Resultados:

Kruskal-Wallis

	χ^2	df	p
IMC	12.1	2	0.002

Dado el p-valor obtenido, se rechaza la hipótesis nula. Existen diferencias entre las medianas de IMC según la valoración de salud.

Observación: Este menú no ofrece opciones para añadir estadísticos descriptivos. Estos se podrían pedir desde el menú **Exploration → Descriptives**, o desde el menú **ANOVA**.



Para determinar entre qué grupos se encuentras estas diferencias, podemos solicitar las comparaciones 2 a 2 ("**DSCF pairwise comparisons**"):

Pairwise comparisons - IMC

		W	p
Mala	Regular	-2.86	0.107
Mala	Buena	-4.70	0.003
Regular	Buena	-2.44	0.195

Las diferencias se encuentran entre las valoraciones de salud 'mala' y 'buena'.

9 TABLAS DE CONTINGENCIA

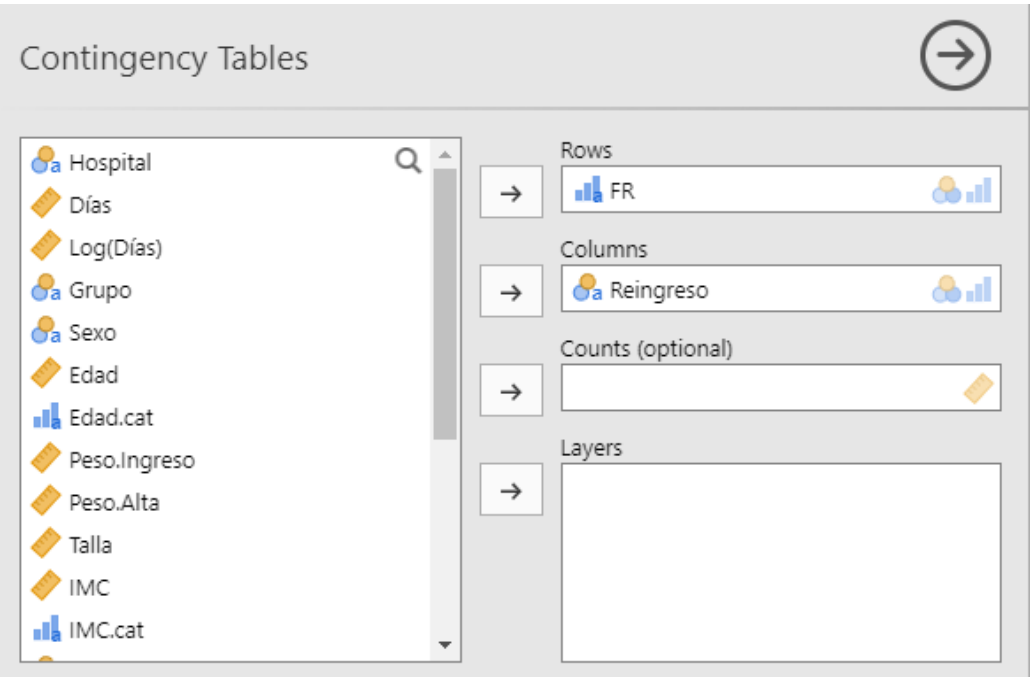
Para comparar una variable respuesta entre dos o más muestras independientes cuando dicha variable es categórica se utiliza la **prueba de χ^2**. En el caso de tener más de un 20% de las casillas con pocas observaciones (valor esperado inferior a 5) se recomienda utilizar la prueba **exacta de Fisher**.

La hipótesis que contrasta es:

- H_0: La variable respuesta es independiente de la variable explicativa (los grupos son homogéneos).
- H_1: La variable respuesta NO es independiente de la variable explicativa (los grupos no son homogéneos).

Ejemplo: Deseamos estudiar si hay relación entre los factores de riesgo (FR) y el reingreso.

Para llevar a cabo dicha prueba seleccionamos **Analyses → Contingency tables → Independent Samples**:



El test que realiza por defecto es el test Chi cuadrado:

χ^2 Tests

	Value	df	p
χ^2	8.61	2	0.013
N	325		

Se observan diferencias estadísticamente significativas (p-valor=0,013). A partir de la tabla de contingencia (perfiles fila o columna) podemos ver cuáles son las categorías que presentan mayores diferencias. En este caso seleccionaremos los perfiles fila (ver apartado 5.4.3):

Contingency Tables

		Reingreso		
FR		**No**	**Sí**	**Total**
0	Observed	62	58	120
	% within row	51.7 %	48.3 %	100.0 %
1	Observed	50	90	140
	% within row	35.7 %	64.3 %	100.0 %
2+	Observed	22	43	65
	% within row	33.8 %	66.2 %	100.0 %
Total	Observed	134	191	325
	% within row	41.2 %	58.8 %	100.0 %

A medida que aumentan los factores de riesgo, el porcentaje de pacientes que reingresa también aumenta.

