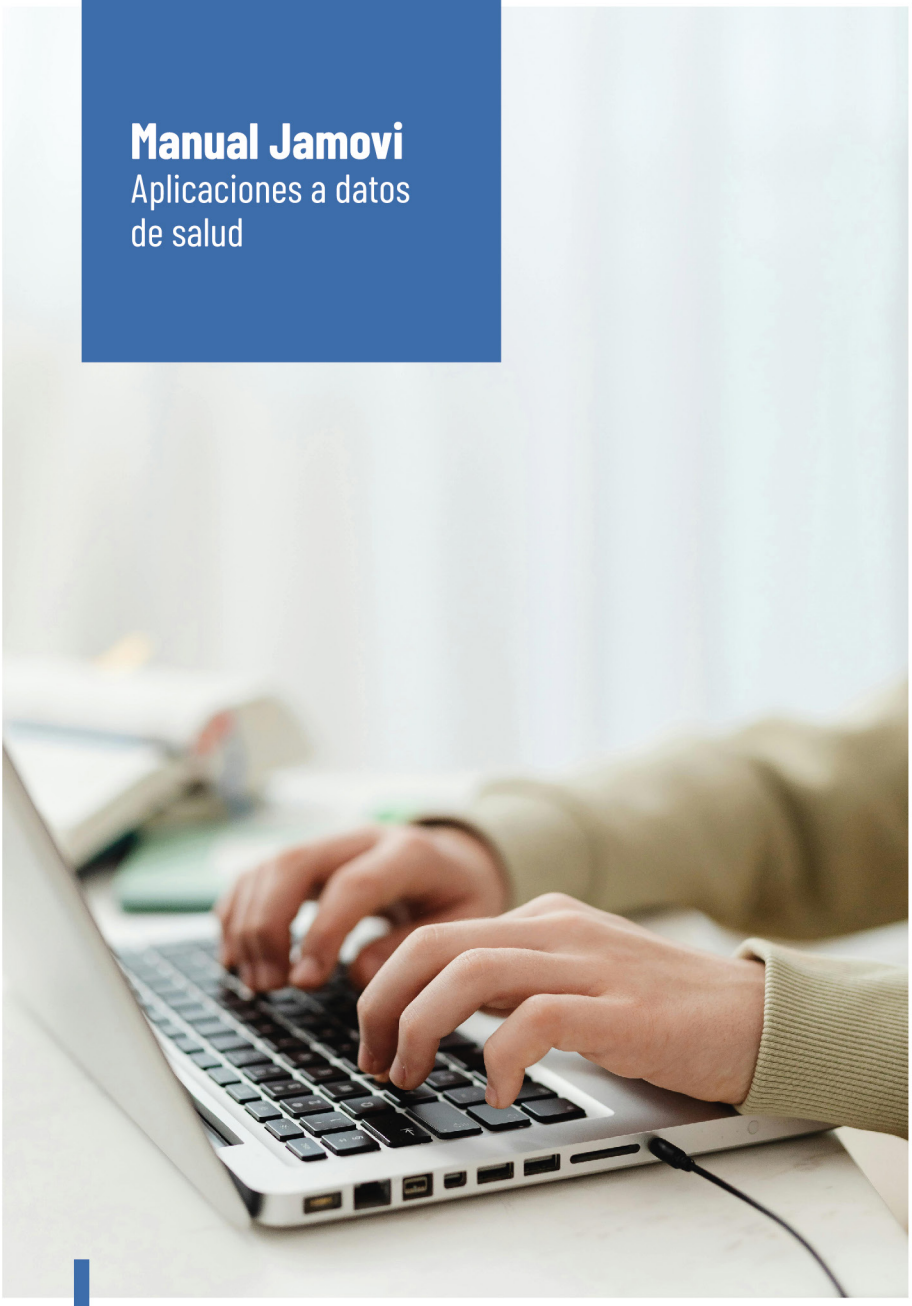
Manual Jamovi

Aplicaciones a datos de salud



Pablo Juan Verdoy
Marc Saez Zafra

Manual Jamovi. Aplicaciones a datos de salud

Manual Jamovi

Aplicaciones a datos de salud

Pablo Juan Verdoy, Marc Sáez Zafra

Documenta Universitaria

Datos CIP recomendados por la Biblioteca de la UdG

CIP 519.2 JUA

Juan Verdoy, Pablo, autor
Manual Jamovi : aplicaciones a datos de salud / Pablo Juan Verdoy, Marc Sáez Zafra. – Girona : Documenta Universitaria : Oficina Edicions UdG, 2024. – 1 recurs electrònic (162 pàgines : il·lustracions, gràfics)
ISBN 978-84-9984-691-0 (Documenta Universitaria).
ISBN 978-84-8458-706-4 (Edicions UdG)

I. Sáez Zafra, Marc, autor 1. Estadística mèdica – Programari
2. Jamovi (Programari) 3. Llibres electrònics

CIP 519.2 JUA

Foto de la cubierta: Kaboompics.com

ISBN
978-84-8458-706-4 – Universitat de Girona - Oficina Edicions UdG
978-84-9984-691-0 – Documenta Universitaria

DOI: 10.33115/b/9788499846910

Girona, diciembre 2024

Índice

Prólogo

Jamovi es un software de análisis de datos gratuito y de código abierto que ofrece una interfaz gráfica intuitiva para llevar a cabo análisis estadísticos. Es una excelente alternativa a programas como SPSS o R, especialmente para aquellos que no tienen experiencia en programación. En lo que respecta a la relación entre Jamovi y R, por ejemplo, pese a que ambos programas de análisis de datos son gratuitos y de código abierto, existen algunas diferencias clave entre ellos: Jamovi tiene una interfaz gráfica intuitiva similar a una hoja de cálculo, ideal para los principiantes y, además, permite un aprendizaje más suave y fácil, ideal para quienes no poseen experiencia en programación.

Jamovi es ideal para principiantes que buscan una herramienta fácil de usar para efectuar análisis básicos, incluso para usuarios experimentados que necesitan mayor flexibilidad y potencia para llevar a cabo análisis complejos.

Las características principales de Jamovi son su interfaz intuitiva, su amplia gama de análisis, el código abierto y sus funciones adicionales. Además, el uso de Jamovi es recomendable tanto para estudiantes, como para investigadores o profesionales.

En el capítulo 1, **Jamovi, instalación y módulos**, se desarrolla la interfaz y los módulos necesarios. En el capítulo 2, **Tratamiento de datos**, podemos ver, a partir de un conjunto de datos presentados como ejemplo, los conceptos básicos de estadística y los tipos de variables. En el capítulo 3, **Descriptiva de una variable**, como su nombre indica, se muestran todos los pasos y las posibilidades de Jamovi cuando queremos conocer, trabajar o mostrar una variable de forma descriptiva. El capítulo 4, **Descriptiva de dos o más variables**, como continuación del anterior, se complementa con las herramientas del análisis de dos o más variables a la vez. De forma más avanzada en investigación estadística, los dos capítulos siguientes, el capítulo 5, **Introducción a la Inferencia con una variable**, y el capítulo 6, **Introducción a la Inferencia de relación entre dos variables**, ayudará al uso de Jamovi para investigaciones más complejas que la misma descriptiva.

Este manual supone una buena introducción al trabajo y análisis estadístico y a su comprensión por parte de profesores, investigadores

y especialmente del alumnado, que necesita ayuda en el uso de herramientas estadísticas de forma sencilla y clara. Más allá de la misma descriptiva, Jamovi proporciona una experiencia de aprendizaje accesible y enriquecedora para usuarios de todos los niveles. Su enfoque intuitivo y sus capacidades avanzadas lo convierten en un aliado indispensable para el análisis de datos en el ámbito académico y profesional. Todo esto nos lo da Jamovi.

Capítulo 1
Jamovi, instalación y módulos

Introducción

Jamovi es un *software* de análisis de datos gratuito y de código abierto que ofrece una interfaz gráfica intuitiva para realizar análisis estadísticos. Es una excelente alternativa a programas como SPSS o R, o más complejos como Python, Stata o Epidat, especialmente para aquellos que no tienen experiencia en programación.

En lo que respecta a la relación entre **Jamovi y R** —ambos programas de análisis de datos gratuitos y de código abierto—, podemos destacar algunas diferencias clave entre ellos.

Interfaz

- **Jamovi:** interfaz gráfica intuitiva similar a una hoja de cálculo, ideal para principiantes.
- **R:** interfaz basada en código, requiere conocimientos de programación.

Flexibilidad

- **Jamovi:** menos flexible que R, con un conjunto de análisis predefinidos.
- **R:** más flexible y potente, permite realizar análisis más complejos y personalizados.

Facilidad de aprendizaje

- **Jamovi:** curva de aprendizaje más suave, ideal para quienes no tienen experiencia en programación.
- **R:** curva de aprendizaje más pronunciada, requiere tiempo y esfuerzo para dominar el lenguaje.

Mientras Jamovi es ideal para principiantes que buscan una herramienta fácil de usar para realizar análisis básicos, R es ideal para usuarios experimentados que necesitan mayor flexibilidad y potencia para realizar análisis complejos.

Se puede ejecutar el código R en Jamovi con el uso de módulos, en concreto el módulo Rj Editor. Otra opción es el paquete jmv, que permite realizar los análisis de Jamovi desde una sesión interactiva de R.

Si nos centramos en las características principales de Jamovi:

- **Interfaz intuitiva.** Se asemeja a una hoja de cálculo, lo que facilita su uso para principiantes.
- **Amplia gama de análisis.** Ofrece una amplia gama de análisis estadísticos, desde pruebas t hasta regresión lineal y análisis de varianza.
- **Código abierto.** Esto significa que es gratuito, transparente y puede ser modificado por la comunidad.
- **Funciones adicionales.** Funciones como la creación de gráficos personalizados, la gestión de datos y la exportación de resultados.

¿Quién puede utilizar Jamovi?

- **Estudiantes.** Herramienta ideal para estudiantes que están aprendiendo estadística por primera vez.
- **Investigadores.** Puede ser utilizado por investigadores para realizar análisis estadísticos complejos.
- **Profesionales.** Puede ser utilizado por profesionales en diversas áreas como la psicología, la sociología, la economía y la educación.

Después de esta pequeña introducción, vamos a empezar con Jamovi.

Para empezar con Jamovi, puedes seguir estos pasos:

1. **Descarga e instala Jamovi** desde el sitio web oficial.
   https://www.JAMOVI.org/
2. **Abre Jamovi** y familiarízate con la interfaz.
3. **Consulta la documentación** de Jamovi para aprender a realizar análisis específicos.
   https://www.JAMOVI.org/user-manual.html
4. **Visita la comunidad de Jamovi** para obtener ayuda y consejos.
   https://forum.JAMOVI.org

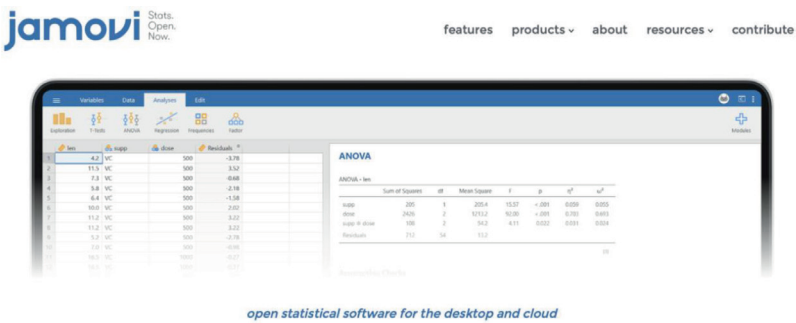


Figura 1. Web de Jamovi

A continuación, se desarrolla el apartado de la instalación propiamente dicha.

Instalación

Jamovi permite su uso en dos formatos, *online* o *desktop*, como podemos ver en su web.

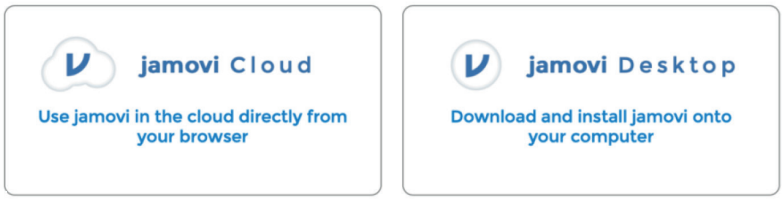


Figura 2. Web de Jamovi

En la opción Cloud, podemos ver la opción de pago, mientras que, en Desktop, tenemos las distintas versiones para descargar el programa para los distintos sistemas:

- **Windows:** descarga el instalador .exe
- **Mac:** descarga el archivo .dmg
- **Linux:** descarga el archivo .AppImage

A continuación, ejecuta el instalador, pero ten en cuenta:

- Asegúrate de tener suficiente espacio libre en disco antes de instalar Jamovi.
- Si tienes problemas para instalar Jamovi, consulta la documentación oficial:
  https://docs.JAMOVI.org/_pages/um_1_installation.html
- También puedes encontrar ayuda en la comunidad de Jamovi:
  https://www.JAMOVI.org/library.html

Los posibles problemas que pueden surgir en la instalación de Jamovi.

Problemas de descarga

- **Error de conexión a internet:** asegúrate de tener una conexión a internet estable.
- **Servidor ocupado:** intenta descargar Jamovi en otro momento.
- **Archivo dañado:** descarga el archivo de instalación de nuevo.

Problemas de instalación

- **No se cumple con los requisitos del sistema:** verifica que tu equipo cumpla con los requisitos mínimos para instalar Jamovi.
- **Archivos de instalación corruptos:** descarga el archivo de instalación de nuevo.
- **Permisos de administrador:** asegúrate de tener permisos de administrador para instalar Jamovi.
- ***Software* antivirus:** desactiva temporalmente el *software* antivirus mientras instalas Jamovi.

Problemas al iniciar Jamovi

- **Archivos de la aplicación dañados:** reinstala Jamovi.
- **Versiones incompatibles de Java:** asegúrate de tener la última versión de Java instalada.
- **Problemas con la configuración del sistema:** consulta la documentación oficial de Jamovi para obtener más información.

Otros problemas

- **Idioma:** Jamovi está disponible en varios idiomas. Asegúrate de descargar la versión en el idioma que deseas usar.
- **Actualizaciones:** Jamovi se actualiza con frecuencia. Asegúrate de tener la última versión instalada.

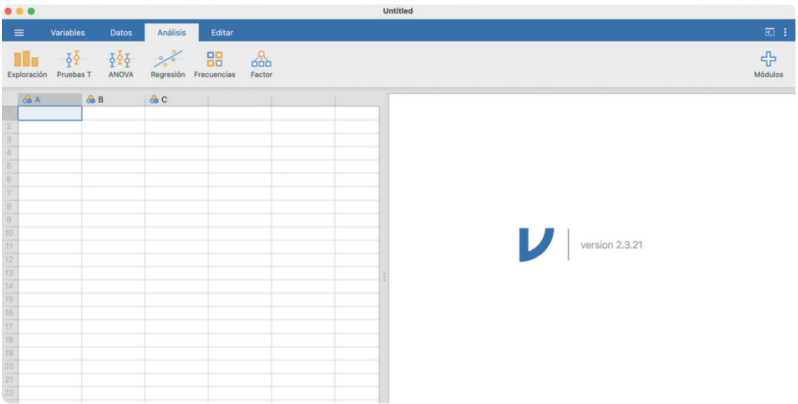


Figura 3. Jamovi al abrir al inicio el programa

Antes de ver los distintos módulos que se añadirán en Jamovi para el trabajo desarrollado en este manual para salud, vamos a ver las distintas partes del *software*.

En la parte superior tenemos el menú de opciones, donde primeramente tenemos el menú de archivos.

Figura 4. Menú de archivos

En esta parte, que desarrollaremos más adelante, podremos ver todo el tratamiento de datos.

También tenemos la gestión de datos, que desarrollaremos en la parte de transformación de estos, además de la ventana de análisis y edición.

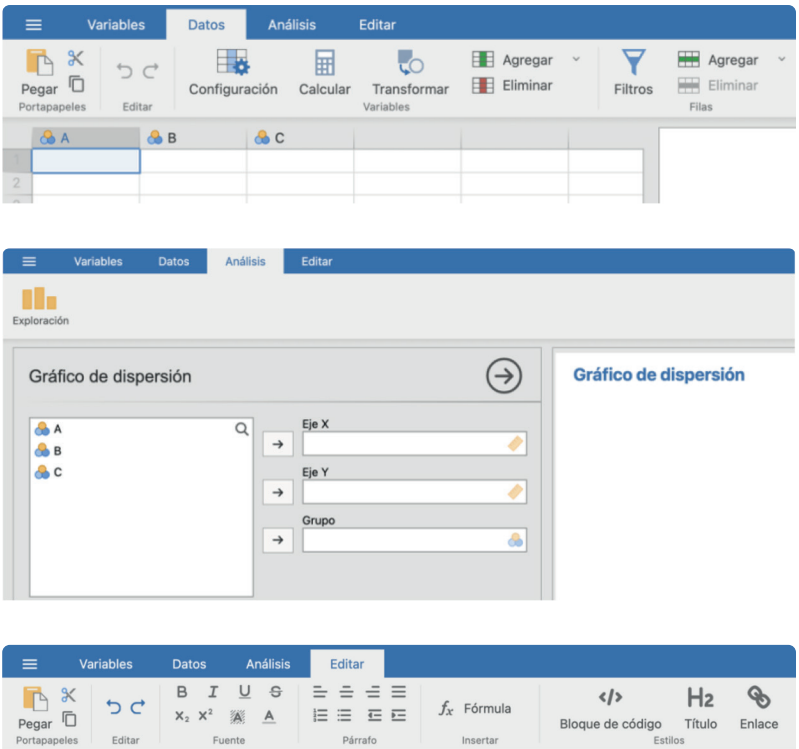


Figura 5. Menús datos, análisis y edición

Finalmente, antes de entrar en el desarrollo de los módulos, que afectarán a las distintas ventanas, comentaremos el menú de configuración de Jamovi que aparece en el desarrollo de los tres puntos verticales de la parte derecha.

Como se puede observar, en este menú se podrán cambiar las configuraciones y visualización de idiomas, datos numéricos, etc. El propio lector puede desarrollar estos puntos con pruebas concretas.

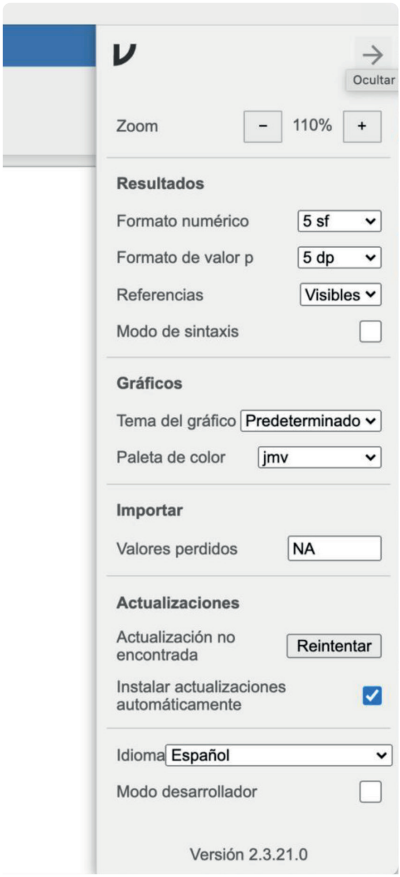


Figura 6. Menú de configuración

Módulos

Los módulos en Jamovi son **extensiones** que se pueden instalar para **ampliar las funcionalidades** del programa. Estos módulos te permiten realizar análisis estadísticos **más especializados** que no son posibles con las herramientas básicas de Jamovi.

¿Dónde encontrarlos?

Hay dos maneras de encontrar módulos:

- **Biblioteca Jamovi.** Accede a ella desde el menú «Módulos». Aquí encontrarás una lista de módulos precargados que puedes instalar con un solo clic.
- **Repositorios externos.** Puedes encontrar módulos adicionales en repositorios como https://github.com/ o https://cran.r-project.org/

¿Dónde encontrarlos?

Instalar módulos en Jamovi es un proceso bastante sencillo que se puede llevar a cabo dentro de la misma aplicación. Los módulos son extensiones que agregan funcionalidades adicionales a Jamovi, que es una aplicación estadística de código abierto diseñada para ser fácil de usar.

Los pasos que seguir:

1. **Abrir Jamovi.** Inicia la aplicación Jamovi en tu computadora. Si no la tienes instalada, puedes descargarla desde su sitio web oficial https://www.JAMOVI.org
2. **Acceso a la biblioteca de módulos.** Una vez abierto Jamovi, busca el icono de la «tienda de módulos» en la barra de menú principal. El icono suele parecerse a una caja abierta o a una bolsa de compras.
3. **Explorar módulos disponibles.** Al hacer clic en este icono, se abrirá la biblioteca de módulos de Jamovi. Aquí podrás explorar los distintos módulos disponibles para su instalación. Puedes navegar por las diferentes categorías o usar la barra de búsqueda para encontrar un módulo específico.
4. **Instalar un módulo.** Cuando encuentres el módulo que deseas instalar, haz clic en él para ver más detalles. Luego, simplemente haz clic en el botón «Instalar» o «Agregar» para iniciar la instalación del módulo.

5. **Esperar la instalación.** El módulo comenzará a descargarse e instalarse automáticamente. Este proceso puede tardar unos minutos, dependiendo del tamaño del módulo y de la velocidad de tu conexión a Internet.
6. **Cerrar y reabrir Jamovi (si es necesario).** Algunos módulos pueden requerir que reinicies Jamovi para completar la instalación. Si es así, cierra la aplicación y vuélvela a abrir.
7. **Usar el módulo instalado.** Una vez instalado el módulo, este debería aparecer en la barra de menú de Jamovi o en el lugar correspondiente dentro de la aplicación. Puedes hacer clic en el módulo para comenzar a utilizar sus funcionalidades adicionales.

Si surge algún problema durante la instalación de un módulo, como mensajes de error o problemas de compatibilidad, revisa si tienes la versión más actualizada de Jamovi. Los desarrolladores de módulos a menudo actualizan sus extensiones para ser compatibles con las versiones más recientes de Jamovi, por lo que mantener tu aplicación actualizada puede solucionar muchos problemas.

¿Cómo puedo saber qué módulo necesito?

Para saber qué módulo necesitas, debes tener en cuenta:

- El tipo de análisis que deseas realizar.
- El tipo de datos que tienes.
- Tu nivel de experiencia en estadística.

Ventajas de usar módulos:

- Amplían las capacidades de Jamovi.
- Permiten realizar análisis más especializados.
- Son gratuitos y de código abierto.
- Están desarrollados por una comunidad activa de investigadores.

Desventajas de usar módulos:

- Algunos módulos pueden ser complejos de usar.
- No todos los módulos están bien documentados.
- Es posible que algunos módulos no sean compatibles con la versión actual de Jamovi.

A continuación, vamos a desarrollar de forma más concreta alguno de los módulos que pueden ser utilizados para el análisis, sobre todo en datos de salud.

Para empezar a instalarlos, podemos ver, en la parte de análisis, una cruz en la esquina derecha que nos permite tratar los módulos.

Figura 7. Instalación de módulo

Módulo jmv

Este módulo, ***jmv* de Jamovi, es un complemento para el análisis de datos multivariante**. *jmv* ofrece una amplia gama de herramientas para:

- **Análisis de Componentes Principales (ACP).** Permite identificar las principales variables que explican la variabilidad en un conjunto de datos.
- **Análisis Factorial (AF).** Similar al ACP, pero con la capacidad de identificar factores latentes que no son directamente observables.
- **Análisis Discriminante (AD).** Permite clasificar a los individuos en diferentes grupos en función de sus características.
- **Escalamiento Multidimensional (MDS).** Representa las relaciones entre las variables en un espacio de menor dimensionalidad.
- **Análisis de Conglomerados (AC).** Permite agrupar a los individuos en función de sus características.
- **Regresión Múltiple.** Permite modelar la relación entre una variable dependiente y un conjunto de variables independientes.

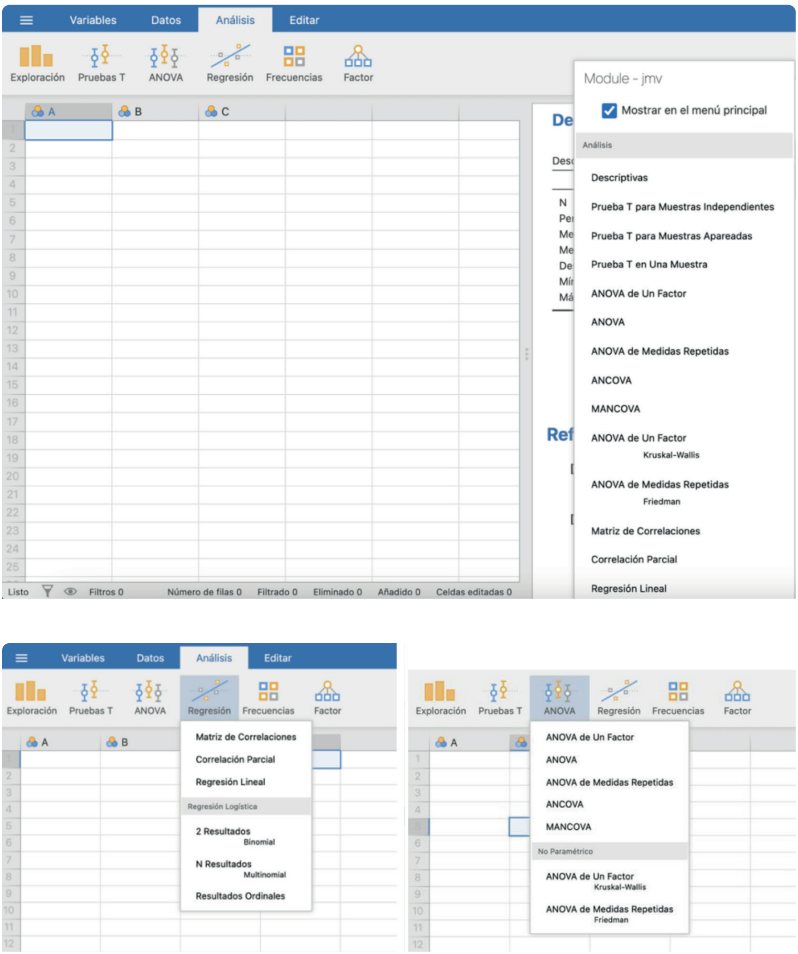


Figura 8. Módulo jmv

Funciones clave del módulo *jmv*

- **Interfaz intuitiva.** *jmv* ofrece una interfaz gráfica de usuario que facilita la selección de las opciones de análisis y la interpretación de los resultados.
- **Amplia gama de análisis.** *jmv* ofrece una amplia gama de técnicas de análisis multivariante, lo que lo convierte en una herramienta versátil para diferentes tipos de investigaciones.

- **Visualización de datos.** *jmv* ofrece diferentes herramientas para visualizar los resultados de los análisis, como gráficos de dispersión, diagramas de *scree plot* y dendrogramas.
- **Exportación de resultados.** *jmv* permite exportar los resultados de los análisis a diferentes formatos, como CSV, SPSS y R.

Módulo jmvBase

Este módulo, *jmvBase*, es un complemento del módulo anterior que nos permite efectuar, como podemos ver en la imagen, análisis más completos (ANOVA, contrastes de independencia, etc.).

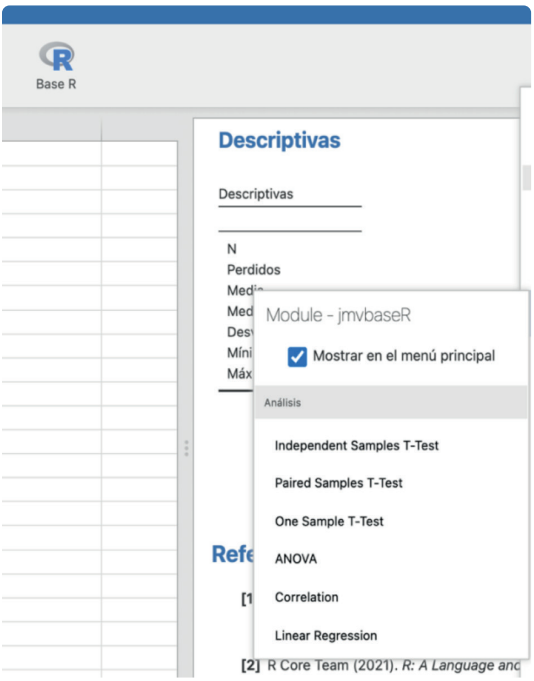


Figura 9. Módulo jmvBase

Módulo MAJOR

Otro de los módulos es *MAJOR*, que se utiliza para el desarrollo de metaanálisis.

Módulo ClinicoPath

El siguiente módulo, cuyas opciones podemos ver en la siguiente imagen, nos permite ampliar el tema descriptivo de nuestros datos (a desarrollar en el apartado de descripción de variables). Cabe notar la gran cantidad de posibilidades que ayudarán en la parte de salud.

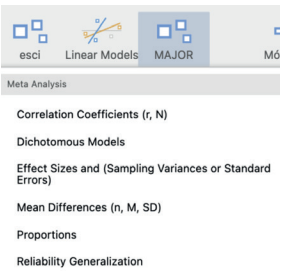


Figura 10. Módulo MAJOR

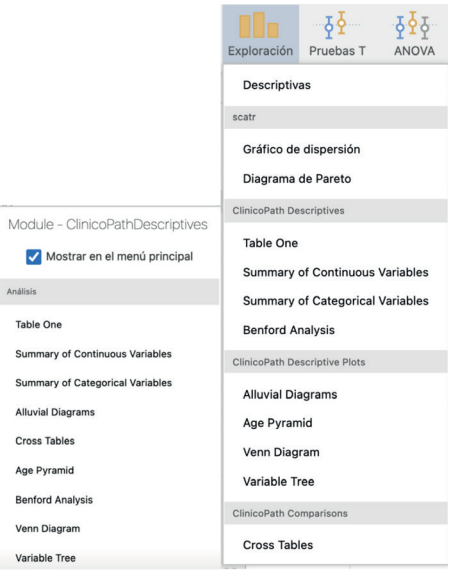


Figura 11. Módulo ClinicoPath

Módulo Statkat

Este módulo nos permite encontrar la técnica estadística adecuada para los datos y preguntas de investigación. Como se muestra en la ventana, nos permite elegir técnica para el caso de una sola variable o el caso en el que tengamos que ver relación entre variables.

Módulo ufs

El módulo *ufs* hace que las funciones del paquete R del mismo nombre estén disponibles en Jamovi. Estas incluyen funciones para calcular intervalos de confianza para tamaños de efecto, calcular el tamaño de muestra requerido para estimar un tamaño de efecto con un intervalo de confianza de un ancho dado, producir gráficos de diamante, crear una tabla de respuestas múltiples, y realizar algunas operaciones básicas como disminuir o aumentar estimaciones de tamaño de efecto.

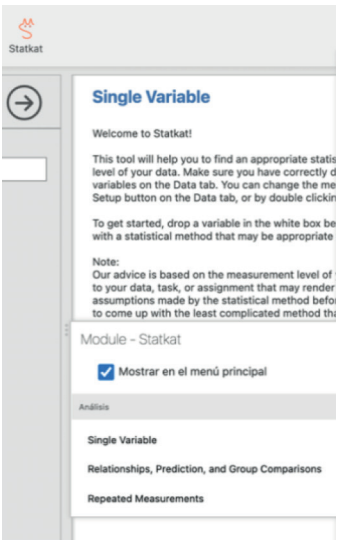


Figura 12. Módulo Statkat

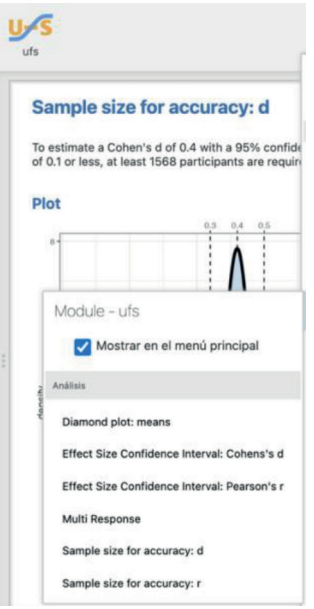


Figura 13. Módulo ufs

Módulo snowcluster, jsurvival y esci

El módulo *snowcluster*, entre otras funciones, permite a los usuarios realizar agrupamientos K-means, correspondencia simple y múltiple, y visualizar los resultados con ejes de PCA, mientras que se puede complementar con el módulo *jsurvival*, que nos permite realizar todo tipo de análisis de supervivencia.

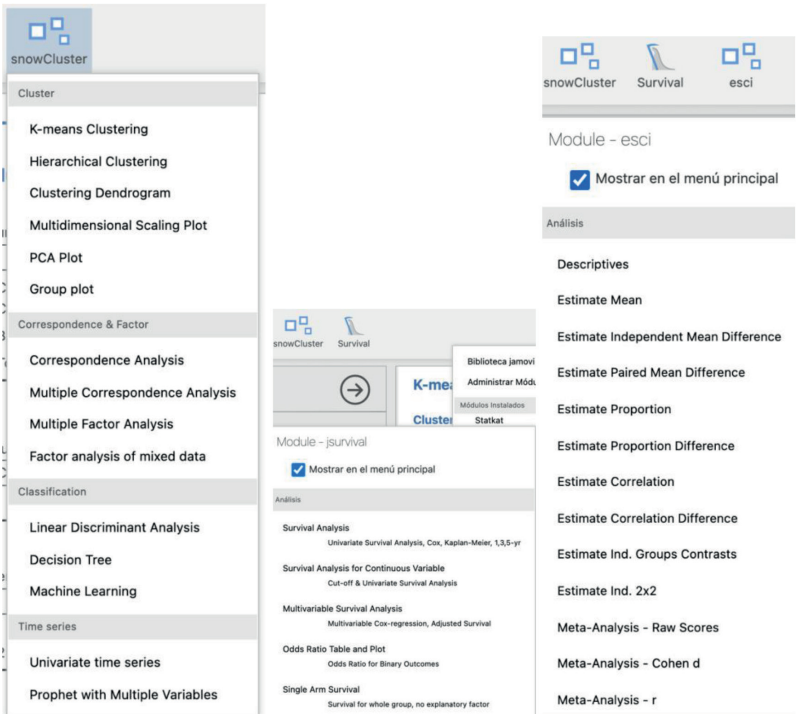


Figura 14. Módulo snowCluster (izquierda), módulo jsurvival (central) y módulo esci (derecha)

esci ofrece un paso fácil hacia las estadísticas de estimación (también conocidas como las «nuevas estadísticas»), un enfoque que enfatiza los tamaños de efecto, los intervalos de estimación y el metaanálisis. Este módulo, *esci,* puede proporcionar estimaciones e intervalos de confianza para la mayoría de los análisis que aprenderías en un curso de estadísticas de pregrado y metaanálisis (lo cual realmente debería ser parte de un buen curso de estadísticas de pregrado, ya que, aunque el módulo este pensado para cualquier usuario, fue especialmente desarrollado

pensando en los estudiantes). La mayoría de los análisis se pueden ejecutar a partir de datos brutos o de datos resumidos (lo que te permite generar estimaciones a partir de artículos de revistas que solo informaron pruebas de hipótesis). Todos los análisis generan visualizaciones agradables que enfatizan los tamaños de efecto y la incertidumbre. *esci* es para todos, pero fue desarrollado especialmente pensando en los estudiantes: proporciona instrucciones paso a paso, retroalimentación clara e intenta prevenir errores de novato (como calcular una media en una variable nominal).

Módulo jpower

Este módulo está desarrollado en el artículo «Power to the people: A beginner's Tutorial to Power Análysis using Jamovi», de James E. Bartlett y Sarah J. Charles, de 2020.

Paquete muy útil para la toma de decisiones referentes al tamaño de muestra en los datos, más concretamente en salud. Además, nos permite elegir entre tamaños de muestra o potencia del muestreo (entre otros factores) para una o más muestras.

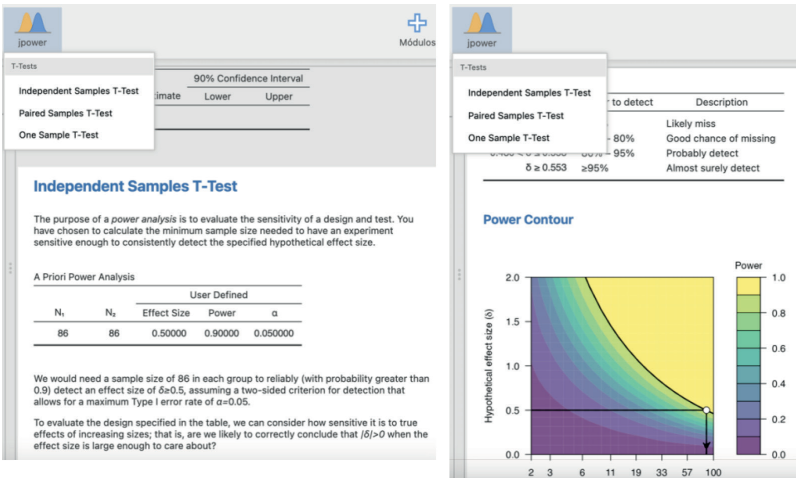


Figura 15. Módulo jpower

Módulo jjSTATPLOT y editor de Rj

El módulo *jjSTATPLOT* sirve para crear gráficos con detalles de pruebas estadísticas incluidos en los propios gráficos ricos en información. En un flujo de trabajo típico de análisis exploratorio de datos, la visualización de datos y el modelado estadístico son dos fases diferentes: la visualización informa al modelado, y el modelado, a su vez, puede sugerir un método de visualización diferente, y así sucesivamente. La idea central de *ggstatsplot* (nombre del *package* en base R) es simple: combinar estas dos fases en una en forma de gráficos con detalles estadísticos, lo que facilita y acelera la exploración de datos.

Para completar la parte gráfica y todos los análisis que realicemos en Jamovi, podemos utilizar el módulo *Rj Editor*, para ver y desarrollar los códigos de R.

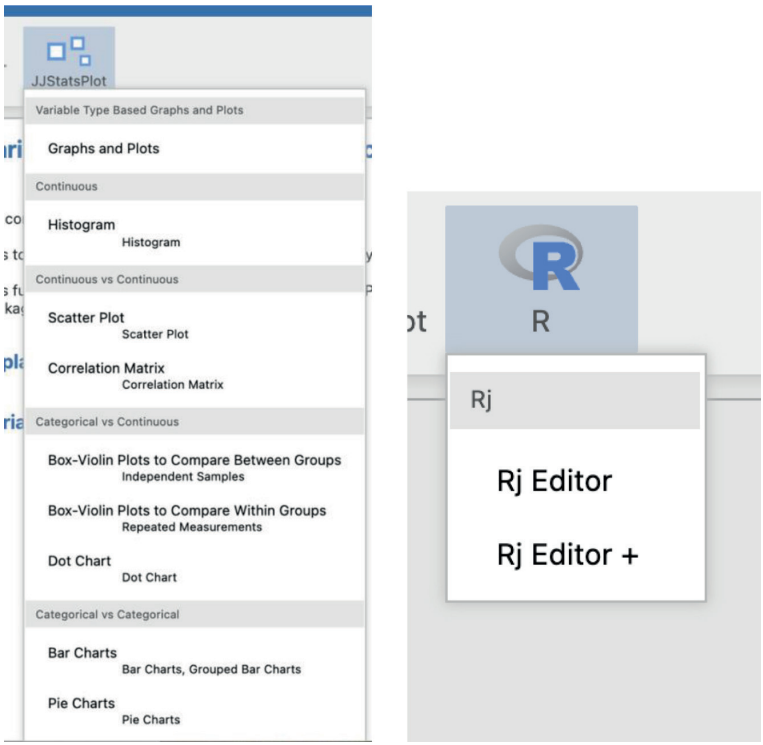


Figura 16. Módulos jjSTATPLOT y Rj Editor

Módulos jsq y seolmatrix

El módulo *jsq* nos permite realizar los distintos análisis bayesianos que podemos ver en la imagen, mientras que el módulo *seolmatrix*, por el contrario, nos sirve para completar otros módulos en la parte de correlación, evaluación o análisis jerárquico.

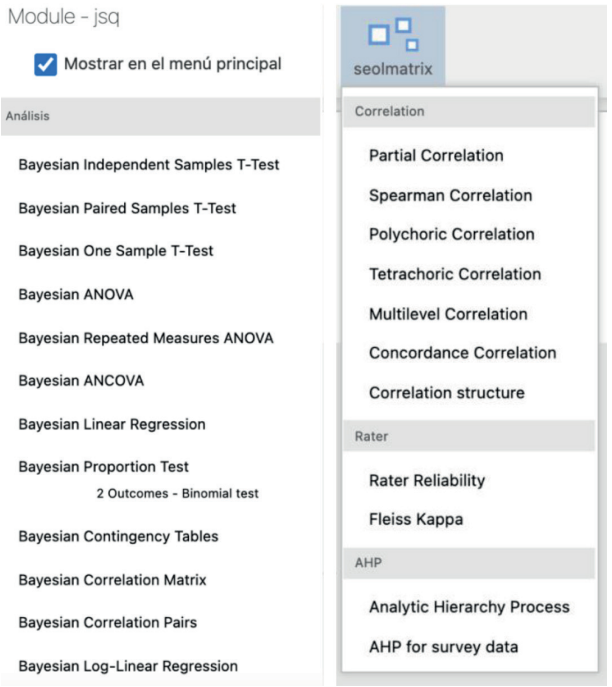


Figura 17. Módulos jsq.y seolmatrix

Módulo distrACTION

El módulo *distrACTION* nos muestra de forma gráfica la forma y desarrollo de las distintas distribuciones de probabilidad continuas o discretas.

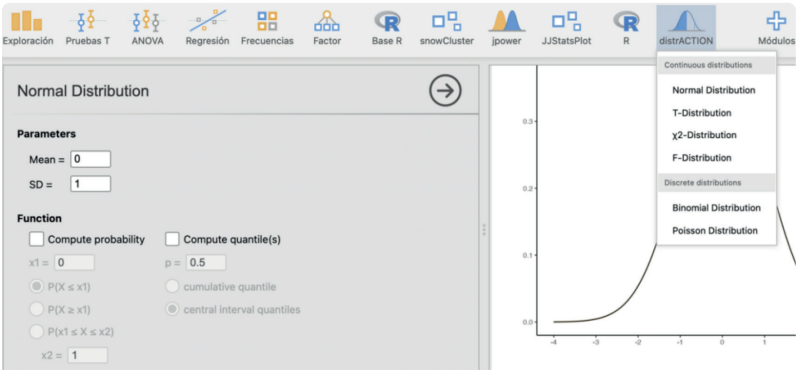


Figura 18. Módulo distrACTION

Módulo Flexplot y módulo bmtest

Estos dos módulos son muy concretos, uno para la realización de modelos lineales generalizados (en algunas de sus posibilidades) y gráficos y, por otra parte, el módulo *bmtest*, que solo realiza el contraste de Brunner-Munzel, que además está publicado en la siguiente dirección: https://www.mdpi.com/2624-8611/5/2/26

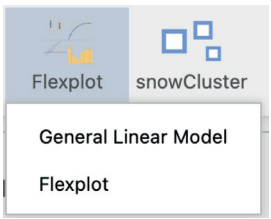


Figura 19. Módulo Flexplot

Módulos Demonstration y el módulo clt

Otros dos módulos complementarios para Jamovi son el que nos muestra demostraciones de los datos (*Demonstration*) y el paquete *clt*, indicado para realizar análisis como correlación o contraste de hipótesis.

Módulo jeva

¿Alguna vez has tratado de hacer un análisis evidencial? Es posible que te haya resultado difícil encontrar una plataforma estadística para ello. Ahora existe el módulo *jeva* de Jamovi, que puede proporcionar razones de verosimilitud para una variedad de pruebas estadísticas comunes.

Además, nos muestra muchas otras posibilidades, como podemos ver en el enlace: https://blog.JAMOVI.org/2023/02/22/jeva.html

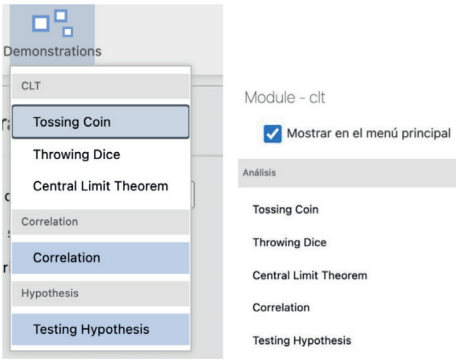


Figura 20. Módulos Demonstrations y clt

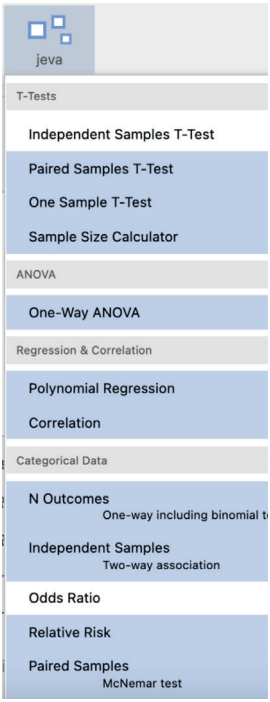


Figura 21. Módulo jeva

Módulo zestvist

Módulo claro y conciso para mostrar gráficamente, y con cálculos, el contraste de una muestra.

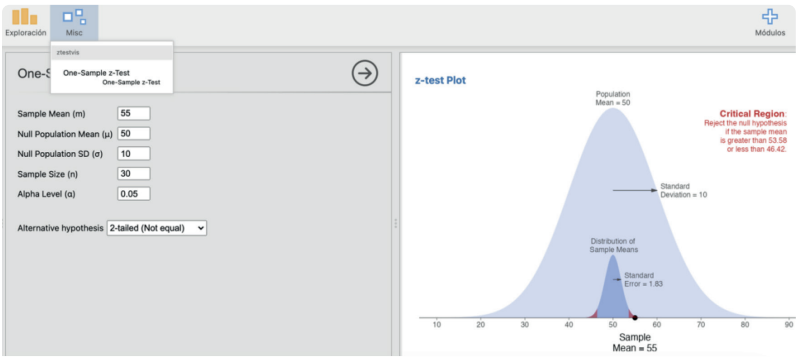


Figura 22. Módulo zestvist

Módulo meddecide

Módulo también con muchos posibles usos en salud, para el desarrollo de distintos análisis de decisión de los datos, como sensibilidad, intervalos de confianza, etc.

Figura 23. Módulo meddecide

Módulos PPDA y MEDA

Como podemos ver en la imagen, el módulo PPDA nos permite realizar descripciones automáticas de modelos univariantes y multivariantes. Por el contrario, el PPDA es mucho más concreto en este tipo de análisis, como, por ejemplo, para efectuar las curvas ROC.

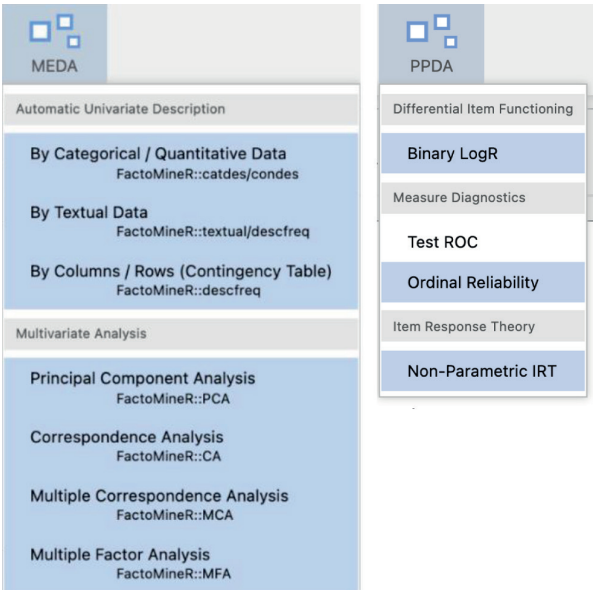


Figura 24. Módulos PPDA y MEDA

Módulo gamlj

El módulo *GAMLj* ofrece múltiples herramientas para realizar todo tipo de análisis multivariantes.

Todo su desarrollo puede verse en el enlace: https://gamlj.github.io

GAMLj ofrece herramientas para estimar, visualizar e interpretar modelos lineales generales, modelos lineales mixtos y modelos lineales generalizados con variables categóricas y/o continuas, con opciones para facilitar la estimación de interacciones, pendientes simples, efectos simples, pruebas *post hoc*, etc.

- Enfoque de ANOVA y regresión.
- Variables independientes continuas y categóricas.
- Prueba F, pruebas de razón de verosimilitud y estimaciones de parámetros.
- Intervalos de confianza, estándar, perfil y de arranque.
- Moderación facilitada.
- Análisis de pendientes simples.
- Análisis de efectos simples.
- Análisis de interacción simple.
- Análisis *post hoc.*
- Gráficos para cualquier orden de interacciones.
- Selección automática de los mejores métodos de estimación y selección de grados de libertad.
- Estimación de tipo III.
- Un amplio conjunto de índices de tamaño de efecto, según el modelo que se esté estimando.

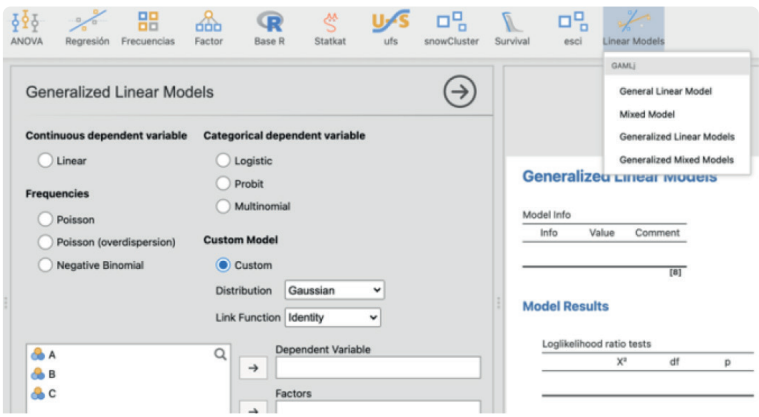


Figura 25. Módulo GAMLj

Los posibles modelos:

- OLS Regression (GLM).
- OLS ANOVA (GLM).

- OLS ANCOVA (GLM).
- Random coefficients regression (Mixed).
- Random coefficients ANOVA-ANCOVA (Mixed).
- Logistic regression (GZLM).
- Logistic ANOVA-like model (GZLM).
- Probit regression (GZLM).
- Probit ANOVA-like model (GZLM).
- Multinomial regression (GZLM).
- Multinomial ANOVA-like model (GZLM).
- Poisson regression (GZLM).
- Poisson ANOVA-like model (GZLM).
- Overdispersed Poisson regression (GZLM).
- Overdispersed Poisson ANOVA-like model (GZLM).
- Negative binomial regression (GZLM).
- Negative binomial ANOVA-like model (GZLM).
- Ordinal regression (GZLM).
- Ordinal ANOVA-like model (GZLM).
- Mixed Logistic regression (GMixed).
- Mixed Logistic ANOVA-like model (GMixed).
- Mixed Probit regression (GMixed).
- Mixed Probit ANOVA-like model (GMixed).
- Mixed Multinomial regression (GMixed).
- Mixed Multinomial ANOVA-like model (GMixed).
- Mixed Poisson regression (GMixed).
- Mixed Poisson ANOVA-like model (GMixed).
- Mixed Overdispersed Poisson regression (GMixed).
- Mixed Overdispersed Poisson ANOVA-like model (GMixed).
- Mixed Negative binomial regression (GMixed).
- Mixed Negative binomail ANOVA-like model (GMixed).
- Mixed Ordinal regression (GMixed).
- Mixed Ordinal ANOVA-like model (GMixed).

Referencias

Introducción a Jamovi, por JD Leongómez:
https://jdleongomez.info/es/post/JAMOVI/

Guía didáctica: Introducción al análisis de datos en estudios biomédicos (Jamovi):
https://sites.google.com/iacs.es/analisisdatosJAMOVI/inicio

Información sobre cómo instalar y usar los módulos:
https://www.JAMOVI.org/user-manual.html

Página web de la biblioteca Jamovi:
https://www.JAMOVI.org/library.html

Canal de YouTube de RecuEdu:
https://www.youtube.com/watch?v=JMsLTnk1_gU

Blog de Matt C. Howard:
https://mattchoward.com/instalacion-de-modulos-en-JAMOVI/

Capítulo 2
Tratamiento de los datos

Conceptos básicos de estadística

Entre las distintas definiciones de la estadística escogemos dos: i) disciplina que estudia la variabilidad; ii) ciencia que estudia cómo se tiene que utilizar la información y cómo dar una guía de acción en situaciones prácticas que presentan incertidumbre. El proceso de investigación estadístico requiere la recolección, organización, análisis, interpretación y presentación de los datos.