En caso de que hubiéramos obtenido más del 20% (en este caso 2) casillas con un valor esperado inferior a 5, tendríamos que utilizar el **Test de razón de verosimilitud** (Likelihood ratio) o bien el **Test exacto de Fisher** (Fisher's exact test) de la pestaña "**Tests**".

Observación: El valor esperado se puede obtener seleccionando la opción "**Expected counts**" en el desplegable "**Cells**".

10 RESUMEN METODOLÓGICO

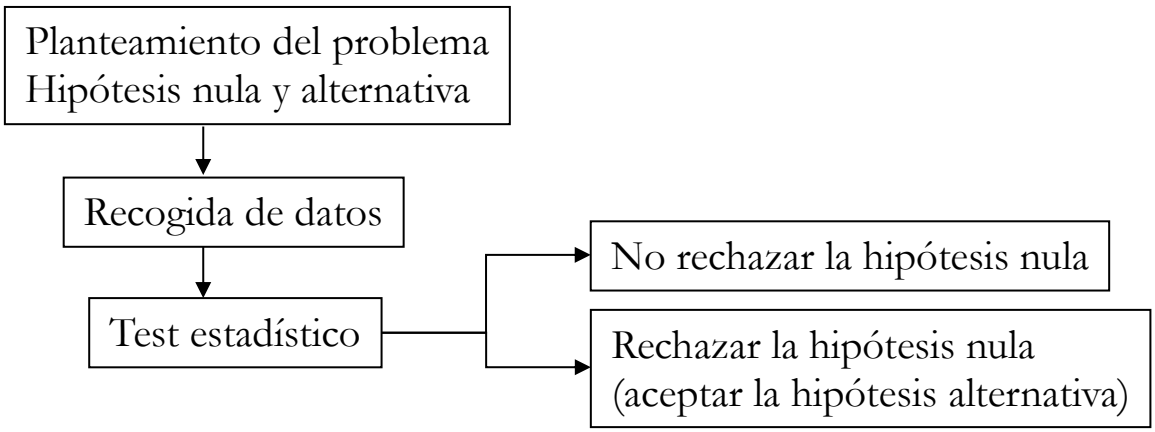
Los datos (**variables**) son características observables de los **individuos** de una población. Pueden ser:

- **CUALITATIVAS o CATEGÓRICAS**: etiquetas que representan el grupo o categoría a la cual pertenece un individuo.
- **CUANTITATIVAS**: valores numéricos para los que tiene sentido realizar aritmética.

En estadística, las variables también las clasificamos en función del papel que tienen dentro del análisis de un determinado proyecto:

- **Variable Respuesta**: variable que queremos explicar en el análisis.
- **Variables Explicativas**: variables que explican la variable respuesta.

Resumen de una prueba de hipótesis



¿Cómo determinar qué prueba es la idónea?

Variable respuesta **categórica** y variable explicativa **categórica,** ambas con dos o más categorías:

- En general, **prueba χ^2**.
- Si el número de casillas de la tabla de contingencia con frecuencia esperada < 5 es superior al 25 %: **Test de razón de verosimilitud** o **Exacto de Fisher**.

Variable respuesta **continua** y variable explicativa **categórica** (**2 grupos**):

- Si la distribución de la respuesta en cada grupo es Normal: **T-Test**.
- Si la distribución de la respuesta en cada grupo es Normal y no hay homogeneidad de varianzas: **T-Test** *con la corrección de Welch*.
- Si la distribución no es Normal pero es continua: **Test de Wilcoxon**.

Variable respuesta **continua** y variable explicativa **categórica** (k grupos):

- Si la distribución de la respuesta en cada grupo es Normal: **ANOVA**.
- Si la distribución de la respuesta en cada grupo es Normal y no hay homogeneidad de varianzas: **ANOVA** *con la corrección de Welch*.
- Si la distribución no es Normal pero es continua: **Prueba de Kruskal-Wallis**.

¿Cómo determinar si las pruebas T-Test o ANOVA son correctas?

Normalidad de la variable respuesta en cada grupo:

- Estudio gráfico
- Prueba de Shapiro-Wilk

Homogeneidad de varianzas:

- Estudio gráfico
- Prueba de Levene

11 BIBLIOGRAFÍA

Moriña D., Utzet M, Nedel F., Martín M. and Navarro A. (2016). Introducción a la estadística para ciencias de la salud con R-Commander. Primera edición. Servei de publicacions UAB.

Moore, D., Notz W. and Flinger M. (2018). The Basic practice of statistics. 8th edition. Freeman.

En la siguiente página web se puede encontrar ayuda sobre ejemplos de código en **R** para usuarios de **R** que se pueden implementar en **Jamovi**: www.statmethods.net