La palabra *estadística*, aunque tiene la raíz latina *status*, que significa 'cifras relacionadas con el Estado', es en realidad la forma femenina del término alemán *statistik*, derivado a su vez del italiano *statista*, que significa 'hombre de estado'.

La estadística puede dividirse en:

- **Estadística descriptiva**: la estadística descriptiva es una rama de la estadística que se encarga de organizar, resumir y presentar de manera sistemática y comprensible un conjunto de datos para extraer conclusiones útiles sobre sus características. Su objetivo principal es describir y analizar los datos mediante técnicas y métodos que permiten entender su distribución, tendencia central, dispersión, forma y otras propiedades relevantes.
- **Inferencia estadística** o, también, **estadística inductiva**: es la parte de la estadística que trata de las condiciones y procedimientos bajo los cuales podemos extraer conclusiones o inferencias de la población a partir de la información muestral.

En este capítulo 2 y en el siguiente, utilizaremos la estadística descriptiva para el análisis de la base de datos Framingham.

Explicación datos ejemplo: Framingham

Para el tratamiento de datos, como es normal, hacen falta datos. Para la explicación de todas las posibles técnicas que se utilizan en Jamovi relacionadas con la salud, vamos a utilizar una base de datos conocida, la de Framingham.

La enfermedad cardiovascular (ECV) es la principal causa de muerte y enfermedad grave en los Estados Unidos. En 1948, se inició el Estudio Framingham sobre el Corazón (en adelante, estudio Framingham) bajo la dirección del Instituto Nacional del Corazón (ahora conocido como Instituto Nacional del Corazón, Pulmón y Sangre o NHLBI, por sus siglas en inglés). En ese momento, se sabía poco sobre las causas generales de las enfermedades cardíacas y los accidentes cerebrovasculares, aunque las tasas de mortalidad por enfermedad cardiovascular estaban aumentando constantemente desde principios del siglo xx, convirtiéndose en una epidemia.

Más tarde, el estudio Framingham se convirtió en un proyecto conjunto del NHLBI y la Universidad de Boston. El objetivo del estudio Framingham era la identificación de los factores o características comunes que contribuyen al desarrollo de enfermedades cardiovasculares en un gran grupo de participantes. Estos participantes aún no habían desarrollado síntomas evidentes de enfermedad cardiovascular ni habían sufrido un ataque cardíaco o un accidente cerebrovascular. Los investigadores reclutaron a 5 209 hombres y mujeres de entre 30 y 62 años, de la ciudad de Framingham, Massachusetts, y comenzaron la primera ronda de exámenes físicos y entrevistas sobre el estilo de vida, que posteriormente analizarían en busca de patrones comunes relacionados con el desarrollo de ECV. Desde 1948, los sujetos fueron controlados cada dos años, recopilando su historial clínico detallado, realizándoles un examen físico, así como pruebas de laboratorio.

En 1971, el estudio incorporó una segunda generación, que consistió en 5 124 hijos adultos de los participantes originales y sus cónyuges. En abril de 2002, el estudio entró en una nueva fase al registrar una tercera generación de participantes, los nietos de la cohorte original, en concreto, 4 095 participantes. Este paso es de vital importancia para ampliar nuestra comprensión de las enfermedades cardíacas y los accidentes cerebrovasculares, y cómo estas condiciones afectan a las familias.

A lo largo de los años, el seguimiento preciso de la población del estudio Framingham ha permitido identificar los principales factores de riesgo de enfermedad cardiovascular: presión arterial alta, colesterol alto, tabaquismo, obesidad, diabetes y la inactividad física, así como una gran cantidad de información valiosa sobre los efectos de factores

relacionados, como los niveles de triglicéridos y colesterol HDL, edad, sexo y problemas psicosociales.

Con la ayuda de otra generación de participantes, el estudio puede acercarse a las causas profundas de las enfermedades cardiovasculares y ayudar en el desarrollo de nuevas y mejores formas de prevenir, diagnosticar y tratar las enfermedades cardiovasculares.

El estudio Framingham ha tenido un impacto significativo en la práctica médica. Previo a la década de 1950, se consideraba normal que las personas con alteraciones en las constantes vitales o resultados de laboratorio no manifestaran signos clínicos evidentes de enfermedad, calificando estas anomalías como benignas si no causaban molestias clínicas.

Un hallazgo relevante del estudio fue la demostración estadística de la relación entre la diabetes y las enfermedades cardiocirculatorias, reconociendo la hiperglicemia como un factor de riesgo. El estudio Framingham se inició con el propósito de evaluar y demostrar los factores personales y ambientales que influían en la aparición temprana de la arteriosclerosis. Las conclusiones del estudio, al confirmar estadísticamente la importancia de los factores de riesgo, han sido ampliamente satisfactorias y su implementación es esencial en la práctica clínica.

Este estudio ejemplifica el trabajo en equipo, y los desafíos enfrentados durante su continuación y expansión resaltan el entusiasmo y la dedicación de los profesionales involucrados. Se logró el objetivo propuesto, validando estadísticamente las percepciones dispersas sobre el riesgo asociado con las alteraciones en las constantes físicas y metabólicas. El estudio Framingham establece de manera contundente el concepto de *factores de riesgo*.

El archivo Framingham.sav es un subconjunto anonimizado de los datos correspondientes a 40 años de seguimiento. Contiene medidas de variables sobre 4 434 pacientes libres de enfermedad coronaria en el inicio del seguimiento (momento basal).

Estos datos, siempre para una finalidad docente, pueden solicitarse en https://biolincc.nhlbi.nih.gov/teaching/ (requiere registro previo).

A continuación, vamos a ver el conjunto de las variables que tenemos en este fichero, para familiarizarnos con él.

Variables identificadoras

- **Randid.** Número de identificación único para cada participante
- **Participación**. 1 una ola, 2 dos olas, 3 tres olas

- **Dónde_participación.** 1 primera (ola), 2 segunda, 3 tercera, 4 primera y segunda (olas), 5 segunda y tercera, 6 primera y tercera, 7 todas (las tres olas)
- **Seguimiento.** Tiempo de seguimiento (en años).

Variables respuesta (variables dependientes)

- **death**. Muerte por cualquier causa.
- **angina**. Angina de pecho.
- **hospmi**. Infarto de miocardio hospitalizado.
- **mi_fchd**. Infarto de miocardio hospitalizado o enfermedad coronaria fatal.
- **anychd**. Angina de pecho, infarto de miocardio (hospitalizado y silencioso o no reconocido), insuficiencia coronaria (angina inestable) o enfermedad coronaria fatal.
- **stroke**. Infarto aterotrombótico, embolia cerebral, hemorragia intracerebral, hemorragia subaracnoidea o enfermedad cerebrovascular fatal.
- **cvd**. Infarto de miocardio (hospitalizado y silencioso o no reconocido), enfermedad coronaria fatal, infarto aterotrombótico, embolia cerebral, hemorragia intracerebral, hemorragia subaracnoidea o enfermedad cerebrovascular fatal.
- **hypertens**. Hipertensión. Definida como el primer examen tratado por presión arterial alta o el segundo examen en el que la presión sistólica es mayor o igual a 140 mmHg o la presión diastólica es mayor o igual a 90 mmHg.
- **timeap**. Número de días desde el examen inicial hasta el primer episodio de angina durante el seguimiento o número de días desde el examen inicial hasta la fecha de censura. La fecha de censura puede ser el final del seguimiento, la muerte o la última fecha de contacto conocida si el sujeto se pierde durante el seguimiento.
- **timemi**. Definido como arriba para el primer evento de HOSPMI durante el seguimiento.
- **timemifc**. Definido como arriba para el primer evento de MI_FCHD durante el seguimiento.
- **timechd**. Definido como arriba para el primer evento de ANYCHD durante el seguimiento.
- **timestrk**. Definido como arriba para el primer evento de STROKE durante el seguimiento.

- **timecvd**. Definido como arriba para el primer evento de CVD durante el seguimiento.
- **timedth**. Número de días desde el examen inicial hasta la muerte si ocurre durante el seguimiento o número de días desde el examen inicial hasta la fecha de censura. La fecha de censura puede ser el final del seguimiento o la última fecha de contacto conocida si el sujeto se pierde durante el seguimiento.
- **timehyp**. Definido como arriba para el primer evento de HYPERTEN durante el seguimiento.

Características y factores de riesgo (variables independientes)

- **sexo**. Sexo del participante.
- **edad basal** – **edad_basal_cat**. Edad en el momento basal (años y categorías, cuartiles).
- **edad – edad_cat**. Edad en el examen (años y categorías, cuartiles).
- **sysbp**. Presión arterial sistólica (media de las dos últimas de tres mediciones) (mmHg).
- **diabp**. Presión arterial diastólica (media de las dos últimas de tres mediciones) (mmHg).
- **cursmoke**. Consumo actual de cigarrillos en el examen.
- **cigpday**. Número de cigarrillos fumados por día.
- **imc**. Índice de masa corporal, peso en kilogramos/altura en metros cuadrados.
- **diabetes**. Diabético según criterios del primer examen tratado o primer examen con glucosa casual de 200 mg/dL o más.
- **bpmeds**. Tratamiento para la hipertensión.
- **glucosa**. Glucosa.
- **prevap**: Angina de pecho prevalente en el examen.
- **prevchd**. Enfermedad coronaria prevalente definida como angina de pecho preexistente, infarto de miocardio (hospitalizado, silencioso o no reconocido) o insuficiencia coronaria (angina inestable).
- **prevmi**. Infarto de miocardio prevalente.
- **prevstrk**. Accidente cerebrovascular prevalente.
- **prevhyp**. Hipertensión prevalente. El sujeto fue definido como hipertenso si estaba siendo tratado o si en el segundo examen la

presión sistólica media era > = 140 mmHg o la presión diastólica media era > = 90 mmHg.

- **heartrte**. Frecuencia cardíaca.
- **totchol**. Colesterol total en suero (mg / dL).
- **hdlc**. Colesterol de lipoproteínas de alta densidad (HDL) (mg / dL).
- **ldlc**. Colesterol de lipoproteínas de baja densidad (LDL) (mg / dL).
- **overweight**. 1 sobrepeso (25 kg/m^2 ≤ IMC ≤ 30 kg/m^2), 0 otro caso.
- **obese**. 1 obesidad (IMC ≥ 30 kg/m2), 0 otro caso.
- **Garrow**. 0 normopeso, 1 sobrepeso, 2 obeso.

A partir del conocimiento de las distintas variables, vamos a trabajarlas, después de conocer un poco los tipos de ficheros que podemos utilizar en Jamovi.

Tipos de ficheros

Jamovi es compatible con una gran variedad de tipos de archivos, lo que lo hace muy versátil para trabajar con diferentes tipos de datos estadísticos.

En el menú «Archivos», podemos ver la forma de abrir los mismos (figura 1).

Las opciones que tenemos:

- Nuevo. Abrir una nueva instancia de Jamovi. Además, lo hace sin cerrar la actual.
- Abrir. Muestra una ventana para seleccionar un archivo de datos para cargarlos dentro de Jamovi.
  - Funciona con una variedad de archivos de aplicaciones comúnmente utilizadas para el análisis de datos cuantitativos (SPSS, Stata, SAS, JASP) y hojas de cálculo (LibreOffice Calc o Excel), así como archivos de R o el formato genérico para intercambio de datos CSV.
  - Además, nos permite abrir datos que nosotros tengamos o datos que nos proporciona el propio Jamovi (biblioteca de datos).

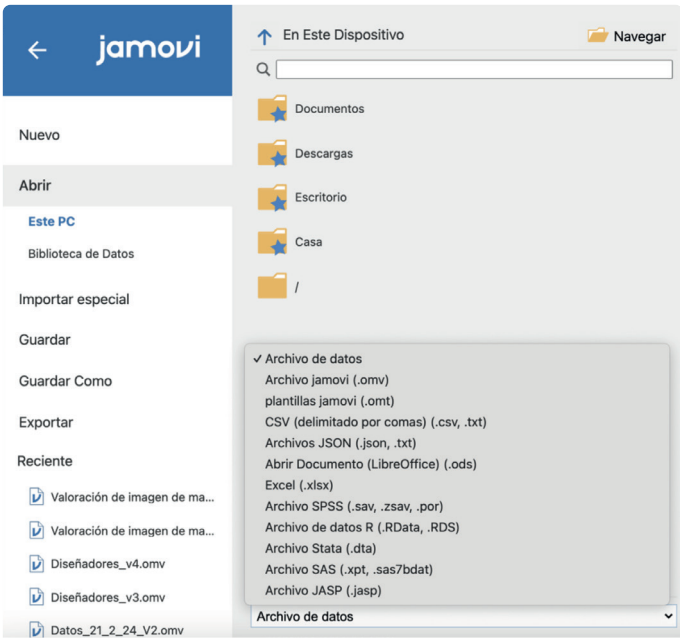


Figura 1. Imagen de Jamovi, para abrir archivos

Figura 2. Abrir archivos de la biblioteca de datos

En la figura 2, podemos ver los distintos ejemplos que, por defecto, nos proporciona Jamovi para poder aplicar las distintas técnicas estadísticas de análisis y descriptivas.

- Importar especial. Muestra una ventana para seleccionar uno o más archivos de datos para agregarlos dentro de Jamovi. Funciona con archivos del tipo OMV.
- Guardar y Guardar como. Guarda en el formato propio de Jamovi (OMV) toda la información de la sesión actual de análisis, no solo la matriz de datos, con las identificaciones de cada variable original y cada variable generada o transformada, así como todo el código R que genera los resultados solicitados, así como todo el texto agregado a los resultados.
- Exportar. Genera diferentes tipos de archivos con contenido específico, según el formato seleccionado para la exportación.
- Resultados. Se pueden generar páginas con todas las tablas y gráficos de los resultados y con los textos adjuntos, en el caso de exportaciones hacia PDF, HTML o LaTeX (paquete zip).
- Datos. En el caso de exportaciones hacia formatos de aplicaciones de análisis, lo que se genera son archivos únicamente con los datos, sin los resultados.

Ahora que conocemos los tipos de archivos que podemos utilizar en Jamovi, pasemos a analizar los diferentes tipos de variables.

Tipos de variables

Antes de conocer los tipos de variables, vamos a conocer los pasos previos para preparar los datos. La hoja de trabajo de Jamovi está organizada en columnas, cada una de las cuales se refiere a una variable. El tipo de análisis a realizar está ligado al tipo de variables. Podemos ver, a continuación, la ventana que se nos abrirá en la pestaña de datos (figura 3).

En lo que respecta al tipo de variable a definir o utilizar, podemos ver que tenemos cuatro tipos (figura 4): nominal, ordinal, continua e ID.

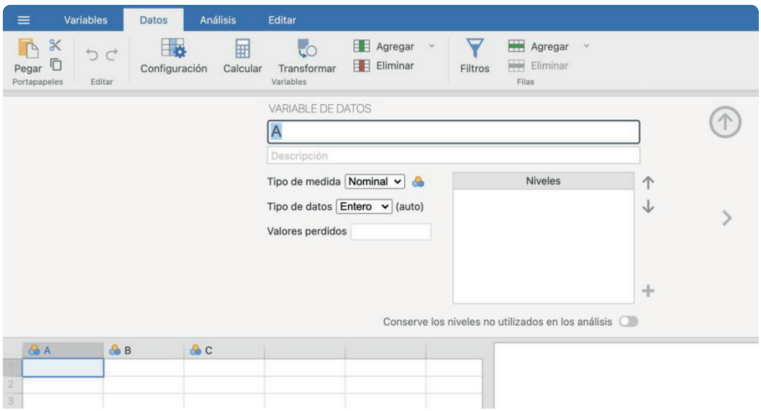


Figura 3. Configuración de datos

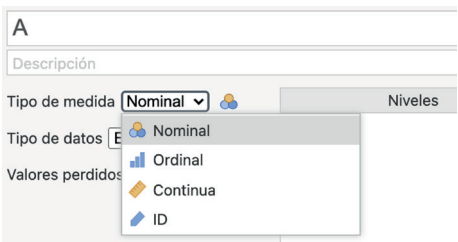


Figura 4. Tipos de datos

Podemos definir, de forma resumida, cada una de ellas:

- Variables identificadoras (ID). Son variables que identifican una observación (sujeto, medida, fecha, etc.). Por ejemplo, las variables identificadoras de la base de datos Framingham.
- Variables cualitativas, categóricas o atributos. Son variables cuyas modalidades no son medibles, expresando únicamente cualidades o categorías. No pueden expresarse mediante valores numéricos, aunque podamos utilizar números para identificar y simplificar cada categoría. Por ejemplo, la variable dependiente de Framingham *death*, que toma los valores 'vivo' y 'muerto', que codificamos como 0 si el sujeto está 'vivo' al final del seguimiento de la cohorte y 1 si 'murió' antes del final del seguimiento. Otro ejemplo, en este caso variable independiente (características y factores de riesgo), Garrow, que toma los valores 'normopeso', 'sobrepeso' y 'obeso', y codificamos 0 'normopeso', 1 'sobrepeso' y 2 'obeso'.

  Las variables cualitativas se clasifican en:

  - Nominales. Solo relaciones de identidad. Permiten obtener información de la igualdad o desigualdad de dichos individuos para una determinada característica.

    Las variables nominales pueden ser tanto dicotómicas, que toma dos valores, por ejemplo, la variable *death* en la base Framingham, como politómicas, más de dos categorías. Una variable politómica nominal serían los tipos de *stroke*: infarto aterotrombótico, embolia cerebral, hemorragia intracerebral, hemorragia subaracnoidea o enfermedad cerebrovascular fatal.

  - Ordinales. Se puede valorar su orden. Nos permiten obtener información de la igualdad o desigualdad de dichos individuos para una determinada característica. Las variables ordinales solo son politómicas (más de dos valores), por ejemplo, la variable *Garrow* en la base Framingham.

- Variables cuantitativas. Sus modalidades son medibles y pueden expresarse con valores numéricos. Estas variables se clasifican a su vez en:

  - Discretas. Son aquellas que toman valores enteros no negativos y con un número finito o infinito numerable de valores distintos. Por ejemplo, la variable *timemi* en la base de datos Framingham, el número de días desde el examen inicial hasta el primer episodio de infarto de miocardio o, si el sujeto no ha padecido un infarto, hasta el final del seguimiento. Esta variable toma valores enteros no negativos (0, 1, 2, etc.) y con un rango no muy grande, siendo un número finito de valores (como máximo el número de días de seguimiento).

  - Continuas. Son aquellas que pueden tomar cualquier valor de los infinitos que hay en un intervalo de valores. Por ejemplo, las variables independientes de la base Framingham *sysbp*, presión arterial sistólica, *diabp*, presión arterial diastólica y *glucosa*, nivel de glucosa.

Veamos un ejemplo (figura 5) donde la misma variable puede ser nominal, donde conservará las clases o continua, donde estas desaparecerán.

Figura 5. Cambio de variable

Tratamiento de datos

Creación

La primera de las posibilidades que ofrece Jamovi en cuanto a tratamiento de datos es la creación de estos. Veamos, pues, cuáles son los pasos a seguir en el supuesto de tener que introducir datos.

Datos (ejemplo número de miembros de una familia, que sería una variable cuantitativa discreta):

3,4,5,4,3,4,2,1,4,1,2,1,1,2,3,1,3,2,3,2.

En las siguientes ventanas (figura 6), podemos ver los pasos simples para la introducción de los datos.

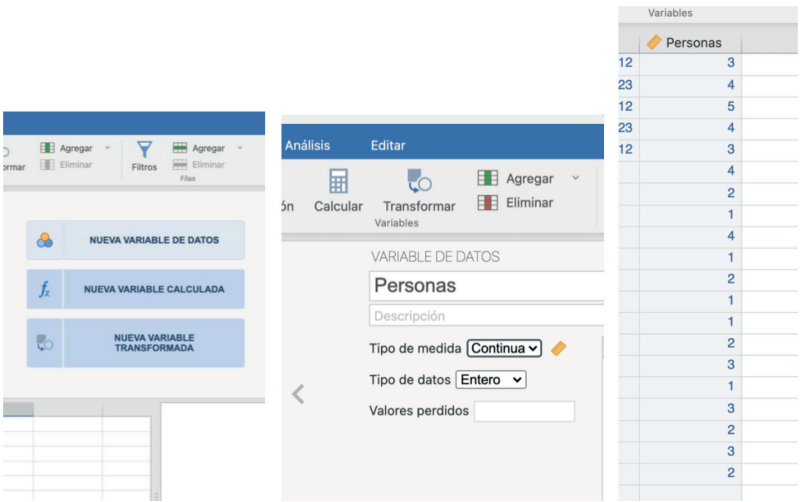


Figura 6. Introducción de datos

Se puede también crear el fichero de los datos, a partir de otro que dispongamos en otro *software* con formato de hoja de cálculo, copiándolos directamente. Esto nos permite trabajar o crear datos que tengamos en filas y los queramos en columnas. Por ello, se aconseja la creación del fichero de datos mediante el proceso de copiado en Jamovi.

Transformación

Para transformar una variable, hay que tener claro que vamos a crear una nueva a partir de una original.

Por ejemplo, si cogemos los datos del apartado anterior, el primer paso será pulsar en el icono de «Transformar» (figura 7, izquierda).

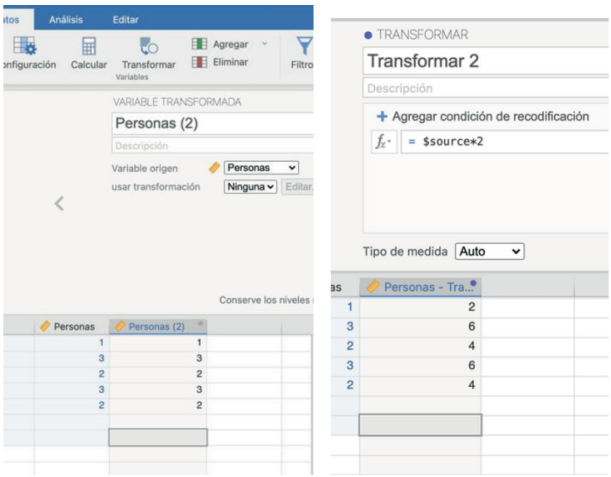


Figura 7. Transformación de variables

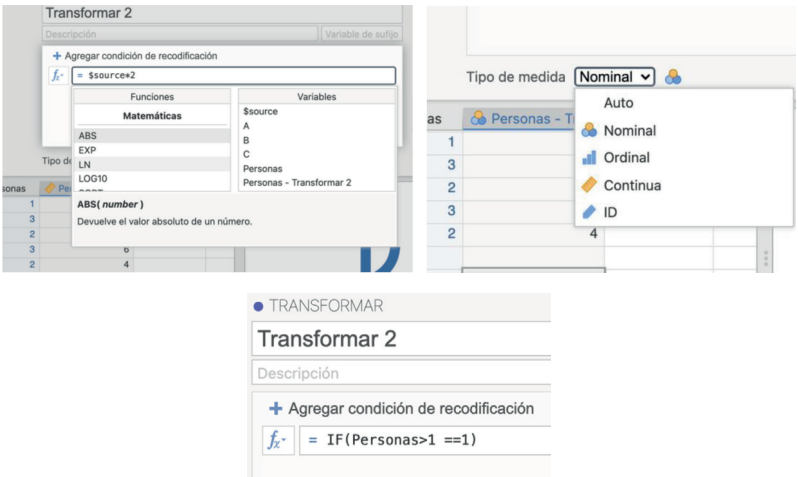


Figura 8. Transformación de variables con la función IF

A continuación, se crea y se aplica la transformación necesaria (figura 7, derecha). Por ejemplo, lo que se ha hecho es multiplicar la variable por 2.

También podemos solicitar que cumpla unas condiciones con la función IF (figura 8). En este caso, lo que se ha conseguido es la creación de una variable que nos da el valor 1 cuando tenemos más de una persona por familia.

Vemos que, además, se pueden utilizar muchas funciones tanto matemáticas como estadísticas o lógicas para la creación de la transformación.

Recodificación de variables

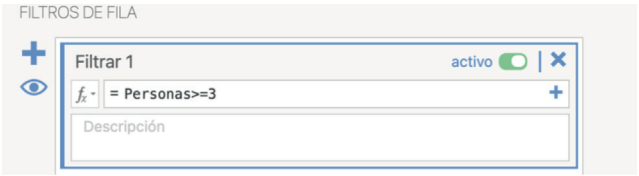
En el ámbito de la codificación de datos, la recodificación de variables es una técnica ampliamente utilizada. Si bien su función principal es agrupar un rango de valores existentes en nuevas categorías, también permite asignar una nueva codificación a los valores originales, facilitando así el análisis posterior.

En cualquier caso, la *recodificación* siempre implica un cambio en los valores de las variables, un cambio que puede o no implicar una modificación en la métrica de la variable. En este sentido, pueden darse cuatro situaciones diferentes, en función del tipo de las variables original y final: pasar de cualitativa a cuantitativa o a una misma cualitativa o al revés, de cuantitativa a cualitativa o cuantitativa. En todos los casos, se debe utilizar el procedimiento de aplicación de una regla de transformación (variable origen + regla de transformación = variable final). Se trata de crear una nueva variable a través de otras previas.

En los dos casos, tanto las transformaciones como la recodificación de variables, es posible que sean utilizadas para casos más complejos. Se irán desarrollando durante el documento cuando estas sean necesarias.

Filtros

En algunos casos, es necesario que seleccionemos solo algunos casos de los datos que tenemos. Por ejemplo, en los datos ejemplo, figura 9, seleccionar solo los casos que sean, de la variable *Personas*, mayor o igual a 3. Observamos que se ha creado una columna donde pone filtro y con los casos seleccionados.



	Filtrar 1	A	B	C	Personas	Perso
1	✓	1	a	12	3	
2	✓	2	b	23	4	
3	✓	3	c	12	5	
4	✓	4	d	23	4	
5	✓	5	e	12	3	
6	✓	5	a	23	4	
7	✗	4	e	12	2	
8	✗	3	c	11	1	
9	✓	4	d	20	4	
10	✗	4	a	15	1	
11	✗				2	

Figura 9. Filtro de variables

Esto nos permitirá seleccionar datos de casos concretos, como, por ejemplo, en bases de datos donde queramos separar por alguna característica: sexo, edad u otras que cumplan una condición.

Cálculos

Finalmente, también podemos hacer, de forma simple, cálculos con las variables. En la imagen (figura 10), podemos ver cómo crear una nueva variable que sea el LN de la columna A.

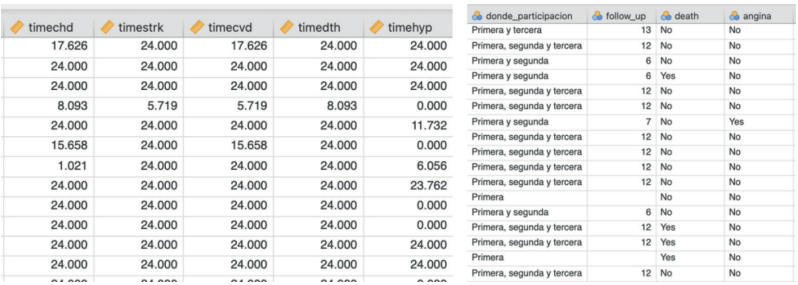
	A	B (2)
1	1	0.000
2	2	0.693
3	3	1.099
4	4	1.386
5	5	1.609
6	5	1.609
7	4	1.386
8	3	1.099
9	4	1.386
10	4	1.386

Figura 10. Cálculo de variables

Ejemplo de tratamiento de datos y variables

Con el fichero Framingham, vamos a ver algún ejemplo de los apartados anteriores, aplicado a estos datos.

Primeramente, podemos ver en el fichero distintos tipos de variables. Por ejemplo, como variables continuas o variables cualitativas (figura 11).



timechd	timestrk	timecvd	timedth	timehyp
17.626	24.000	17.626	24.000	24.000
24.000	24.000	24.000	24.000	24.000
24.000	24.000	24.000	24.000	24.000
8.093	5.719	5.719	8.093	0.000
24.000	24.000	24.000	24.000	11.732
15.658	24.000	15.658	24.000	0.000
1.021	24.000	24.000	24.000	6.056
24.000	24.000	24.000	24.000	23.762
24.000	24.000	24.000	24.000	0.000
24.000	24.000	24.000	24.000	0.000
24.000	24.000	24.000	24.000	24.000
24.000	24.000	24.000	24.000	24.000

donde_participacion	follow_up	death	angina
Primera y tercera	13	No	No
Primera, segunda y tercera	12	No	No
Primera y segunda	6	No	No
Primera y segunda	6	Yes	No
Primera, segunda y tercera	12	No	No
Primera, segunda y tercera	12	No	No
Primera y segunda	7	No	Yes
Primera, segunda y tercera	12	No	No
Primera, segunda y tercera	12	No	No
Primera, segunda y tercera	12	No	No
Primera, segunda y tercera	12	No	No
Primera		No	No
Primera y segunda	6	No	No
Primera, segunda y tercera	12	Yes	No
Primera, segunda y tercera	12	Yes	No
Primera		Yes	No
Primera, segunda y tercera	12	No	No

Figura 11. Variables continuas (izquierda) y variables cualitativas (derecha)

Vamos a crear un filtro de las mujeres (figura 12), seleccionaremos solo los que son hombres (*Male)* (función sex==«Male»).

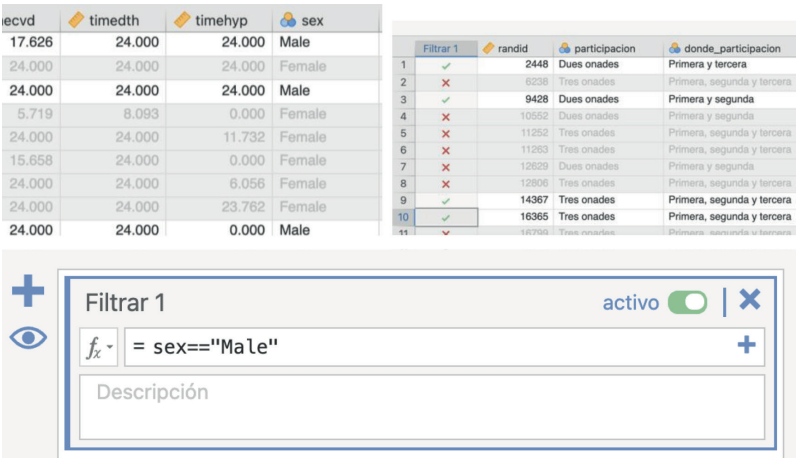


Figura 12. Filtro de los datos por sexo

Otro ejemplo es transformar los datos de una variable, y que pasen a ser 0 y 1, si cumple o no una condición (figura 13).

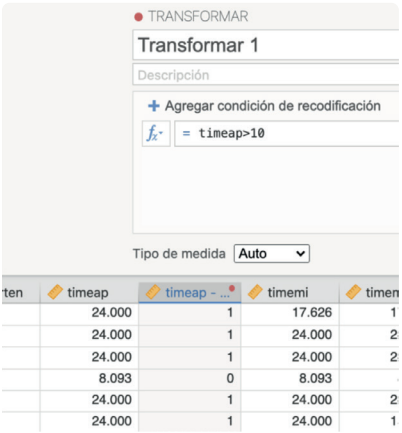


Figura 13. Transformación variable timeap

Estos son solo dos simples ejemplos en los que se utilizan los datos del fichero Framingham. Durante el documento, se realizarán las transformaciones, cálculos o filtros que sean necesarios.

Referencias

Balcells, M. (2016). El estudio Framingham. *Neurosciences and History*. 4(1), 43-46.

The Framingham Heart Study. *The NHLBI Biologic Specimen and Data Repository Information Coordinating Cetner (BioLINCC)*, 2015.

Capítulo 3
Descriptiva de una variable

En este capítulo vamos a desarrollar la descriptiva de una variable. Para ello debemos saber qué es la estadística descriptiva y qué elementos la componen. En este caso nos centraremos en una variable; en el siguiente capítulo lo haremos con más de una variable y realizando la separación, por ejemplo, por clases, tomándolo como una segunda variable.

Las técnicas comunes utilizadas en estadística descriptiva incluyen:

- **Medidas de tendencia central**. Como la media, la mediana y la moda, que representan valores típicos o centrales del conjunto de datos.
- **Medidas de dispersión**. Como la varianza y la desviación estándar, que indican la dispersión o la variabilidad de los datos alrededor de la medida central.
- **Medidas de posición**. Como los percentiles y los cuartiles, que dividen el conjunto de datos en partes iguales o proporcionales.
- **Representaciones gráficas**. Como histogramas, diagramas de barras, diagramas de caja (*boxplots*), diagramas de dispersión, entre otros, que ofrecen una visualización intuitiva de la distribución de los datos y sus características.

La estadística descriptiva proporciona un primer vistazo a los datos y es fundamental en la etapa inicial de cualquier análisis estadístico o estudio, ya que ayuda a entender la naturaleza y la estructura de los datos antes de realizar inferencias más profundas mediante técnicas de estadística inferencial.

A continuación, desarrollaremos el tema de la descriptiva, pero separando en función del tipo de variable, lo que hará que los resultados y la forma de mostrarlos sean diferentes. Además, a partir de aquí, lo desarrollaremos a través de los datos del fichero que tomamos como ejemplo, los datos Framingham.

Variable cuantitativa

Para entender el caso de describir variables continuas, utilizaremos algunas de las variables que hemos descrito en el capítulo anterior

(*age_basal*, *sysbp1*, *diabp1*, *bmi1*, *timecvd*). Además, utilizaremos una descripción que maneje gran cantidad de elementos mostrados en los distintos paquetes de Jamovi para que sea más completa.

Empezaremos por las medidas de **tendencia central**.

La **media** (o media aritmética o promedio) se obtiene con la suma de un conjunto de valores dividida entre el número total de sumandos. Suele representarse por el símbolo $\bar{x}$. Así, se calcularía como

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

donde Σ denota la suma, i es el orden de la observación ($i = 1, 2, \ldots n$), n es el número total de sumandos y x_i es el valor de la variable x para la observación i.

Por ejemplo, consideremos la edad, al inicio del seguimiento, de los primeros 11 sujetos de la base de datos Framingham.

Tabla 1. Edad de 11 sujetos de la base de datos Framingham

randid	age_basal
2448	39
6238	46
9428	48
10552	61
11252	46
11263	43
12629	63
12806	45
14367	52
16365	43
16799	50

La media sería 48,73 años.

$$\bar{x} = \frac{\sum_{i=1}^{n} age\_basal_i}{n} = \frac{39 + 46 + 48 + 61 + 46 + 43 + 63 + 45 + 52 + 43 + 50}{n} = 48,73$$

La **mediana** es la medida de tendencia central que, ordenados los datos de menor a mayor, deja igual número de valores a su izquierda que a su derecha.

Siguiendo el ejemplo anterior, en primer lugar ordenaremos la edad de menor a mayor (también podría ser al revés).

Tabla 2. Edad de 11 sujetos de la base de datos Framingham. Datos ordenados

randid	age_basal
2448	39
11263	43
16365	43
12806	45
6238	46
11252	46
9428	48
16799	50
14367	52
10552	61
12629	63

En este caso, la mediana sería 46, por cuanto deja el mismo número de valores (5 valores) por debajo (o izquierda) que por arriba (o derecha). Si tuviésemos un número de valores par, por ejemplo, solo los diez primeros casos.

Tabla 3. Edad de 10 sujetos de la base de datos Framingham

Sin ordenar

randid	age_basal
2448	39
6238	46
9428	48
10552	61
11252	46
11263	43
12629	63
12806	45
14367	52
16365	43

Ordenadas

randid	age_basal
2448	39
11263	43
16365	43
12806	45
6238	46
11252	46
9428	48
14367	52
10552	61
12629	63

La mediana se calcularía como la media aritmética de los valores en la posición 5 y 6, es decir

$$\text{Mediana} = \frac{46 + 46}{2}$$

que en este caso también es igual a 46.

La **moda** es el valor de la variable que más veces se repite, es decir, el que presenta mayor frecuencia absoluta.

En el ejemplo que utilizamos hay dos modas, 43 y 46, que son las únicas que se repiten dos veces. En este caso se diría que la distribución es bimodal.

Existen varias **medidas de posición** como el valor **mínimo** (39 años en nuestro ejemplo, véase tabla 2) y el **máximo** (63 años en nuestro ejemplo).

Además, los **cuantiles** o **percentiles** son las medidas de posición que generalizan el concepto de mediana. Ordenadas las observaciones de menor a mayor, se define el percentil $p\ (0 < p < 1)$ como el valor que deja $p\%$ observaciones a su izquierda y $(1 - p)\%$ observaciones a la derecha.

En la tabla 2, por ejemplo, el percentil 20 (también denominado segundo decil), corresponde a 43, el tercer valor de edad, ya que deja arriba (izquierda) un 20 % de los valores y abajo (a la derecha) un 80 %.

Algunos percentiles tienen nombre propio, como los deciles (el primer decil acumula el 10 % de los valores, el segundo decil el 20 %, etc.), los terciles (el primer tercil acumula el 33 % de los valores, el segundo el 66 %, etc.) y, sobre todo, los **cuartiles**. El primer cuartil (o Q1) acumula el 25 % de los valores, el segundo (o Q2 o mediana) acumula el 50 % de los valores, el tercer cuartil (o Q3) el 75 % de los valores y el cuarto cuartil (o Q4 o máximo) el 100 % de los valores.

Volviendo a nuestro ejemplo en la tabla 2, el primer cuartil corresponde al tercer valor, igual a 43, ya que acumula el 25 % de los valores; el segundo cuartil o la mediana ya vimos que era igual a 46; y el tercer cuartil corresponde al noveno valor, igual a 52, ya que acumula el 75 % de los valores. Los cuartiles se interpretan como sigue: primer cuartil: el 25 % de los 11 sujetos tienen como máximo 43 años; segundo cuartil o mediana: el 50 % de los 11 sujetos tienen como máximo (y en este caso también como mínimo) 46 años; y tercer cuartil: el 75 % de los 11 sujetos tienen como máximo 52 años.

Por lo que se refiere a las **medidas de dispersión**, explicaremos primero el **rango**, **recorrido o amplitud**, calculado como la diferencia entre el valor máximo y el mínimo. Volviendo a la tabla 2, el rango, máximo-mínimo, $63 - 39$, es de 24 años.

La medida de dispersión de la media es la **varianza** (s^2), que no es más que el promedio de las desviaciones de cada valor de la variable respecto a su media, es decir:

$$\text{Varianza} = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$

El problema es que la varianza no tiene las mismas unidades que la media, por lo que se utiliza su raíz cuadrada, denominada **desviación típica** o **desviación estándar** (s).

Volviendo a nuestro ejemplo:

Tabla 4. Edad de 11 sujetos de la base de datos Framingham

randid	age_basal	age_basal – $\bar{x}$	(age_basal – $\bar{x}$)^2
2448	39	39 – 48,73 = –9,73	94,6729
6238	46	46 – 48,73 = –2,73	7,4529
9428	48	48 – 48,73 = –0,73	0,5329
10552	61	61 – 48,73 = 12,27	150,5529
11252	46	46 – 48,73 = –2,73	7,4529
11263	43	43 – 48,73 = –5,73	32,8329
12629	63	63 – 48,73 =14,27	203,6329
12806	45	45 – 48,73 = –3,73	13,9129
14367	52	52 – 48,73 = 3,27	10,6929
16365	43	43 – 48,73 = –5,73	32,8329
16799	50	50 – 48,73 = 1,27	1,6129
Suma	536	–0,03	556,181
Suma/n	$\bar{x} = \frac{536}{11} = 48,73$		$s^2 = \frac{556,181}{11} = 50,562$ $s = \sqrt{50,562} = 7,1$

La media es igual a 48,73 años, la varianza es igual a 50,562, y la desviación típica igual a 7,1 años. Para ver si la distribución de la variable es muy o poco dispersa, es mejor comparar la desviación típica con la media, lo que resulta en el denominado coeficiente de variación `(CV = (s/) * 100)`. En nuestro caso

```
CV = (7,1 / 48,73) * 100 = 14,57 %
```

Es decir, la distribución de la variable edad entre esos 11 sujetos no es muy dispersa.

La medida de dispersión de la mediana es el **rango (o recorrido) intercuartílico** (IQR), que se calcula como el valor del tercer cuartil menos el valor del primer cuartil, `Q3 – Q1`.

En nuestro caso, `IQR = Q3 – Q1 = 52 – 43 = 9 (años)`. Comparando el IQR con la mediana, que era de 46 años, tampoco parece una distribución muy dispersa (9 es un 19,56 % de 46).

Las **medidas de forma** tratan de identificar ciertas desviaciones en la forma de la distribución de frecuencias de una variable con respecto a un modelo de referencia: la distribución normal, cuya función de densidad adopta una forma característica conocida como campana de Gauss.

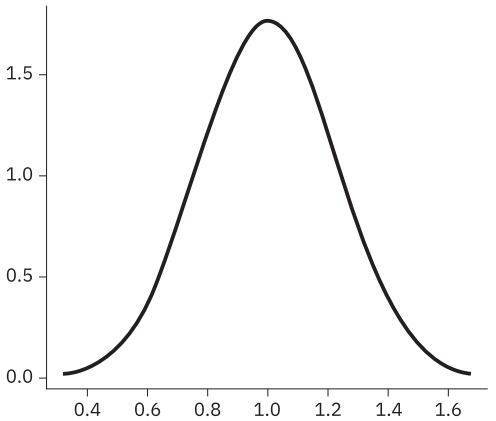


Figura 1. Función de densidad de la distribución normal, campana de Gauss

Las variables cuantitativas continuas siguen una distribución de probabilidad normal, con una función de densidad simétrica, como vemos en la figura 1. Pero esto es en teoría. En la práctica, las variables cuantitativas continuas siguen una distribución similar a la normal, dependiendo del número de observaciones (cuantas más observaciones tenga la variable más se parecerá la distribución) y de la existencia de valores extremos (cuantos menos valores extremos más se parecerá).

Para explicar las medidas de forma es mejor referirse primero a una de las representaciones gráficas, el **histograma**. Un histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados.

La representación del histograma de la variable bmi3, índice de masa corporal correspondiente a la tercera medida, es bastante simétrico, aunque parece más apuntado (veremos más adelante qué significa) que el que correspondería a una variable que se distribuyese como una distribución normal.

Pero, además, se pueden utilizar algunas medidas de forma.

Coeficiente de **asimetría**. Mide si las observaciones están dispuestas de forma simétrica respecto a la media. Si el coeficiente de asimetría es igual a 0, la distribución es simétrica (también denominada sesgada). Cuando el coeficiente sea mayor que 0, la distribución será asimétrica a la derecha (o presenta un sesgo positivo) y cuando sea menor que 0 la distribución será asimétrica a la izquierda (o presenta un sesgo negativo). Cuando la distribución sea simétrica, la media será muy parecida a la mediana y, si es unimodal (solo hay una moda), también a la moda.

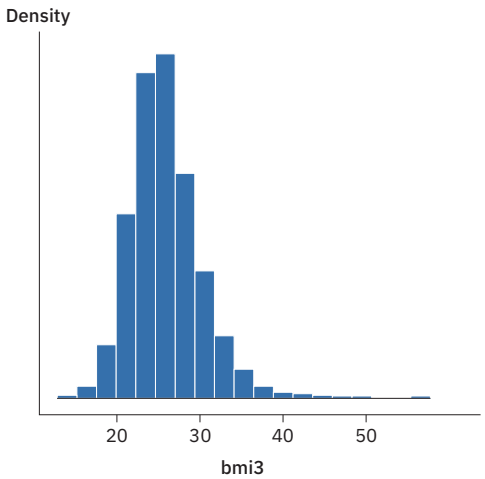


Figura 2. Histograma de la variable bmi3 en la base de datos Framingham

En el caso de la variable bmi3 (tabla 5), vemos que su distribución es bastante simétrica, quizás con cierta asimetría a la derecha, pero no excesiva, ya que la media y la mediana son prácticamente iguales.

Tabla 5. Algunos estadísticos descriptivos para la variable bmi3

Descriptives	
	bmi3
N	3246
Missing	1188
Mean	25.9
Median	25.5
Skewness	0.916
Std. error skewness	0.0430

La distribución de la variable `x2` (simulada), sin embargo, es asimétrica a la izquierda, con un coeficiente de asimetría negativo muy diferente de cero, −3,05. Cuando la distribución es asimétrica a la izquierda la media está a la izquierda de la mediana. En este caso, la media de `x2` es igual a 1,524, mientras que la mediana es igual a 1,622.

Contrariamente, la distribución de la variable `x3` (simulada) es asimétrica a la derecha, con un coeficiente de asimetría positivo muy diferente de cero, 1,47. Cuando la distribución es asimétrica a la derecha la media está

a la derecha de la mediana. En este caso, la media de `x3` es igual a 3,092, mientras que la mediana es igual a 2,497.

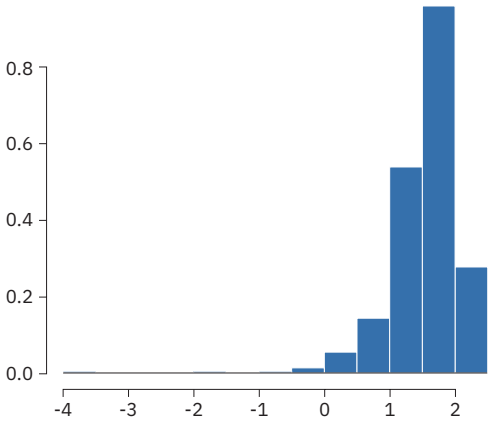


Figura 3. Histograma de la variable simulada x2

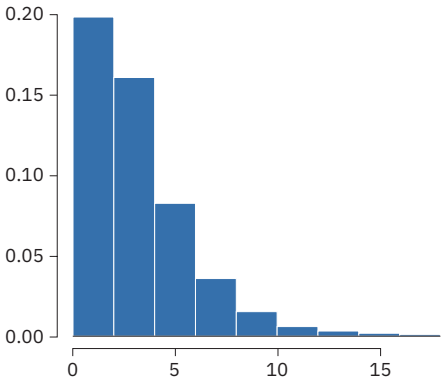


Figura 4. Histograma de la variable simulada x3

De hecho, cuando la distribución de una variable cuantitativa continua no es simétrica, no se deben utilizar la media ni la desviación típica para describirla. En este caso son más representativas la mediana y el rango intercuartílico.

Por lo que se refiere al **apuntamiento** (también denominado curtosis), cuando el coeficiente de apuntamiento sea igual a cero, la distribución será igual de apuntada que la normal (o mesocúrtica), cuando sea mayor

que cero más apuntada (o leptocúrtica) y cuando sea menor que cero menos apuntada o más aplastada (o platicúrtica).

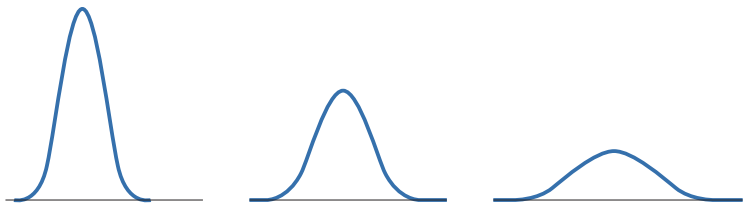


Figura 5. Densidad de variables con distinto apuntamiento

Utilizando Jamovi describiremos, a continuación, un conjunto de variables cuantitativas de la base de datos Framingham.

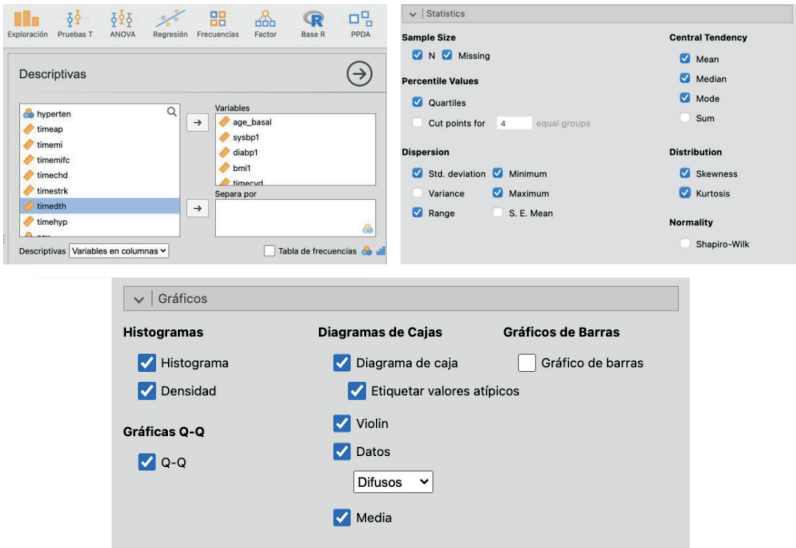


Figura 6. Ventanas que aparecen (Jamovi) en Exploración > Descriptiva

Tabla 6. Variables cuantitativas

	age_basal	sysbp1	diabp1	bmi1	timecvd
N	4434	4434	4434	4415	4434
Perdidos	0	0	0	19	0
Media	49.926	132.91	83.084	25.846	18.667
Mediana	49.000	129.00	82.000	25.450	24.000
Moda	40.000	120.00	80.000	23.480	24.000
Desviación estándar	8.6769	22.422	12.056	4.1018	7.6778
RIC	15.000	26.500	15.000	5.0000	9.9131
Mínimo	32.000	83.500	48.000	15.540	0.0000
Máximo	70.000	295.00	142.50	56.800	24.000
Asimetría	0.19244	1.1479	0.74286	0.98066	–1.1981
Error est. asimetría	0.036773	0.036773	0.036773	0.036852	0.036773
Curtosis	–1.0262	2.0865	1.3312	2.6028	0.039049
Error est. curtosis	0.073530	0.073530	0.073530	0.073688	0.073530
25percentil	42.000	117.50	75.000	23.090	14.087
50percentil	49.000	129.00	82.000	25.450	24.000
75percentil	57.000	144.00	90.000	28.090	24.000

Ahora podemos comentar los resultados de esta tabla de forma genérica. Se observa que tenemos 4434 casos y que solo tenemos datos perdidos en bmi1, que nos podrían mostrar valores faltantes en algún caso. En lo que respecta a la media, son valores que varían según el tipo de datos, desde 18 a 132, pero siendo variables de configuración muy diferente. Las medianas y las modas, muy similares entre ellas.

Esta tabla también nos muestra las medidas de variabilidad de las variables, con una desviación estándar alta y similar y con alto rango intercuartílico. Para finalizar con la tabla, podemos ver que los datos están distribuidos de forma muy parecida a la normal (gaussiana).

En resumen, la tabla descriptiva nos puede mostrar mucha información inicial de los datos para saber cómo tratarlos con posterioridad en los distintos análisis.

Tratamiento gráfico

En lo que respecta al tratamiento gráfico, podemos ver en la figura 6 que, para variables cuantitativas, los gráficos descriptivos pueden ser los histogramas, densidades, gráficos Q-Q y diagramas de cajas. Los diagramas de densidad, también conocidos como gráficos de densidad

de kernel o gráficos de densidad de trazado, son herramientas valiosas para visualizar la distribución de datos continuos. A diferencia de los histogramas, que presentan los datos en barras divididas en intervalos, los diagramas de densidad ofrecen una representación más fluida y precisa de la forma de la distribución. Los gráficos Q-Q nos representan los datos para ver si las variables son o no normales, y los diagramas de cajas, también conocidos como diagramas de cajas y bigotes o *boxplot*, son una herramienta gráfica estandarizada para representar la distribución de un conjunto de datos numéricos.

Componentes de un diagrama de cajas

- **Caja.** La caja principal del diagrama representa el rango intercuartil (IQR) de los datos. El IQR abarca el 50 % central de los datos, excluyendo los valores atípicos.
- **Mediana.** La línea que divide la caja en dos mitades representa la mediana del conjunto de datos.
- **Bigotes.** Se extienden desde la caja hasta los valores mínimo y máximo, sin incluir los valores atípicos.
- **Límites de los bigotes.** Suelen corresponder al primer y tercer cuartil (Q1 y Q3), respectivamente.
- **Valores atípicos.** Los puntos individuales que se encuentran más allá de los bigotes se representan como círculos o estrellas y se consideran valores atípicos.

En las siguientes figuras (7 y 8), podemos ver los gráficos de las variables. En el caso de la figura 7, apreciamos dos distribuciones bastante diferentes, aunque en los dos casos cumplen normalidad. En el caso de *sysbp1*, se muestran valores atípicos, marcados con recuadros de números en los diagramas de cajas, que en este caso son, por ejemplo, los puntos 501 o 3649. Esto nos permite apreciar posibles errores en las muestras o valores atípicos que podrían (o no) ser eliminados previamente a los análisis. También hay que comentar que la simetría de las distribuciones es muy diferente y con mayor apuntamiento de los datos de *sysbp1*, es decir, más concentrados y con menor variabilidad.

En el caso de la figura 8, donde tenemos tres variables, se ha añadido una temporal, *timecvd*, que tiene un comportamiento similar a todas las variables tiempo del fichero, por lo que su comportamiento es similar, apareciendo en el histograma la mayor densidad en los valores más altos. Podemos ver, también, que los histogramas de las variables *bmi1* y *diabp1* son muy similares y característicos de las variables con distribución gaussiana, pero, aunque tengan un apuntamiento similar,

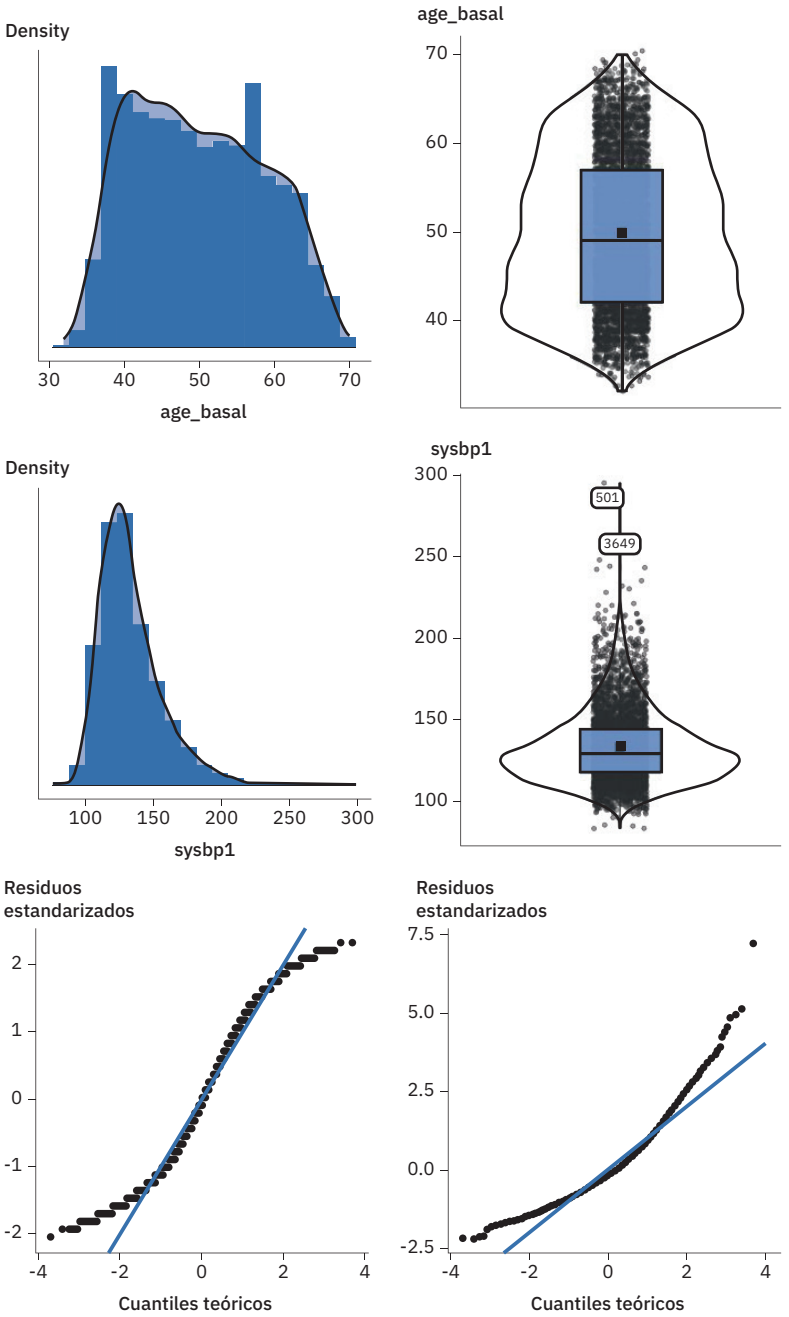


Figura 7. Histograma con densidad (izquierda) y diagrama de cajas (derecha). En la fila final, diagramas Q-Q. Age basal y sysbp1

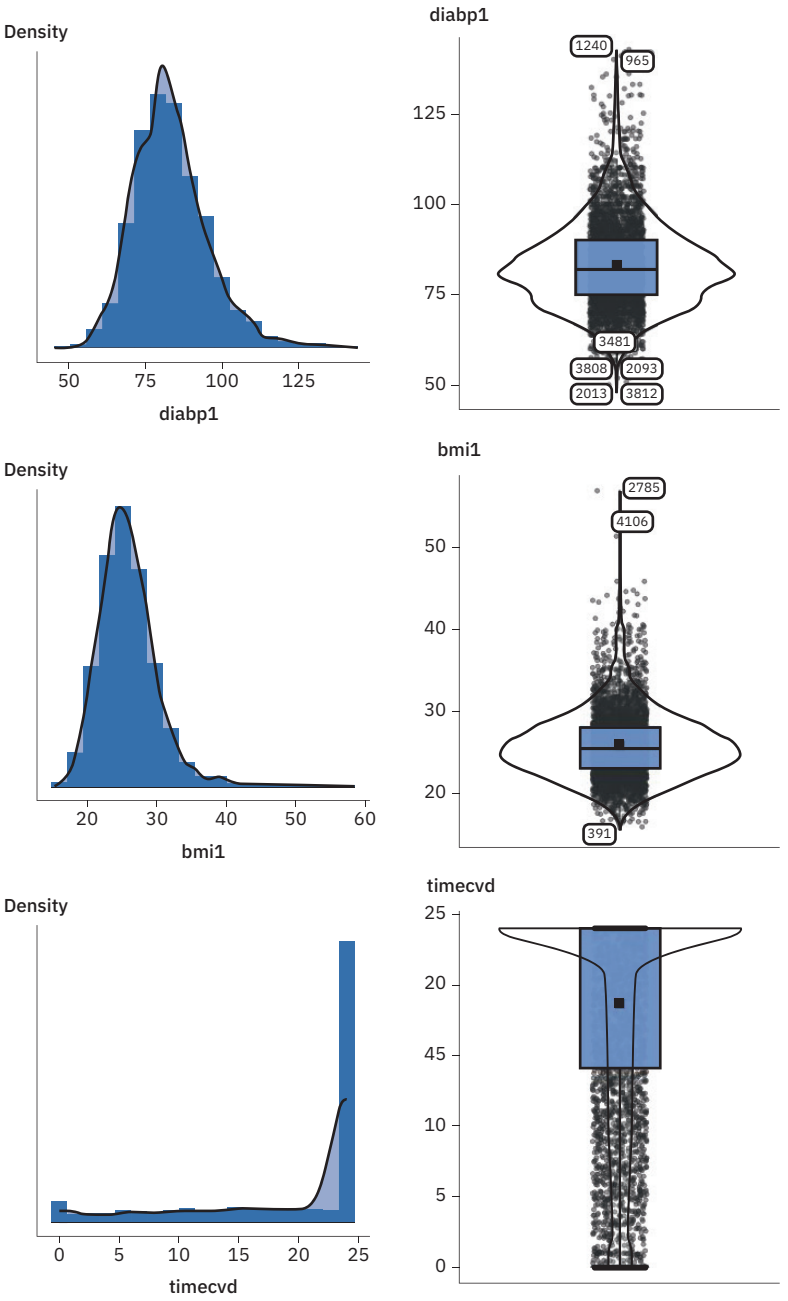


Figura 8. Diagrama Q-Q (izquierda) y diagrama de cajas con datos y violín (derecha). Diabp1, Bmi1 y Timecvd

sysbp1 tiene valores hacia su derecha, por lo que parece que el máximo está desplazado a la izquierda.

Si observamos también los diagramas de caja de la figura 8, podemos ver que la variable *diabp1* tiene mayor número de datos atípicos, que podrían ser estudiados de forma separada.

En todos los diagramas de cajas, se está introduciendo la máxima información, como el punto negro central, que es la media, el tamaño de caja, el conjunto de los puntos que son los valores, los datos atípicos y la densidad en los dos lados en forma de violín, que nos muestra qué valores se repiten más.

Comentar finalmente en estas gráficas, más concretamente en la figura 7, los gráficos que se muestran Q-Q de las variables *age_basal* y *sysbp1*, para ver si estos datos se distribuyen de forma normal, es decir, que los datos de los residuos están lo más próximo a la diagonal, algo que ocurre en los dos casos.

Variables cualitativas

Para esta segunda parte, las variables cualitativas que utilizaremos también serán las descritas en el capítulo anterior como variables dependientes no continuas (*death*, *angina*, *hospmi, mi_fchd*, *anychd*, *stroke*, *cvd* y *hypertens*). Podemos ver que estas ocho variables solo tienen dos opciones, si-no, por lo que tendremos variables dicotómicas, es decir, nominales. Por otro lado, para tomar como ejemplos variables cualitativas politómicas, trabajaremos también con dos variables ordinales: *age_basal_cat* y *Garrow*.

En lo que respecta al procedimiento, es el mismo seguido para las variables cuantitativas.

Variables dicotómicas

Dentro de las variables cualitativas, tenemos las dicotómicas, que podemos ver las tablas de las comentadas con anterioridad (tabla 7).

Tabla 7. Frecuencia de las variables dicotómicas

Frequencias de death

death	Frecuencias	% del total	% acumulado
No	2884	65.043 %	65.043 %
Yes	1550	34.957 %	100.000 %

Frequencias de mi.fchd

mi.fchd	Frecuencias	% del total	% acumulado
No	3703	83.514 %	83.514 %
Yes	731	16.486 %	100.000 %

Frequencias de angina

angina	Frecuencias	% del total	% acumulado
No	3709	83.649 %	83.649 %
Yes	725	16.351 %	100.000 %

Frequencias de hospmi

hospmi	Frecuencias	% del total	% acumulado
No	3980	89.761 %	89.761 %
Yes	454	10.239 %	100.000 %

Frequencias de anychd

anychd	Frecuencias	% del total	% acumulado
No	3194	72.034 %	72.034 %
Yes	1240	27.966 %	100.000 %

Frequencias de stroke

stroke	Frecuencias	% del total	% acumulado
No	4019	90.641 %	90.641 %
Yes	415	9.359 %	100.000 %

Frequencias de cvd

cvd	Frecuencias	% del total	% acumulado
No	327	73.906 %	73.906 %
Yes	1157	26.094 %	100.000 %

Frequencias de hyperten

hyperten	Frecuencias	% del total	% acumulado
No	1182	26.658 %	26.658 %
Yes	3252	73.342 %	100.000 %

En las tablas anteriores podemos ver que todas estas variables son dicotómicas, es decir, dos posibilidades (No-Sí). Además, tienen distintas frecuencias para los dos casos, no son homogéneas.

Para su mejor comprensión, podemos ver la forma de los diagramas de barras de cada una de ellas (figuras 9 y 10). Como es normal en este tipo de datos, hay más casos de NO que de SÍ (que cumplan la condición).

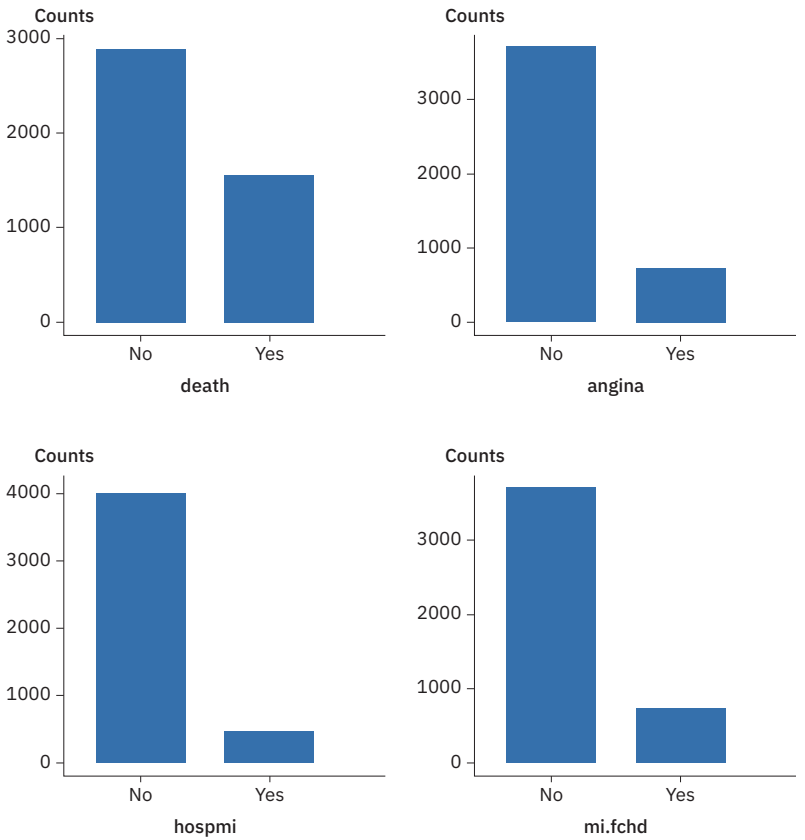


Figura 9. Diagramas de barras (I)

Un diagrama de barras, también conocido como gráfico de barras o gráfico de columnas, es una forma común de representar gráficamente datos categóricos o discretos. Sus componentes son:

- **Barras.** El elemento principal del diagrama está formado por barras rectangulares cuya altura o longitud es proporcional a la frecuencia o magnitud de cada categoría.

- **Eje X.** El eje horizontal (X) representa las categorías o variables cualitativas que se están analizando.
- **Eje Y.** El eje vertical (Y) representa la frecuencia o magnitud de cada categoría. La escala del eje Y debe ser apropiada para los valores que se representan.
- **Títulos y etiquetas.** El diagrama debe incluir un título claro que describa la variable que se está representando, así como etiquetas en los ejes que identifiquen cada categoría.

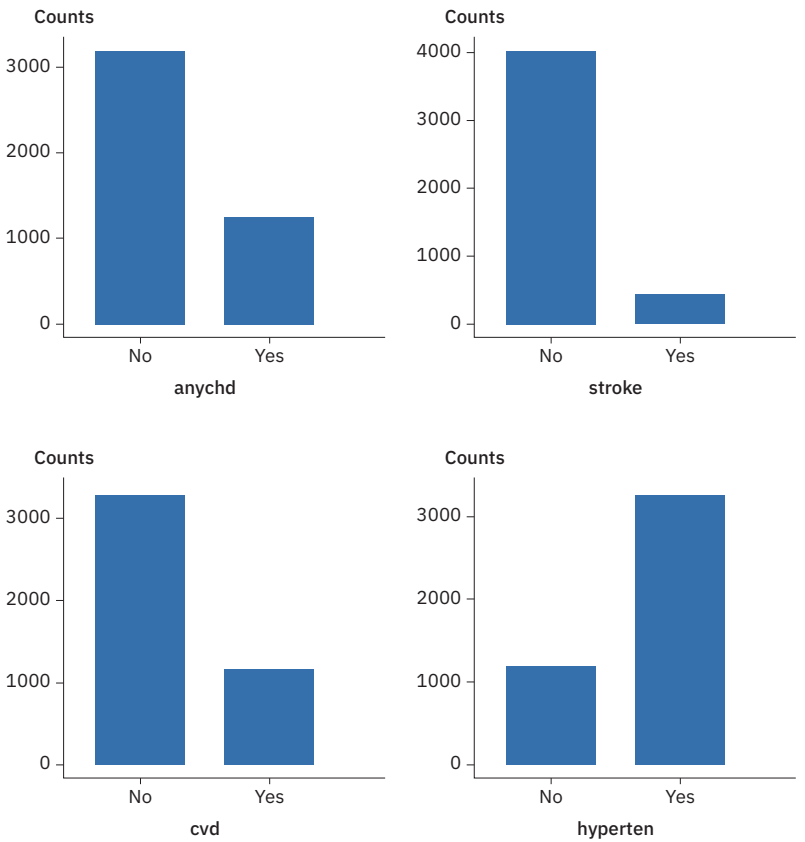


Figura 10. Diagramas de barras (II)

Variables politómicas

Para analizar las variables cualitativas politómicas, describiremos las variables *Garrow* y *age_basal_cat*, que tienen sus valores ordenados. Veremos primero una descriptiva de las mismas, en formato tabla, de la misma forma que se ha efectuado con las variables continuas. Tenemos la descriptiva de los casos en la tabla 3, donde nos marca el número de perdidos distinto a 0 para cada caso de *Garrow*, algo normal por el tipo de datos. Las tablas de frecuencia (tabla 8), nos muestran la distribución de los casos en los *Garrow*, que son tres, y en los *age_basal_cat*, que son cuatro. En ambos casos, ordenados por ser variables ordinales.

Tabla 8. Frecuencias de las variables Garrow y age_basal_cat

Frequencias de Garrow1

Garrow1	Frecuencias	% del total	% acumulado
Normoweight	1993	45.142 %	45.142 %
Overweight	1845	41.789 %	86.931 %
Obese	577	13.069 %	100.000 %

Frequencias de Garrow2

Garrow2	Frecuencias	% del total	% acumulado
Normoweight	1744	44.298 %	44.298 %
Overweight	1662	42.215 %	86.513 %
Obese	531	13.487 %	100.000 %

Frequencias de Garrow3

Garrow3	Frecuencias	% del total	% acumulado
Normoweight	1462	45.040 %	45.040 %
Overweight	1324	40.789 %	85.829 %
Obese	460	14.171 %	100.000 %

Frequencias de age_basal_cat

age_basal_cat	Frecuencias	% del total	% acumulado
Q1 - 32-42	1110	25.034 %	25.034 %
Q2 42-49	1141	25.733 %	50.767 %
Q3 49-57	1019	22.982 %	73.748 %
Q4 57-70	1164	26.252 %	100.000 %

La figura 11 nos muestra la distribución gráfica, mediante el diagrama de barras, de las distintas posibilidades de estas dos variables ordinales, una más homogéneamente distribuida, la *age_basal_cat*, y otra con diferencia de casos, *Garrow*.

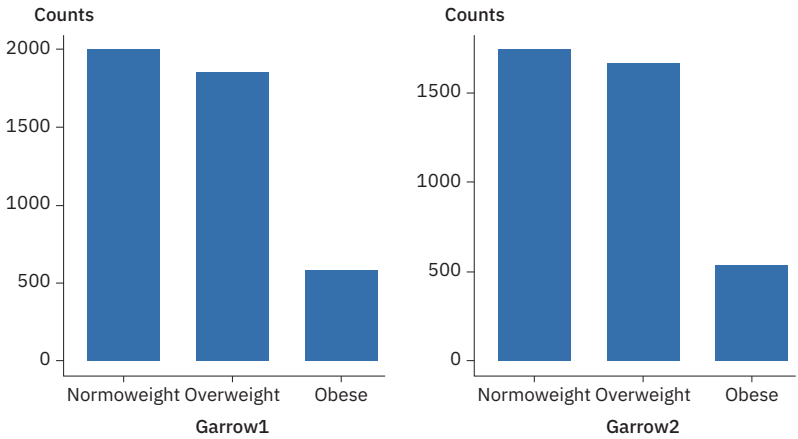


Figura 11. Diagramas de barras (III)

Referencias

The Jamovi project (2022). *Jamovi*. (Versión 2.3) [Computer Software]. Recuperado de: https://www.JAMOVI.org

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

Linlin Yan (2020). *Venn Diagram by ggplot2, with really easy-to-use API*. [R package]. Recuperado de: https://github.com/yanlinlin82/ggvenn

Capítulo 4
Descriptiva de dos o más variables

La técnica de estratificación resulta muy útil para trabajar con conjuntos de datos. Estratificar consiste en repetir el análisis, en este caso descriptivo, tanto de las tablas como de los gráficos, por categorías de variables cualitativas. Por ejemplo, el primer caso, tenemos descritas las variables cuantitativas *age_basal*, *sysbp1*, *diabp1*, *bmi1*, separadas según si la variable *death* es No o Sí (tabla 1). A continuación, los diagramas de cajas de las mismas variables, que en este caso se denominan diagramas de cajas en paralelo, igualmente estratificado por la variable *death* (figura 1).

Tabla 1. Variables

Descriptivas					
	death	**age_basal**	**sysbp1**	**diabp1**	**bmi1**
N	No	2884	2884	2884	2878
	Yes	1550	1550	1550	1537
Media	No	47.241	127.92	81.367	25.544
	Yes	54.922	142.18	86.278	26.411
Mediana	No	46.000	125.00	81.000	25.110
	Yes	56.000	138.00	85.000	26.050
Desviación estándar	No	7.6705	18.538	10.795	3.9110
	Yes	8.2194	25.822	13.541	4.3829

Otra posibilidad es no solo separar o estratificar por una clase, sino por dos. En este siguiente caso, las variables *age_basal* y *timesysbp1hyp*, separadas según *death* y *stroke* a la vez (figura 2).

En el caso de dos variables cualitativas, por ejemplo, el caso de *Garrow1* y *sex*, podemos hacer el diagrama de barras de la primera y ver cómo varía por la segunda, en la parte superior, o ver los casos en diagramas de barras de *Garrow1* según la *edad_basal_cat* (figura 3).

Es interesante observar gráficamente la variación en el conteo de los casos de *Garrow1* al separar por género, ya que el caso de *Overweight* tiene un significado diferente en comparación con *Normoweight* y *Obese*, mientras que en el caso de la separación según *age_basal_cat*, podemos ver un comportamiento diferente de *Normoweight* respecto a los otros dos casos.

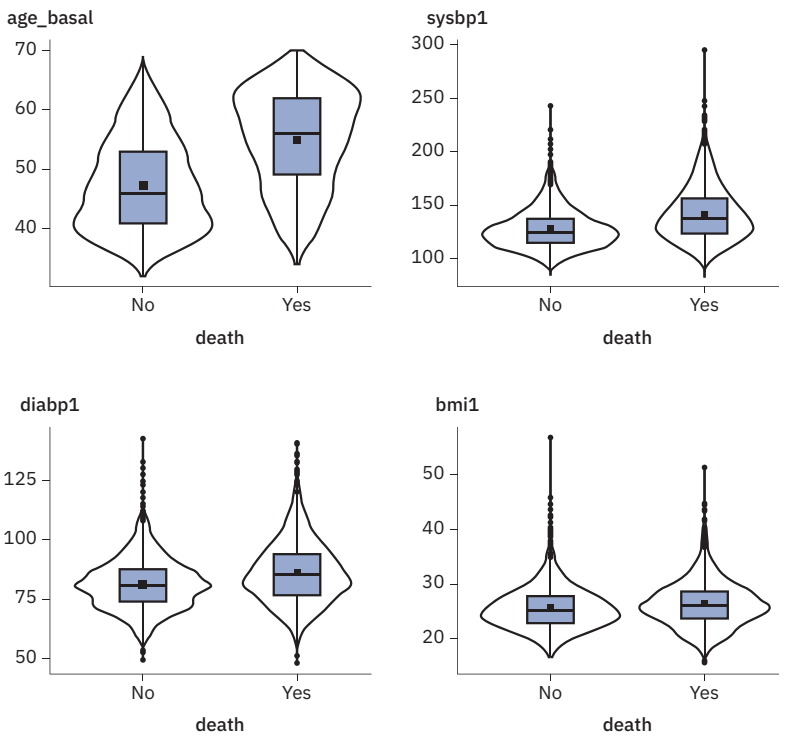


Figura 1. Diagramas de cajas en paralelo, de las variables age_basal, sysbp1, diabp1, bmi1, según si la variable death

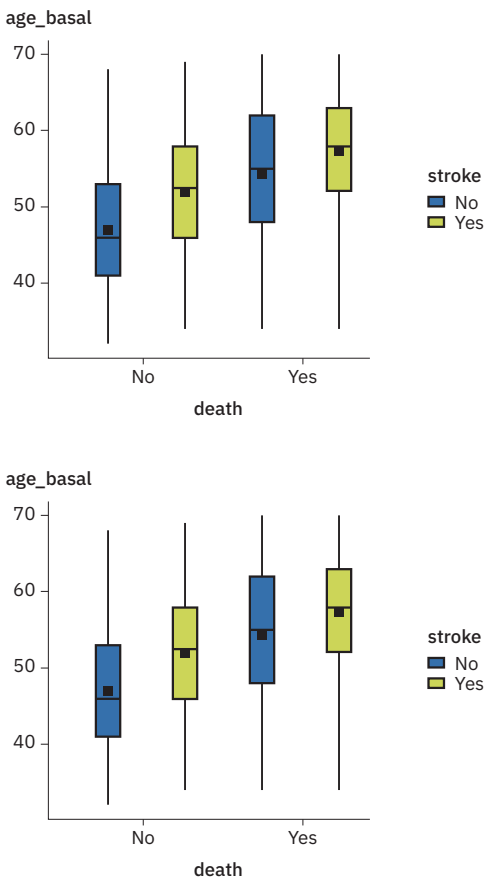


Figura 2. Diagramas de cajas en paralelo (variables age_basal y timesysbp1hyp, separadas según death y stroke)

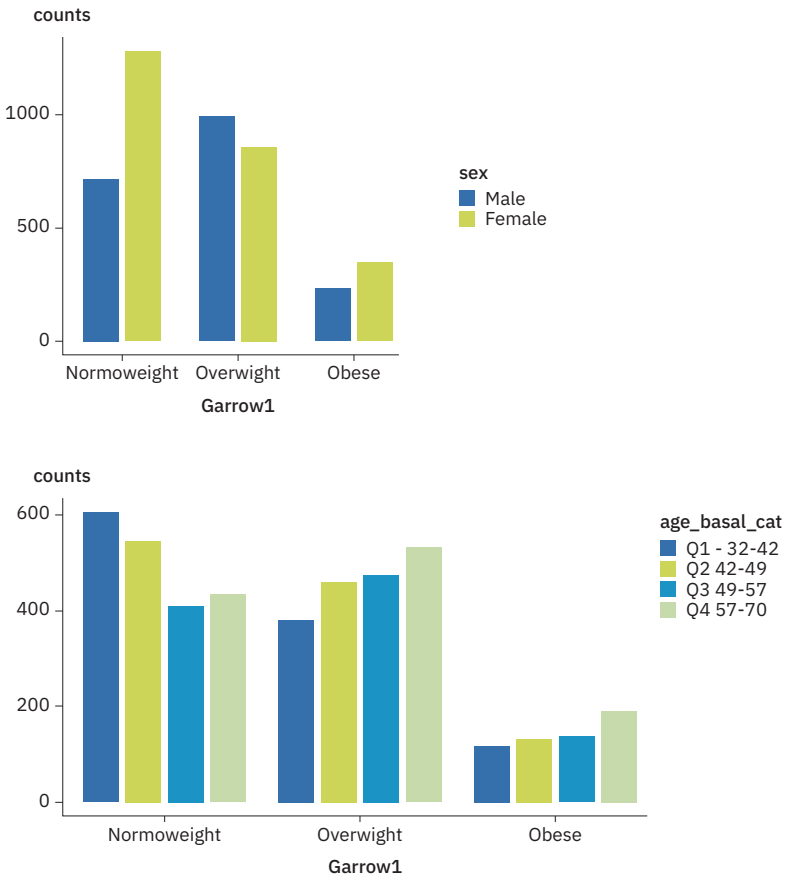


Figura 3. Diagramas de barras. Superior: Garrow 1, según sexo. Inferior: Garrow1, según age_basal_cat

Otros tratamientos gráficos

Además de los diagramas de barras, histogramas o gráficos de cajas ya mostrados, tenemos otras posibilidades descriptivas.

Una opción inicial es utilizar diagramas de Venn para las variables dicotómicas y analizar las relaciones entre ellas, ya sea en cantidades o en porcentajes.

Podemos trabajar con dos, tres o cuatro variables, como se aprecia en los gráficos (figura 4).

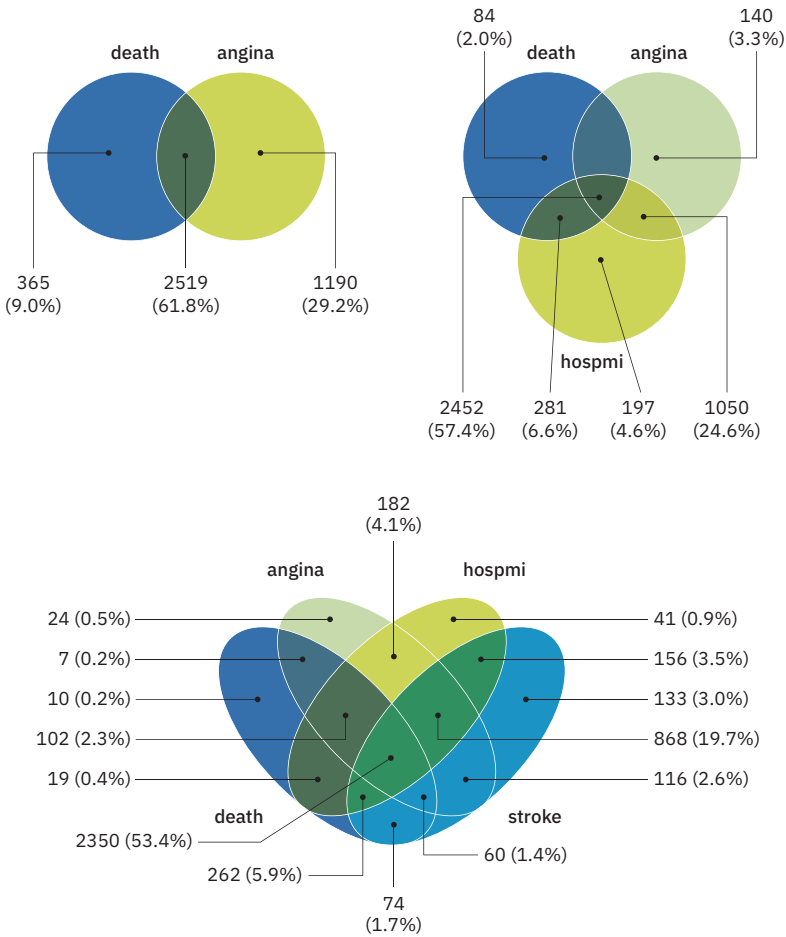


Figura 4. Diagramas de Venn

Estos gráficos nos muestran tanto las relaciones como las cantidades y porcentajes de casos, lo que puede ayudar a ver la relación entre variables. Si nos centramos en los primeros diagramas, vemos que el 61,8 % de los casos del fichero que tenían anginas ha fallecido y que solo un 9 % de los que han fallecido no es por anginas. En el siguiente diagrama de Venn, podemos ver que en un 57,4 % de los casos coinciden las anginas con *hospmi*, que han fallecido. Finalmente, en el último caso

se puede ver la relación en diagramas de Venn de cuatro variables, donde en un 53,4 % coinciden todas.

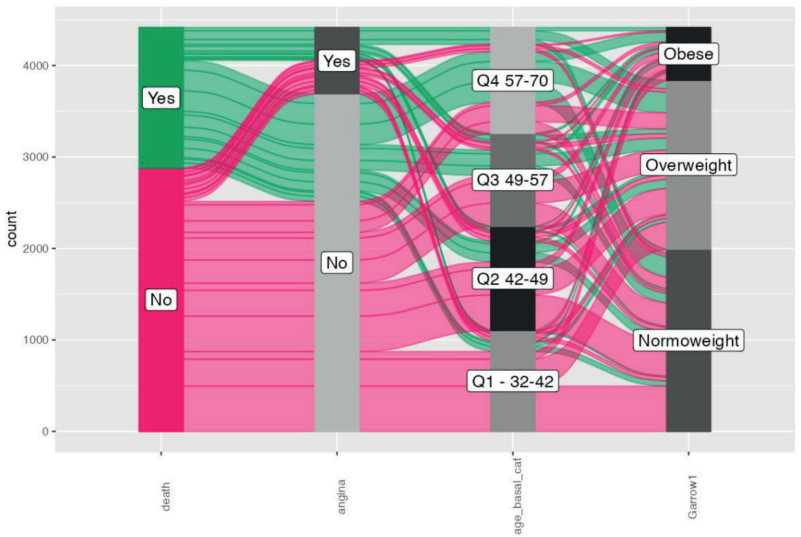


Figura 5. Diagramas Alluvial (I)

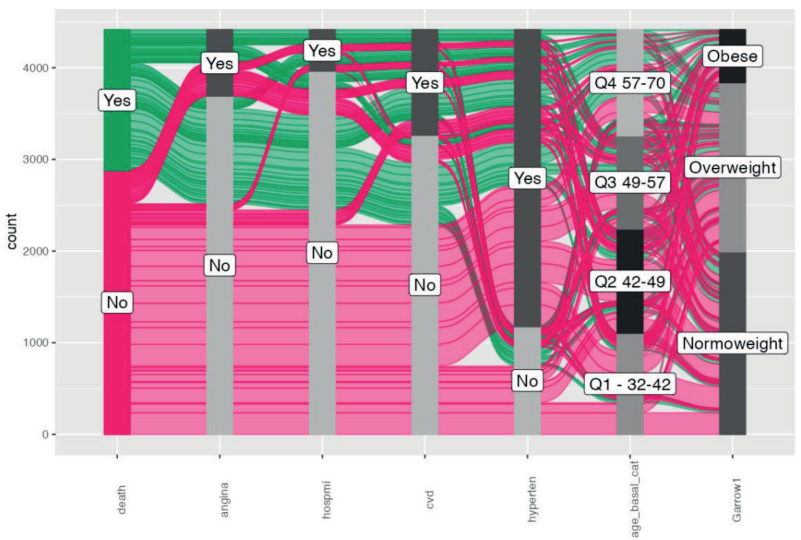


Figura 6. Diagramas Alluvial (II)

Otra opción es el gráfico Alluvial. En el primer caso relacionamos solo cuatro variables (figura 5) mientras que, en el siguiente, la relación es de más variables (figura 6).

¿Qué utilidad tendrán, en la descriptiva de variables, los gráficos Alluvial? Nos permitirán, entre otras cosas, ver la cantidad de relaciones diferentes que hay entre variables y ver cuál se repite más. Por ejemplo, el paso de *death* a *angina* muestra los casos (en verde) de *death* que son con anginas (Yes) y los que no tienen anginas (No). Además, podemos ver que tenemos casos de *death* y *no death* tanto para gente con anginas como para gente sin anginas. En el supuesto que tuviésemos una relación entre dos variables en las que un caso no se diese —por ejemplo, *no death* con *no angina*— podría establecerse una colinealidad entre variables. En resumen, se trata de un gráfico descriptivo entre variables cualitativas y en el que se muestran relaciones directas.

Population	Female	Male
(75,76]	.	1
(70,75]	56	87
(65,70]	206	272
(60,65]	224	351
(55,60]	281	370
(50,55]	311	419
(45,50]	370	446
(40,45]	236	276
(35,40]	7	17

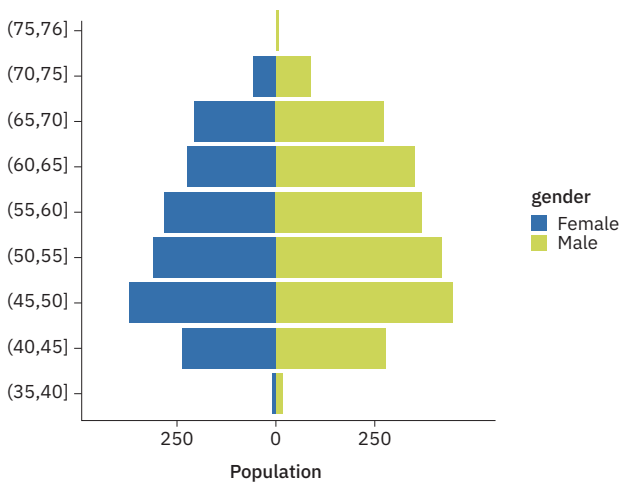


Figura 7. Diagramas pirámide poblacional

Finalmente, se presenta un ejemplo de pirámide de población, con la utilización de los datos de las variables *edad* y *sexo* de Framingham (figura 7), además del árbol de relaciones (figura 8).

Como ejemplo final, se observa en la figura 7, la población femenina supera a la masculina en todos los rangos de edades en este estudio, de forma general tenemos 2490 mujeres y 1944 hombres. En el caso de la figura 8, esta nos muestra cómo se distribuyen los casos, a partir de *angina* (Yes-No) para cada fase, gráfico que podemos realizar con cualquiera de las variables tanto dicotómicas como polinómicas. Con ello, hemos visto algunas pinceladas de la obtención de gráficos descriptivos de variables cuantitativas y cualitativas que nos van a servir como paso previo al análisis más pormenorizado de los datos que haremos más adelante y que, además, nos sirven para conocer mejor las variables.

Referencias

The Jamovi project (2022). *Jamovi*. (Versión 2.3) [Computer Software]. Recuperado de https://www.JAMOVI.org

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

Serdar Balci (2022). *ClinicoPath Jamovi Module doi:10.5281/zenodo.3997188*. [R package]. Recuperado de: https://github.com/sbalci/ClinicoPathJAMOVIModule https://www.serdarbalci.com/ClinicoPathJamoviModule/

Bjoern Koneswarakantha (2019). *easyalluvial: Generate Alluvial Plots with a Single Line of Code.* [R package]. Recuperado de: https://CRAN.R-project.org/package=easyalluvial

Nick *Barrow*man (2020). *vtree: Display Information About Nested Subsets of a Data Frame*. [R package]. Recuperado de: https://CRAN.R-project.org/package=vtree

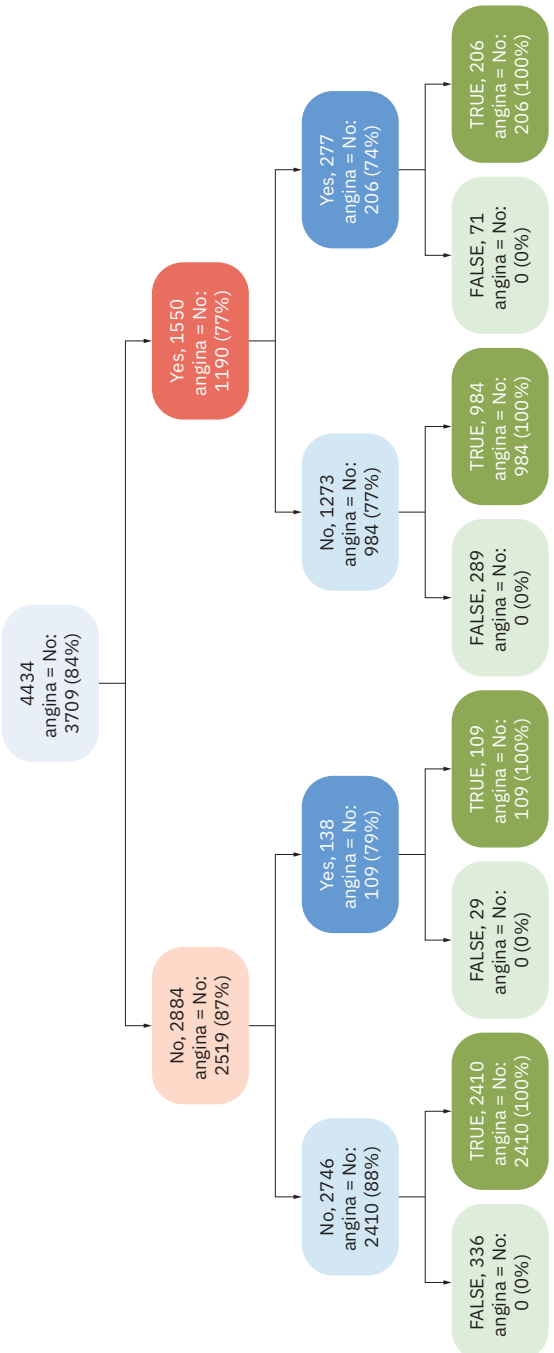


Figura 8. Diagrama de árbol

Capítulo 5
Introducción a la inferencia con una variable

En este capítulo desarrollaremos la inferencia estadística con una variable. El objetivo del capítulo es guiar en la realización de operaciones con Jamovi relativas a la inferencia matemática o estimación de parámetros poblacionales a partir de datos muestrales.

Primero se tratará la construcción de intervalos de confianza para una variable y, a continuación, los contrastes de hipótesis de una variable en relación con un parámetro teórico. En ambos casos se distinguirá entre tipos de variables diferentes: estimación de proporciones para variables cualitativas nominales/ordinales, estimación de medias para variables continuas).

Para dicho propósito, trabajaremos con las variables continuas siguientes:

- **sysbp**. Presión arterial sistólica (media de las dos últimas de tres mediciones) (mmHg).
- **diabp**. Presión arterial diastólica (media de las dos últimas de tres mediciones) (mmHg).
- **imc**. Índice de masa corporal, peso en kilogramos/altura en metros cuadrados.

Además de las ya comentadas en capítulos anteriores, como son *hyperten* y *angina,* que son variables cualitativas dicotómicas, es decir, que solo tienen dos niveles, yes/no.

Conceptos básicos: población, muestra, estimación y contrastes

La **inferencia estadística** es la parte de la estadística que trata de las condiciones y procedimientos bajo los cuales podemos extraer conclusiones o inferencias de la población a partir de la información muestral. Su objetivo es obtener conclusiones (estimaciones, decisiones, predicciones u otras generalizaciones) útiles para hacer razonamientos deductivos sobre una totalidad, basándose en la información numérica dada por la muestra.

Vamos a dar algunas definiciones para los conceptos que acabamos de nombrar.

- **Universo**. Es el conjunto de individuos u objetos que consideramos en un estudio estadístico. Dentro de un mismo universo podemos distinguir varias poblaciones.
- **Población**. Son todos los valores que adopta la característica (o características) que queremos estudiar en el universo. Por ejemplo: si estamos analizando solo una característica, la edad de los españoles, la población sería el conjunto de las edades de todos los españoles y el universo sería el conjunto de todos los españoles.
- **Muestra**. Es cualquier subconjunto de la población. Para que la muestra nos sirva para extraer conclusiones sobre la población deberá ser lo más representativa posible.
- **Muestreo** (o **técnicas de muestreo**). Son los métodos para elegir las muestras.

La inferencia (estadística) pueden tomar la forma de estimaciones de unas características numéricas (**estimación**), respuestas a preguntas sí/no o verdadero/falso (**contraste o prueba de hipótesis**), pronósticos de futuras observaciones, descripciones de asociación (o correlación) o modelización de relaciones entre variables, así como análisis de la varianza, series de tiempo y minería de datos.

La estimación se divide en **estimación puntual** y **estimación por intervalo**.

La **estimación puntual** consiste en dar un valor numérico a un parámetro poblacional desconocido (por ejemplo, la proporción poblacional, la media poblacional, etc.). Aunque existen infinitos estimadores puntuales para cada parámetro poblacional solo uno de ellos es óptimo. Un estimador óptimo, denominado ELIO, es un estimador **lineal** (obtenidos como una combinación lineal), **insesgado** (se espera que su valor coincida con el verdadero valor del parámetro poblacional) y de **varianza mínima** (de entre todos los estimadores insesgados el que tenga mínima varianza).

Ilustraremos estos conceptos con la figura 1. En ella tenemos las dianas de 4 tiradores sobre las que han realizado 10 disparos cada uno. Si el parámetro poblacional desconocido fuese el centro de la diana y el estimador cada uno de los disparos, los tiradores 1 y 3, con todos los disparos en torno al centro de la diana serían insesgados, mientras que los tiradores 2 y 4, con disparos lejos de la diana serían sesgados. Entre los estimadores insesgados el tirador 3 tiene menos dispersión, es decir menos varianza, y el tirador 1 mayor dispersión (varianza). El estimador ELIO sería por tanto el tirador 3.

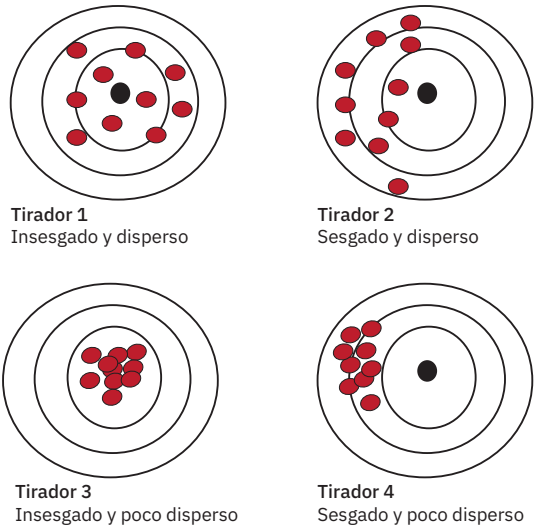


Figura 1. Estimadores sesgados e insesgados, con distinta varianza

El estimador óptimo de la media poblacional (representado por la letra griega μ) es la media muestral (representado por $\bar{x}$) y el de la proporción poblacional (representado por la letra griega π) es la proporción muestral (representada por p). Es decir, si dispusiésemos de una muestra (representativa) de una población con información del sexo y de la edad de un conjunto de sujetos, la proporción de mujeres en la muestra sería un estimador (óptimo) de la proporción de mujeres en la población y la edad media de los sujetos un estimador de la edad media de la población. Lo que ocurre es que si dispusiésemos de otra muestra de la población (por ejemplo, obtenida en otro momento del tiempo), ni la proporción de mujeres en la muestra ni la edad media de los sujetos serían la misma (aunque podrían parecerse). Ahora bien, si esta segunda muestra fuese representativa, los estimadores obtenidos a partir de ella también serían óptimos. Este fenómeno se conoce como **variabilidad muestral** que, teóricamente, se define como la variación de los estimadores (puntuales) de un parámetro entre muestras representativas de una población. De hecho, si un estimador no tiene variabilidad muestral tenemos un error. La variabilidad muestral se mide mediante el error estándar.

Pondremos un ejemplo para aclarar todos estos conceptos. Supongamos que tenemos una población compuesta por 5 sujetos, 1,2, 3, 4 y 5. La media poblacional (que desconocemos) es, por tanto, igual a 3

$$\mu = \frac{1 + 2 + 3 + 4 + 5}{5} = 3$$

Para estimar este parámetro poblacional haremos todas las muestras de tamaño 2 (sin repetición) que podamos de la población (tabla 1).

Tabla 1. Muestras de tamaño 2 de la población ficticia del ejemplo

(1,2)	(2,3)	(3,4)	(4,5)
(1,3)	(2,4)	(3,5)	
(1,4)	(2,5)		
(1,5)			

Como dijimos, el estimador óptimo de la media poblacional es la media muestral. Así que calcularemos las medias muestrales de cada una de las muestras (tabla 2). Así, por ejemplo, en la muestra (1,2), la media muestral sería

$$\bar{x} = \frac{1+2}{2}, \text{ etc.}$$

Tabla 2. Medias muestrales de las muestras de la población ficticia del ejemplo

	$\bar{x}$		$\bar{x}$		$\bar{x}$		$\bar{x}$
(1,2)	1,5	(2,3)	2,5	(3,4)	3,5	(4,5)	4,5
(1,3)	2	(2,4)	3	(3,5)	4		
(1,4)	2,5	(2,5)	3,5				
(1,5)	3						

De la tabla podemos observar, en primer lugar, que las medias muestrales varían mucho, es decir, que existe variabilidad muestral. En segundo lugar, que excepto en las muestras (1,5) y (2,4), las medias muestrales no son iguales que el valor de la media poblacional (que, recordemos, era igual a 3). Entonces, ¿qué significa que la media muestral sea un estimador insesgado de la media poblacional? Si hacemos la media de todas las medias muestrales

$$\bar{x} = \frac{1,5+2+2,5+3+3,5+4+4,5}{10} = \frac{30}{10}$$

Es decir, significa que si tuviésemos todas las muestras posibles de una población el valor (medio) del estimador sería igual que el valor del parámetro. Pero nunca se dispone de todas las muestras de una población, sino usualmente de una sola (a veces, con suerte de 2 o 3, pero nunca de todas).

Siguiendo con nuestro ejemplo, si tuviésemos la muestra (1,5) o la (2,4), el valor del estimador coincidiría con el verdadero valor del parámetro. Pero ¿y si trabajamos con la muestra (2,5), por ejemplo? ¿Podemos estar seguros de que el valor del estimador (puntual) (igual a 3,5) es igual al valor del parámetro poblacional? No. Lo único que sabemos es que 3,5 es un estimador insesgado del verdadero parámetro poblacional.

Por eso necesitamos acompañar la estimación puntual de la **estimación por intervalo**. Se trata de un rango de valores en el que con una determinada probabilidad (denominada grado de confianza) está contenido el parámetro poblacional.

En nuestro ejemplo, si utilizamos la muestra (2,5), el estimador puntual será igual a 3,5 y el estimador por intervalo al 95 % es igual a (–0,88, 7,88) (la fórmula del estimador por intervalo de la media muestral la explicaremos más adelante). Es decir, con un 95 % de probabilidad el verdadero valor del parámetro estará situado entre –0,88 y 7,88. Observad cómo el intervalo es muy amplio. Eso es así porque el tamaño muestral es muy pequeño (igual a 2). Cuanto más grande sea el tamaño muestral, menor será el intervalo y cuanto menor sea el intervalo mayor precisión tendrá el estimador.

En este caso utilizamos dos muestras. Una, la muestra (2,5), que ya hemos utilizado, y la segunda, una nueva muestra formada por la unión de las muestras (1,2), (1,3) y (1,4), es decir muestra (1,2,3,4). El estimador puntual de la media poblacional utilizando esta nueva muestra es 2,5

$$\overline{x} = \frac{1+2+3+4}{4} = \frac{10}{4} = 2,5$$

y el intervalo de confianza al 95 % es (–0,604,5,604). Podemos plantearnos varias preguntas:

- ¿Por qué la amplitud del intervalo en la muestra (2,5) (7,88 – (–0,88) = 8,76) es mayor que en la muestra (1,2,3,4) (5,604 – (–0,604) = 6,208)?
- Porque el tamaño de la muestra es menor (2 vs. 4)
- ¿Qué estimador es más válido? ¿el de la muestra (2,5), igual a 3,5, o el de la muestra (1,2,3,4), igual a 2,5?

Los dos son igual de válidos, lo que ocurre es que el de la muestra (1,2,3,4) es más preciso.

Estadística descriptiva

Primeramente, veamos cómo son las variables continuas de forma descriptiva, datos estadísticos e histogramas. Hacemos notar que esas variables también se han obtenido en los tres períodos del estudio, por lo que aplicaremos inferencia en los tres momentos para las tres variables.

Podemos ver en la tabla 3 que, mientras que BMI (IMC) se mantiene de forma media constante en los tres tiempos, tanto *sysbp* como *diabp* han variado, la primera aumentando en el tiempo y la segunda con disminución en el último período (bajada bastante significativa).

Tabla 3. Descriptiva de las variables en los tres momentos temporales del estudio

	sysbp1	sysbp2	sysbp3	diabp1	diabp2	diabp3	bmi1	bmi2	bmi3
N	4434	3930	3263	4434	3930	3263	4415	3914	3246
Media	132.91	136.95	140.22	83.084	84.020	81.793	25.846	25.898	25.895
Desviación estándar	22.422	22.544	22.928	12.056	11.427	11.271	4.1018	4.1225	4.0807
Mínimo	83.500	88.000	86.000	48.000	47.000	30.000	15.540	15.330	14.430
Máximo	295.00	282.00	267.00	142.50	150.00	130.00	56.800	56.800	56.800

Podemos ver en los histogramas de las tres variables (figura 2), que hay más simetría en *diabp* durante el tiempo y que las demás variables tienen asimetría hacia la derecha.

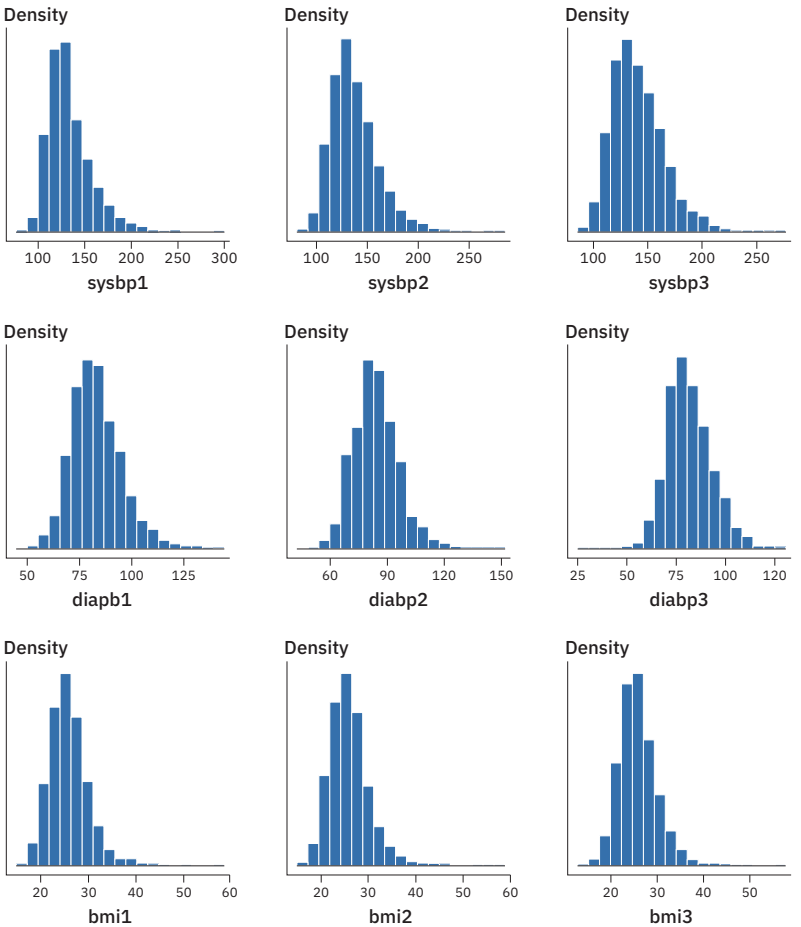


Figura 2. Histograma de las tres variables para cada período de tiempo

Construcción de un intervalo de confianza para una proporción

Como primer paso de inferencia, introducimos la construcción del intervalo de confianza para una proporción. Para ello, utilizaremos las variables dicotómicas *hyperten* y *angina* como ejemplo.

La fórmula para calcular el intervalo de confianza para una proporción es la siguiente (siguiendo la aproximación a la normal):

$$\left(p - z_{a/2}\sqrt{\frac{p(1-p)}{n}} \, , p + z_{(1-a/2)}\sqrt{\frac{p(1-p)}{n}} \right)$$

Donde `p` es la proporción muestral (el estimador puntual de la proporción poblacional), `n` es el tamaño de la muestra y $z_{\alpha/2}$, $z_{(1 - \alpha/2)}$, son los percentiles $\alpha/2$ y $(1 - \alpha/2)$ de la distribución normal tipificada `(N(0,1))` (véase figura 3), siendo α = `1 – grado de confianza` (en tanto por uno).

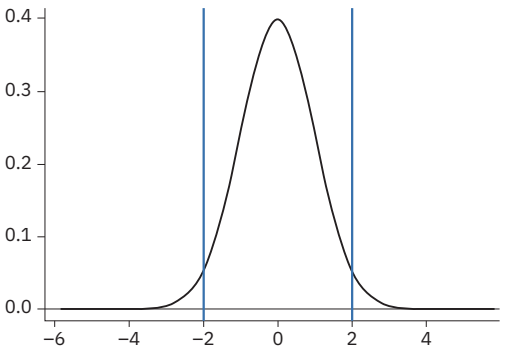


Figura 3. Distribución normal tipificada, en la que se señalan los percentiles 0,025 (–1,959) y 0,975 (1,959), que corresponden a un grado de confianza del 95 %

Así, si trabajamos al 95 % de confianza (lo habitual), α = `1 – 0,95 = 0,05` y α = `0,05 / 2 = 0,025` que se corresponde con el valor –1,959 para $z_{\alpha/2}$ y al 1,959 para el percentil $z_{(1 - \alpha/2)}$.

Los pasos que seguir en Jamovi para la obtención de los resultados será primero elegir «Análisis» (figura 4), donde seleccionaremos «Prueba para la proporción en una muestra» y «2 Resultados. Prueba binomial», seleccionar las variables (figura 4) y, finalmente, decidir qué resultados queremos obtener (figura 5).

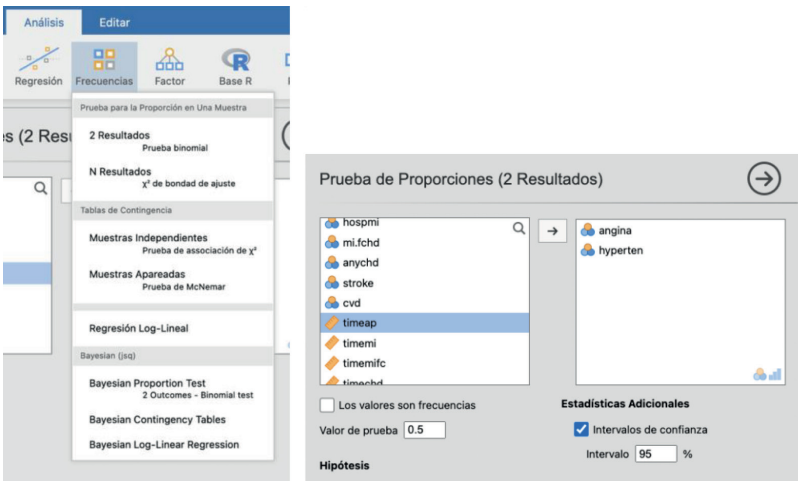


Figura 4. Obtención de proporciones (I)

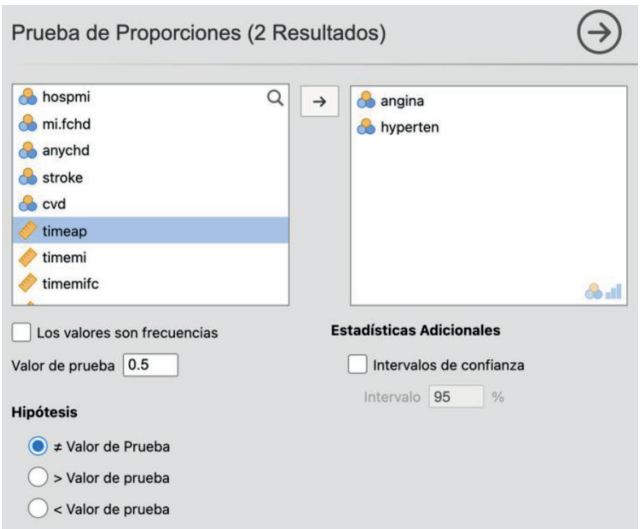


Figura 5. Obtención de proporciones (II)

Los resultados obtenidos al utilizar las dos variables binomiales aparecen en la siguiente tabla.

Tabla 4. Resultados de la prueba de proporciones de una variable, para el caso de angina e hyperten

Prueba binomial

						Intervalo de confianza al 95 %	
	Nivel	**Frecuencia**	**Total**	**Proporción**	**p**	**Inferior**	**Superior**
angina	No	3709	4434	0.83649	< .00001	0.82527	0.84726
	Yes	725	4434	0.16351	< .00001	0.15274	0.17473
hyperten	No	1182	4434	0.26658	< .00001	0.25361	0.27986
	Yes	3252	4434	0.73342	< .00001	0.72014	0.74639

Nota. H_{α} es proporción ≠ 0.5

Obtenemos el número de casos para cada posibilidad, total de la muestra, proporción de cada uno de los dos casos y el p-valor para el contraste de esta proporción, concepto que veremos más adelante. Además, podemos ver la estimación puntual y por intervalo. Por ejemplo, el resultado nos dice que la proporción de hipertensos es de 73,34 %, estimador puntual de la proporción de hipertensos en la población. Por otra parte, con un 95,00 % de confianza, la proporción de hipertensos en la población está entre el 72,01 % y el 74,64 %. La proporción de sujetos que han padecido angina de pecho es del 16,35 % y, con confianza al 95,00 %, la proporción está entre el 15,27 % y el 17,47 %. Observad cómo la amplitud de los intervalos es bastante pequeña, por lo que los estimadores son bastante precisos. Podemos cambiar el nivel de confianza, por ejemplo, al 99,00 %, como se observa en la figura 5.

Construcción de un intervalo de confianza para la media

A continuación, veremos cómo construir un intervalo de confianza para una media. Se trata de un caso que implica inferencia de una variable del tipo continua (o de los tipos escala o numérico).

La fórmula para construir un intervalo de confianza para una media es la siguiente:

$$\left(\bar{x} - t_{n-1,a/2}\frac{s}{\sqrt{n}}, x + t_{n-1,(1-a/2)a/2)}\frac{s}{\sqrt{n}}\right)$$

Donde $\bar{x}$ es la media muestral (el estimador puntual de la media poblacional), s es la desviación típica, n es el tamaño de la muestra y

$t_{n-1,\alpha/2}$, $t_{n-1,(1-\alpha/2)}$ son los percentiles $\alpha/2$ y $(1 - \alpha/2)$ de la distribución t de Student con n-1 grados de libertad, respectivamente (véase figura 6), siendo α = 1 – grado de confianza (en tanto por uno).

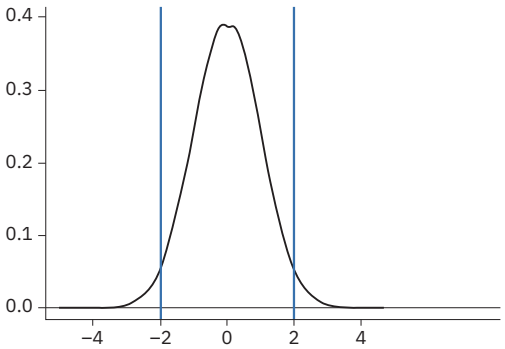


Figura 6. Distribución t de Student con 4433 grados de libertad (4433 – 1), en los que se señalan los percentiles 0,025 (–1,960) y 0,975 (1,960), que corresponden con un grado de confianza del 95 %

En este caso, si utilizamos la muestra de la base de datos de Framingham, con n = 4434 sujetos, trabajamos al 95 % de confianza (lo habitual), α = 1 – 0,95 = 0,05 y α = 0,05 / 2 = 0,025, y, por simetria (1 – α / 2) = 0,975, se corresponde con los valores –1,960 para $t_{4434-1,0,025}$ y, 1,96 a $t_{4434-1,0,975}$.

La distribución t de Student es más apuntada que la distribución normal y cuantos más grados de libertad, más apuntada. Así, en la figura 7, la curva negra corresponde a una distribución con 100 grados de libertad y en azul a una distribución con 4 grados de libertad.

Antes de construir el intervalo de confianza es conveniente comprobar que la variable no tenga valores imposibles. Para ello, en el análisis descriptivo de la variable pediremos el máximo y el mínimo y comprobaremos que todos los valores sean lógicos y, de lo contrario, definiremos como perdidos los valores imposibles.

Así, por ejemplo, seleccionamos en Jamovi «Análisis», «Exploración» y «Descriptivos» y comprobamos los valores imposibles de la edad en el momento de inclusión en la cohorte (*age_basal*) (figura 8).

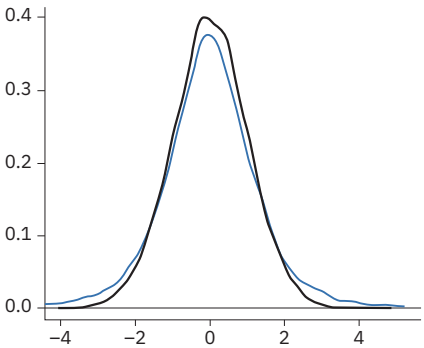


Figura 7. Distribuciones t de Student con 100 grados de libertad (en negro) y con 4 grados de libertad (en azul)

En la figura 8 vemos que la edad máxima es de 165 años, valor imposible. Lo que hacemos es eliminar el valor 165 (es decir, hacerlo *missing*) y repetir el procedimiento a ver si hay otros valores imposibles (figura 9). Como vemos, no parece haber valores imposibles para *age_basal*.

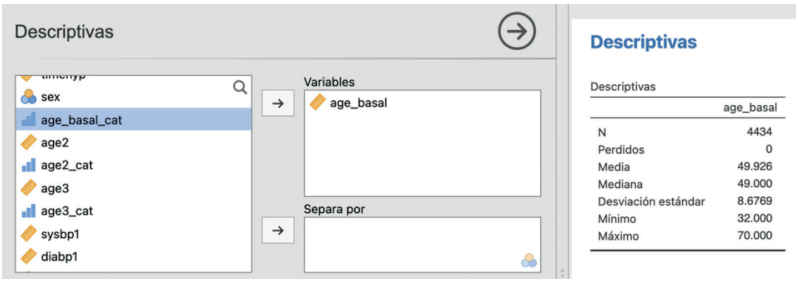


Figura 8. Algunos descriptivos de age_basal

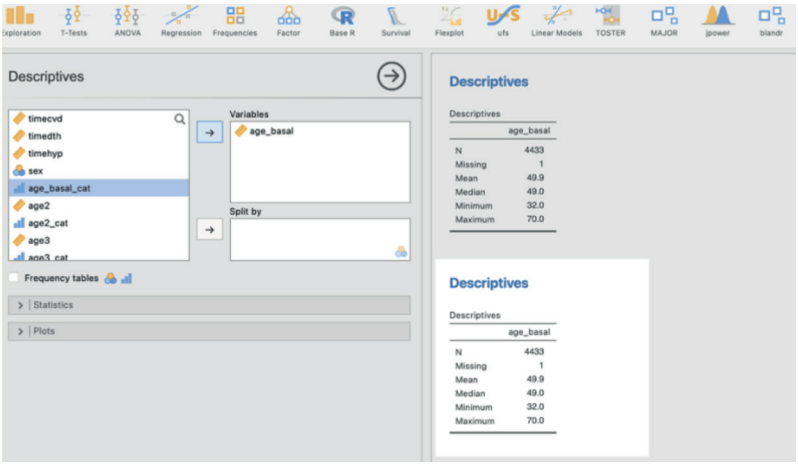


Figura 9. Algunos descriptivos de age_basal, después de hacer faltante (missing) el valor 165 años

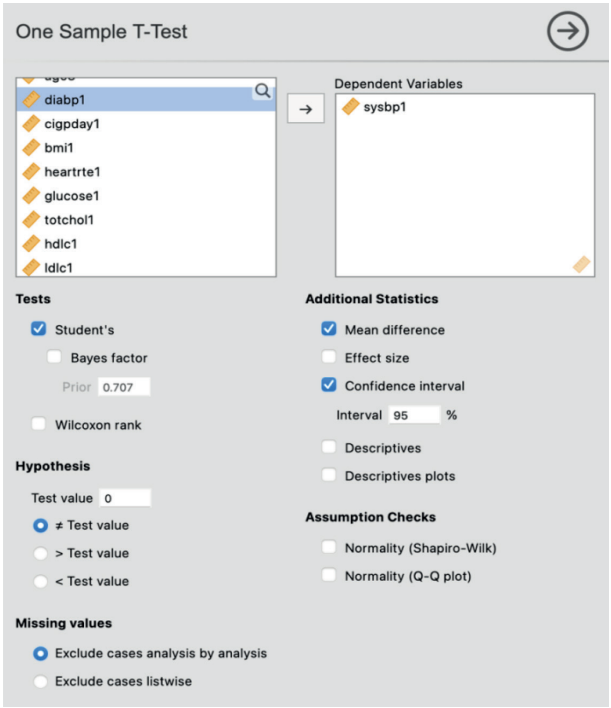


Figura 10. Algunos descriptivos de age_basal, después de hacer faltante (missing) el valor 165 años

Tras comprobar que no existen valores imposibles para las variables *sysbp*, *diabp* y *imc* en sus tres observaciones, construiremos, utilizando Jamovi, intervalos de confianza para sus medias.

Seleccionaremos «Análisis», «Prueba T en una muestra» (figura 10).

Tabla 5. Resultados de la prueba de inferencia para la media de tres variables en tres momentos distintos

Prueba T en una muestra

						Intervalo de confianza al 95%	
		Estadístico	**gl**	**p**	**Diferencia de medias**	**Inferior**	**Superior**
sysbp1	T de Student	394.71	4433.0	<.00001	132.908	132.248	133.568
diabp1	T de Student	458.89	4433.0	<.00001	83.084	82.729	83.439
bmi1	T de Student	418.68	4414.0	<.00001	25.846	25.725	25.967
sysbp2	T de Student	380.82	3929.0	<.00001	136.947	136.242	137.653
diabp2	T de Student	460.93	3929.0	<.00001	84.020	83.662	84.377
bmi2	T de Student	393.02	3913.0	<.00001	25.898	25.769	26.027
sysbp3	T de Student	349.34	3262.0	<.00001	140.216	139.429	141.003
diabp3	T de Student	414.52	3262.0	<.00001	81.793	81.406	82.180
bmi3	T de Student	361.54	3245.0	<.00001	25.895	25.754	26.035

Nota. $H_{\alpha}\ \mu \neq 0$

En los resultados de la tabla 5 podemos ver que, con un 95 % de confianza, la media de *sysbp1* está entre 132,24 y 133,56. Lo mismo podemos decir para cada una de las variables que aparecen en la tabla. Tenemos los valores del estadístico, los grados de libertad, la media y su intervalo de confianza correspondiente.

Contrastes de hipótesis

Los intervalos de confianza se pueden utilizar para contrastar hipótesis. Por ejemplo, utilizando los resultados de la tabla 4 podríamos contrastar si la prevalencia poblacional de angina de pecho es igual al 20 %. Utilizando la base de datos Framingham como muestra, estimamos que la proporción (muestral) de la angina de pecho, es decir, la prevalencia de la angina en la muestra es igual a 16,351 %, con un intervalo de confianza al 95 % igual a (15,274 %, 17,473 %). El 20 % no está contenido en el intervalo. Es decir, con un 95 % de probabilidad, rechazaremos que la

prevalencia de angina de pecho para esa población sea igual al 20 %. Sin embargo, no es habitual utilizar los intervalos de confianza para realizar contrastes de hipótesis. Lo más habitual es utilizar contrastes de hipótesis por sí solos.

Una prueba de **contraste de hipótesis** (también denominado contraste de hipótesis o prueba de significación) es un procedimiento para juzgar si una propiedad que se supone en una población estadística es compatible con lo observado en una muestra de dicha población. Otra definición, en la que utilizaremos varios conceptos básicos de contrastación de hipótesis, es que un contraste calcula la probabilidad de que los resultados obtenidos en una investigación puedan ser debidos al azar en el supuesto de que la hipótesis nula sea cierta.

Teóricamente, el proceso de contraste se divide en varias etapas:

1. Planteamiento de la hipótesis nula (H_0) y de la hipótesis alternativa (H_A o H_1).

Una hipótesis es una determinada conjetura que puede tomar un parámetro poblacional.

Siguiendo con nuestro ejemplo, π = 20 %. Observad que utilizamos la nomenclatura poblacional, ya que queremos contrastar si la prevalencia poblacional de angina de pecho es igual al 20 %. Concretamente,

$$H_0: \pi = 0$$

$$H_A: \pi \neq 0$$

Si en la hipótesis nula los símbolos son igual (=), como en nuestro caso, el contraste se denomina **bilateral** o **a dos colas**. Si tuviese un símbolo diferente al igual, se denominaría **unilateral** o **a una cola**. Que sería este caso:

$$H_0: \pi > 0$$

$$H_A: \pi \leq 0$$

Habitualmente se trabaja con contrastes bilaterales o a dos colas, como haremos nosotros.

Desde el punto de vista científico, nos interesa rechazar la hipótesis nula.

2. Decisión de un estadístico que resuma adecuadamente la información muestral del parámetro de interés.

Ese estadístico, denominado **estadístico de contraste**, es el estimador ELIO del parámetro. En nuestro caso, el estimador ELIO de la proporción poblacional (π) es la proporción muestral (p).

3. División del espacio muestral en dos regiones no solapadas, la región crítica y la región de aceptación.

Ese apartado se divide en tres subetapas:

3a. Representación gráfica de la distribución muestral del estadístico.

En nuestro caso, la proporción muestral se distribuye como una distribución chi-cuadrado que, como vemos en la figura 11, es asimétrica.

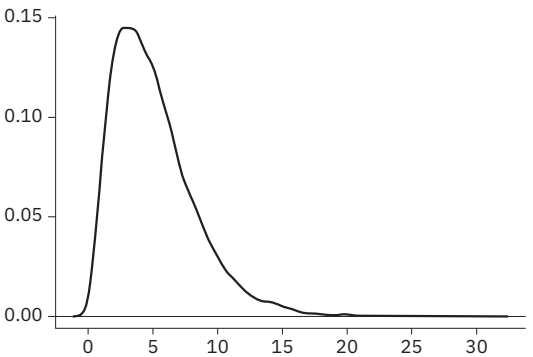


Figura 11. Distribución chi-cuadrado

3b. Elección del riesgo del contraste.

El **riesgo** del contraste, también denominado **nivel de significación**, es la probabilidad de rechazar la hipótesis nula cuando esta es cierta. Estos conceptos se pueden explicar mejor en la siguiente tabla (tabla 6).

Tabla 6. Cuadro de decisión

		Decisión	
		No rechazar	**Rechazar**
Hipótesis nula (H_0)	**Cierta**	Correcto	Error tipo I
	Falsa	Error tipo II	Correcto

El **error tipo I** (también denominado **falsos positivos**) se comete al rechazar la hipótesis nula siendo cierta. La probabilidad de cometer error tipo I se denomina **riesgo, nivel de significación, nivel alfa** o **α**. Además, el grado de confianza no es más que 1 – α .

El **error tipo II** (también denominado **falsos negativos**) se comete al no rechazar la hipótesis nula siendo falsa. La probabilidad de cometer error tipo II se denomina **beta**, ***β***. La **potencia del contraste** es igual a 1 – β. Es decir, la potencia del contraste es la probabilidad de rechazar la hipótesis nula siendo falsa.

Se suele trabajar con un riesgo del 5 % (grado de confianza del 95 %), que se corresponde con un β = 20 % y una potencia del 80 %.

3c. Dibujar la región crítica y la región de aceptación.

Sobre la representación de la distribución muestral del estadístico (en nuestro caso chi-cuadrado) se dibujan los percentiles α = 0,05 / 2 = 0,025 y (1 – α / 2) = 0,975. El riesgo se divide entre 2, porque el contraste es a dos colas. En nuestro caso, véase la figura 12.

Nótese que, al ser una distribución asimétrica, las regiones de rechazo no son de igual tamaño. Esto no pasaría cuando la distribución fuese simétrica (como una t de Student).

4. Obtención de una muestra para medir el parámetro de interés.

En nuestro ejemplo, la base de datos de Framingham.

5. Cálculo del estadístico de contraste en la muestra escogida.

Cuando se contrastan proporciones (una o más de una), el contraste del estadístico es el contraste chi-cuadrado.

En Jamovi, seleccionamos «Análisis», «Frecuencias», «Prueba para la proporción en una muestra» y «2 Resultados. Prueba binomial». En «Valor del contraste» poner 0.2 (el 20 % en tanto por uno) y señalar «Intervalo de confianza al 95 %» (figura 13).

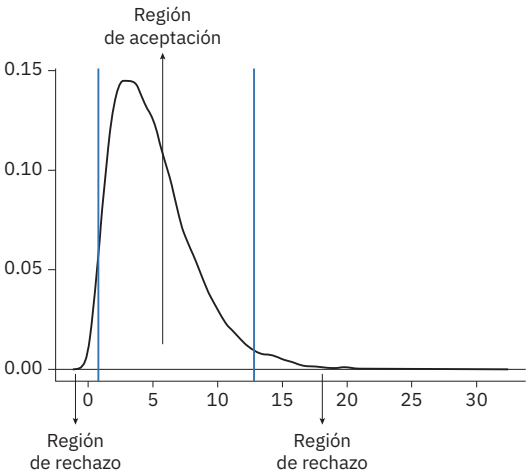


Figura 12. Distribución chi-cuadrado en la que hemos dibujado las regiones de aceptación y de rechazo

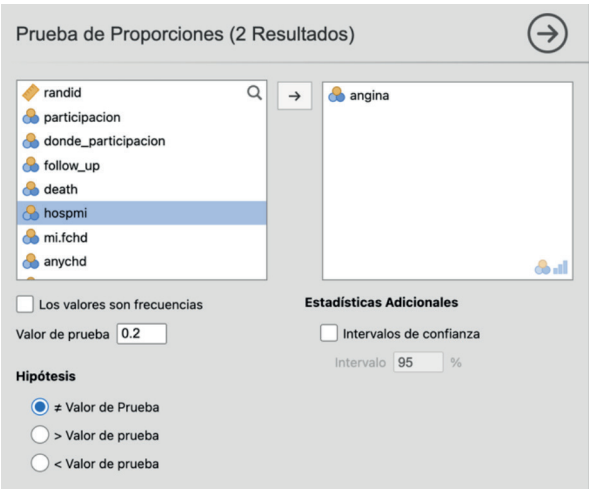


Figura 13. Contraste de hipótesis de una proporción en Jamovi

El resultado del contraste está en la figura 14.

Prueba de proporciones (2 resultados)

Prueba binomial

	Nivel	Frecuencia	Total	Proporcion	p
angina	No	3709	4434	0.83649	<.00001
	Yes	725	4434	0.16351	<.00001

Nota. H_a es proporción ≠ 0.2

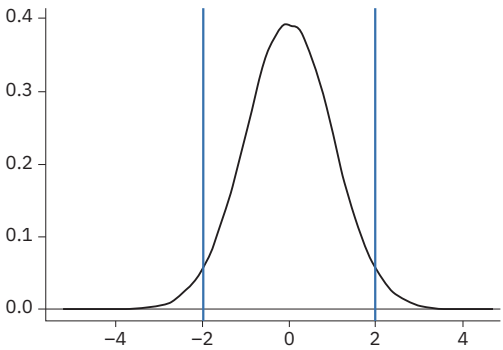


Figura 14. Resultado del contraste de hipótesis de una proporción en Jamovi

6. Resolución del contraste.

La resolución se podría hacer gráficamente (figura 15). Si el valor muestral del estadístico de contraste se sitúa en algunas de las regiones de rechazo (como en nuestro caso), se rechazará la hipótesis nula.

Por el contrario, si el valor muestral del estadístico de contraste se sitúa en la región de aceptación, no se rechaza la hipótesis nula.

Numéricamente, en el pasado se utilizaban las tablas estadísticas, como la que mostramos en la tabla 7.

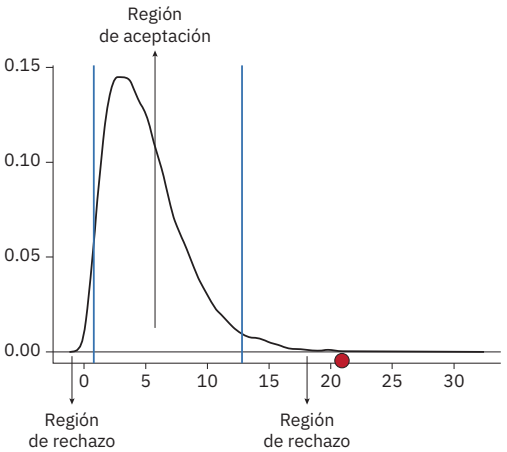


Figura 15. Resolución del contraste. Rechazo de la hipótesis nula

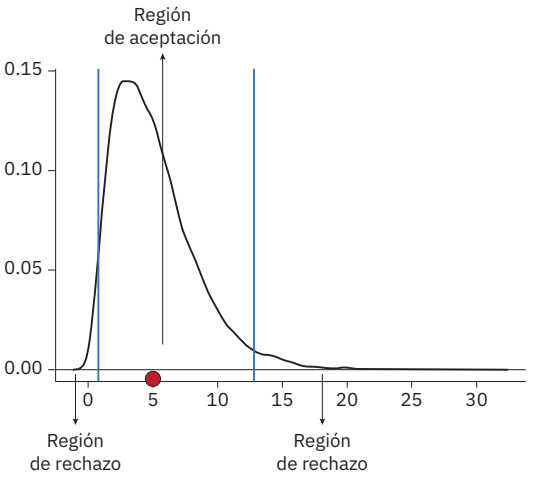


Figura 16. Resolución del contraste. No rechazo de la hipótesis nula

Pero es mucho más cómodo utilizar el **p-valor** (o valor p). El p-valor podría ser definido como la probabilidad de observar los resultados del estudio, u otros más alejados de la hipótesis nula, si la hipótesis fuese cierta.

Tabla 7. Parte de los valores críticos de la distribución chi-cuadrado según los grados de libertad (GL) y el riesgo

GL	α = 0,10	α = 0,05	α = 0,01
1	2,7	3,8	6,6
2	4,6	6,0	9,2
3	6,3	7,8	11,3
4	7,8	9,5	13,3
5	9,2	11,1	15,1
6	10,6	12,6	16,8
7	12,0	14,1	18,5
8	13,4	15,5	20,1
9	14,7	16,9	21,7
10	16,0	18,3	23,2
11	17,3	19,7	24,7
12	18,5	21,0	26,2
13	19,8	22,4	27,7
14	21,1	23,7	29,1
15	22,3	25,0	30,6

Cada contraste en particular tiene asociado su p-valor. En nuestro caso, observando la figura 14, el p-valor del contraste es $p < 0{,}001$, es decir, menor del 1 por 1000. En términos prácticos, si el p-valor es menor que el riesgo, se rechaza la hipótesis nula y, en caso contrario, no se rechaza la hipótesis nula. Así pues, como $0{,}001 < 0{,}05$ (asumimos un riesgo del 5 %), rechazamos la hipótesis nula. Es decir, al 95 % de confianza, no podemos decir que la prevalencia de angina de pecho en la población sea el 20 %.

Cabe señalar que el p-valor solo sirve para rechazar o no la hipótesis nula. No proporciona ninguna información sobre la significación estadística. Por ejemplo, tenemos dos contrastes de una proporción, uno con un p-valor igual a 0,03 y otro con un p-valor igual a 0,01. En los dos casos rechazamos la hipótesis nula, pero en ningún caso podemos decir que el segundo contraste es más significativo que el primero.

Contraste de una media

En este caso, también podemos hacer un análisis sobre un valor concreto de la media, ya que aquí hemos comparado en todos ellos con el valor 0 como valor de la prueba para la media. Por ejemplo, podemos ver si el

caso del índice de masa corporal observado en el momento 3 (variable *bmi3*) se diferencia con un valor concreto, que podría ser 24,9 a partir del cual se presupone sobrepeso.

Volvamos a plantear las etapas de contrastación. En la práctica, no suelen seguirse todas.

1. Planteamiento de la hipótesis nula (H_0) y de la hipótesis alternativa (H_A o H_1).

$$H_0: \mu = 24,9$$

$$H_A: \mu \neq 24,9$$

Se trata, como antes, de un contraste bilateral o a dos colas.

2. Decisión de un estadístico que resuma adecuadamente la información muestral del parámetro de interés.

El estimador ELIO de la media poblacional (μ) es la media muestral ($\bar{x}$).

3. División del espacio muestral en dos regiones no solapadas, la región crítica y la región de aceptación.

3a. Representación gráfica de la distribución muestral del estadístico.

En nuestro caso, la media muestral se distribuye como una distribución t de Student.

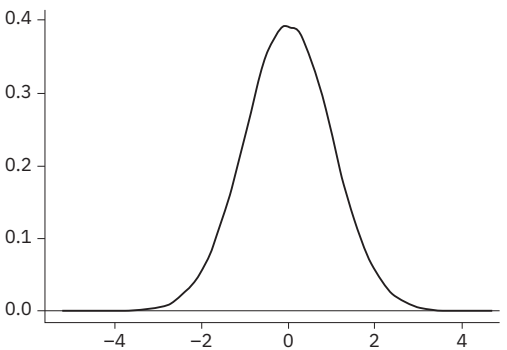


Figura 17. Distribución t de Student

3b. Elección del riesgo del contraste.

Trabajaremos con un riesgo del 5 %.

3c. Dibujar la región crítica y la región de aceptación.

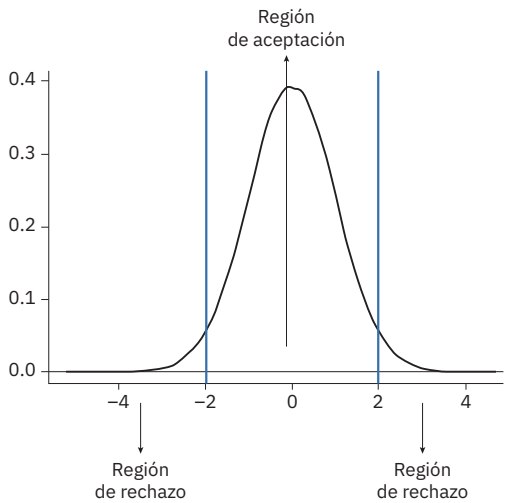


Figura 18. Distribución t de Student en la que hemos dibujado las regiones de aceptación y de rechazo

Nótese que, al ser una distribución simétrica, las regiones de rechazo son de igual tamaño.

4. Obtención de una muestra para medir el parámetro de interés.

En nuestro ejemplo, la base de datos de Framingham.

5. Cálculo del estadístico de contraste en la muestra escogida.

Cuando se contrastan proporciones (una o más), el contraste del estadístico es el contraste t de Student.

En Jamovi, seleccionamos «Análisis», «Contrastes t» y «Prueba t en una muestra». En «Valor del contraste» poner 24.9 y señalar «Intervalo de confianza al 95 %» (figura 19).

Figura 19. Contraste de hipótesis de una media en Jamovi

El resultado del contraste está en la tabla 8.

Tabla 8. Resultado del contraste de hipótesis de una media en Jamovi

Prueba T en una muestra

							Intervalo de confianza al 95 %	
		Estadístico	gl	p		Tamaño del efecto	Inferior	Superior
bmi3	T de Student	13.889	3245.0	<.00001	d de Cohen	0.24378	0.20885	0.27867

Nota. H_α $\mu \neq 24.9$

6. Resolución del contraste.

La resolución se podría hacer gráficamente (figura 20). En nuestro caso, como el valor muestral del estadístico de contraste (igual a 13,9, véase *statistic* en la figura 20) se sitúa en algunas de las regiones de rechazo (como en nuestro caso), se rechazará la hipótesis nula.

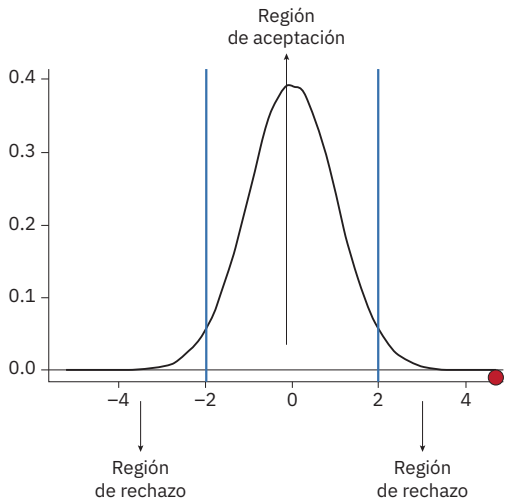


Figura 20. Resolución del contraste. Rechazo de la hipótesis nula

Pero es mucho más cómodo utilizar el **p-valor** (o valor p). En nuestro caso, observando la figura 20, el p-valor del contraste es p < 0,001, menor que el riesgo (0,05). Es decir, al 95 % de confianza no podemos decir que el IMC en el momento 3 sea 24,9.

Referencias

The Jamovi project (2022). *Jamovi*. (Versión 2.3) [Computer Software]. Recuperado de:
https://www.JAMOVI.org

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

Linlin Yan (2020). *Venn Diagram by ggplot2, with really easy-to-use API*. [R package]. Recuperado de:
https://github.com/yanlinlin82/ggvenn

Serdar Balci (2022). *ClinicoPath Jamovi Module doi:10.5281/zenodo.3997188*. [R package]. Recuperado de:
https://github.com/sbalci/ClinicoPathJAMOVIModule
https://www.serdarbalci.com/ClinicoPathJamoviModule/

Bjoern Koneswarakantha (2019). *easyalluvial: Generate Alluvial Plots with a Single Line of Code.*. [R package]. Recuperado de: https://CRAN.R-project.org/package=easyalluvial

Nick Barrowman (2020). *vtree: Display Information About Nested Subsets of a Data Frame*. [R package]. Recuperado de: https://CRAN.R-project.org/package=vtree

JASP Team (2018). *JASP*. [Computer software]. Recuperado de: https://jasp-stats.org

Morey, R. D. y Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs*. [R package]. Recuperado de: https://cran.r-project.org/package=BayesFactor

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., y Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237.

Edwards, A. W. F. (2001). *Occam's Bonus*. Cambridge University Press.

Glover, S. (2018). *Likelihood Ratios; A Tutorial*. https://osf.io/preprints/metaarxiv/g3j2k/

Cahusac, P.M.B. (2020). *Evidence-Based Statistics*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119549833

Capítulo 6
Introducción a la inferencia de relación entre dos variables

Cuando se consideran hipótesis explicativas, es necesario trabajar con más de una variable. Una de ellas, a la que llamaremos dependiente, presenta una distribución concreta de valores que no es aleatoria, sino que está determinada por la distribución de otra (u otras) variables. Según un modelo de análisis determinado, fundamentado en un marco de referencia teórico, estas otras variables «explicativas», a las que llamaremos independientes, tienen una relación con la variable dependiente que puede ser de simple covariación o de asociación e influencia causal.

Esta relación explicativa también puede ser contrastada empíricamente con datos cuantitativos, hasta el punto de afirmar o desmentir hipótesis. Esta estrategia corresponde a hipótesis explicativas respecto a la población.

En la siguiente tabla, podemos ver los distintos análisis que podemos utilizar según la relación entre las variables.

Tabla 1. Combinación de variables y análisis correspondiente

Combinación de variables	Posible análisis
V. Cualitativa → V. Cualitativa	Tablas de contingencia. Contraste de la chi-cuadrado
V. Cualitativa → V. Cuantitativa	Comparación de medias. Análisis de la varianza
V. Cuantitativa → V. Cuantitativa	Análisis de regresión. Regresión lineal
V.Cuantitativa → V. Cualitativa	Regresión logística

De la misma forma que hemos trabajado en el capítulo anterior, vamos a utilizar las variables cuantitativas continuas *sysbp*, *diabp* y *imc*, y las variables cualitativas *age_basal_cat, angina* y *death*.

Previo a sus análisis, vamos a describir las variables de forma conjunta, con el uso del paquete JJStat de Jamovi.

Primero lo haremos, como ejemplo, para dos continuas, tanto el *scatter plot* como la matriz de correlaciones. Podemos ver el resultado en los gráficos siguientes.

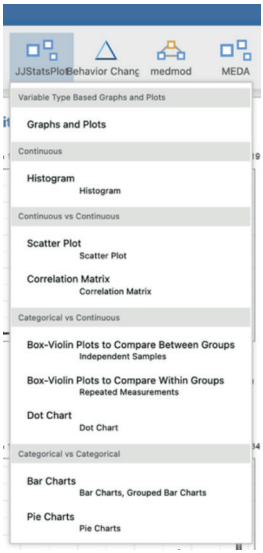


Figura 1. Posibilidades de JJStat

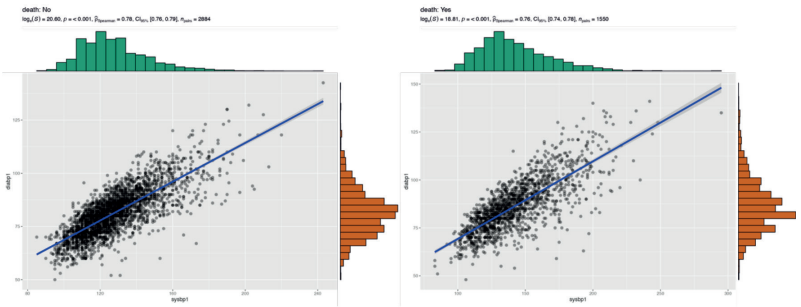


Figura 2. Scatter plot entre sysbp1 y diabp1, separadas por death

Estas figuras 2 y 3, nos muestran, previamente a realizar análisis de inferencia, cómo están relacionadas las variables.

Las siguientes figuras, 4 y 5, muestran la relación entre una variable continua y otra variable discreta (*sysbp1* y *age_basal_cat*). Podemos obtener mucha información a través de ellos, pero nos centraremos en la posibilidad de hacer, a partir de ellos, inferencia para dos o más variables.

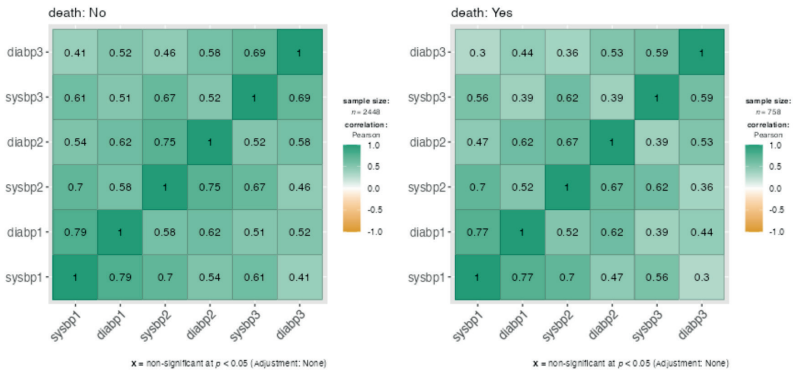
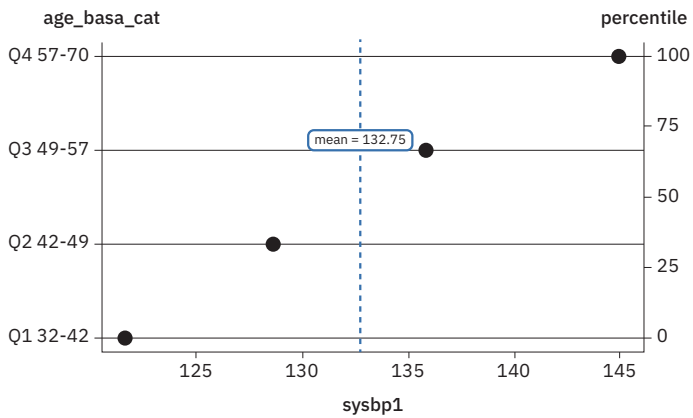


Figura 3. Matriz de correlaciones entre sysbp y diabp, separadas por death, para los tres estados

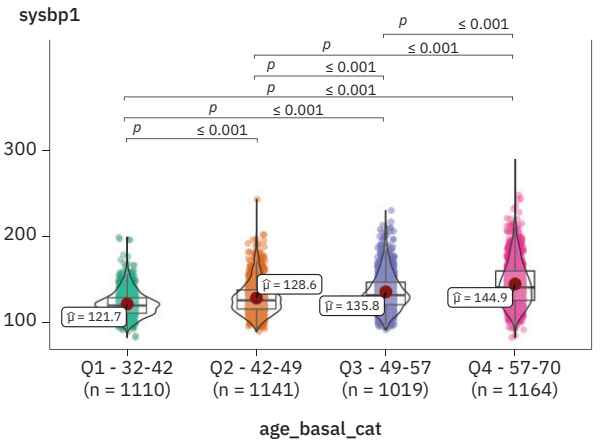
$t_{\text{Student}}(3) = 26.65$, $p = < 0.001$, $\hat{g}\text{Hedge} = 9$



In favor of null: $\log_e(BF_{01}) = -4.91$, $r^{JZS}_{Cauchy} = 0.71$

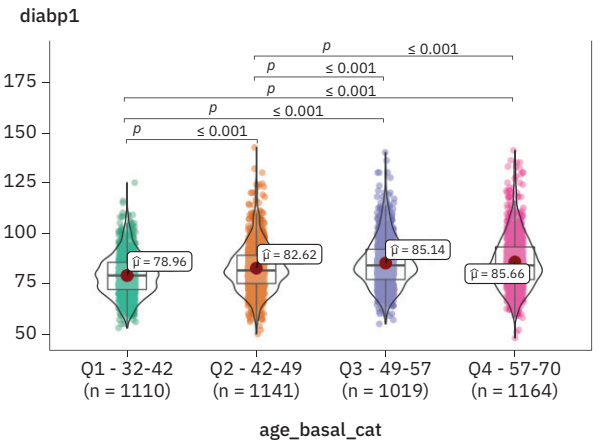
Figura 4. sysbp1 y age_basl_cat

F_{Welch} (3,2405.73) = 270.47, p = < 0.001, $\widehat{\omega^2_p}$= 0.15, $CI_{95\%}$ [0.14, 1.00], n_{obs}



In favor of null: $\log_e(BF_{01})$ = −353.91, r^{JZS}_{Cauchy} = 0.71
Pairwise comparisons: **Games-Howell test**; Adjustment (p-value): **Holm**

F_{Welch} (3,2441.44) = 84.37, p = < 0.001, $\widehat{\omega^2_p}$= 0.05, $CI_{95\%}$ [0.14, 1.00], n_{obs} =



In favor of null: $\log_e(BF_{01})$ = −353.91, r^{JZS}_{Cauchy} = 0.71
Pairwise comparisons: **Games-Howell test**; Adjustment (p-value): **Holm**

Figura 5. sysbp1 y age_basl_cat. (gráfico de violín)

Finalmente, como ejemplo, si tenemos dos variables continuas, podemos ver su relación previa en la figura 6.

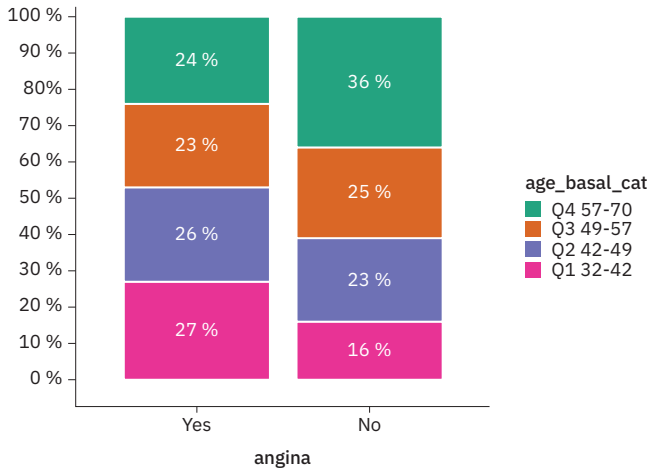


Figura 6. angina y age_basl_cat

Comparación de proporciones. Tablas de contingencia

Después de conocer cómo se pueden relacionar las variables, empezaremos con las tablas de contingencia, donde relacionaremos dos variables en formato factor. Para ello, en Jamovi, podemos observar la figura 7 de la ventana de «Frecuencias», centrándonos en «Muestras independientes» para obtener la tabla de contingencia.

Seleccionaremos dos variables, en este caso *angina* y *death*, que son dos a dos casos, lo que nos permitirá obtener los odds ratio y riesgo relativo, desarrollado en el siguiente capítulo (figura 8).

Primeramente (figura 8), podemos ver la descriptiva en formato de tabla de los datos de las dos variables, en filas y columnas, pero incluyendo, en este caso, los porcentajes por filas. También los valores de las medidas comparativas como son los odds ratio y los riesgos relativos con sus respectivos valores e intervalos.

Figura 7. Procedimiento «Tablas de contingencia»

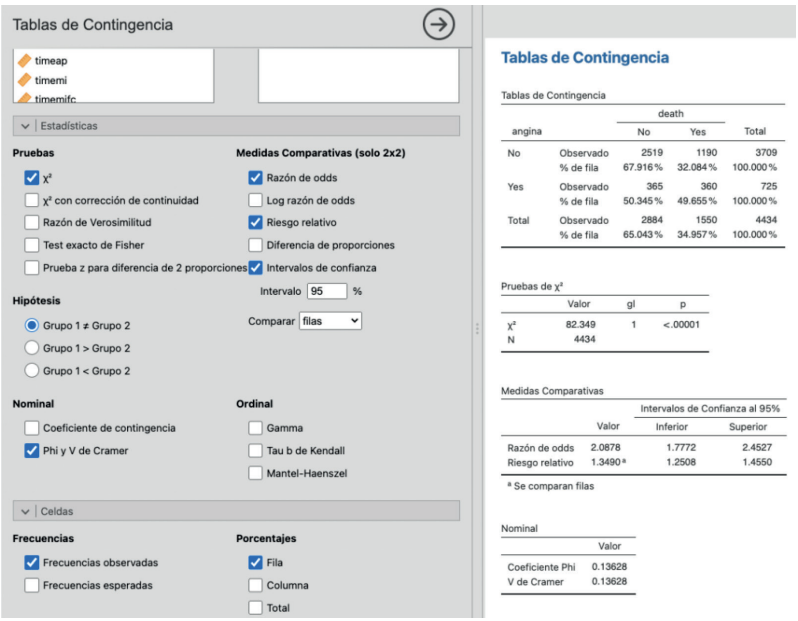


Figura 8. Resultados de la tabla de contingencia

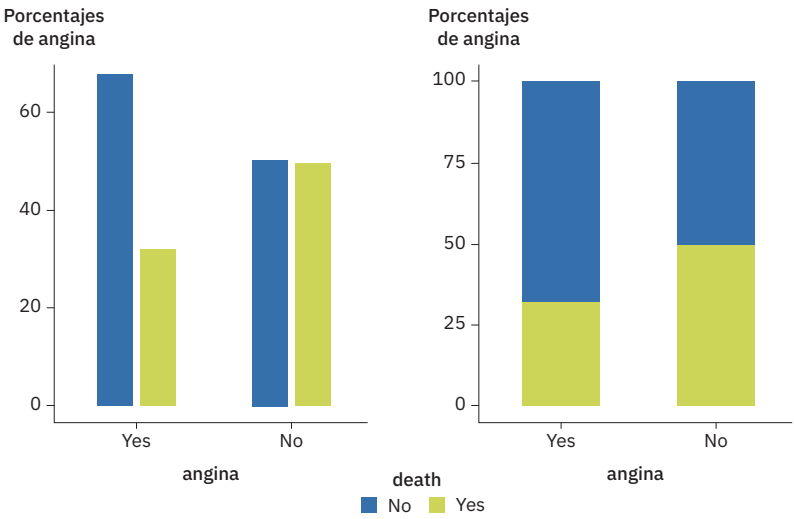


Figura 9. Gráficos ejemplo de la tabla de contingencia

En la figura 9, podemos ver la representación gráfica que nos permite esta ventana, separando por porcentajes por filas. En la parte izquierda, uno al lado de otro y, en la derecha, alineados, que son las opciones que nos permite Jamovi.

El objetivo de la estadística inferencial, expresada aquí a partir del contraste de la khi cuadrado ($\chi 2$), es evaluar hasta qué punto estas diferencias son extrapolables al conjunto de la población.

La conclusión puede ser:

- Las diferencias **son significativas**. Llegaremos a esta conclusión cuando las diferencias difícilmente sean debidas al azar. Si unas diferencias observadas en la muestra no son debidas al azar, es que se dan en la población general y, por tanto, es muy probable que exista relación entre las dos variables consideradas.
- Las diferencias **no son significativas**. Las diferencias pueden deberse al azar, dado que los datos han sido obtenidos de una muestra, de modo que las conclusiones derivadas del análisis no pueden extrapolarse al conjunto de la población.

El valor que nos informa de esta influencia del azar es la significación del contraste de la khi cuadrado, también llamado p-valor. Esta significación nos informa sobre la probabilidad de que las diferencias observadas sean debidas al azar:

- Cuando la significación de $\chi 2$ es grande no podemos descartar que las diferencias sean debidas al azar y, por tanto, no extrapolables a la población.
- Mientras que si la significación de $\chi 2$ es pequeña es que las diferencias difícilmente serán debidas al azar y, por tanto, son extrapolables a la población.

Si el p-valor es igual o menor (p-valor ≤ 0,05) diremos que podemos aceptar que existen diferencias significativas entre las categorías (y, por tanto, que hay relación entre las variables) con un 95 % de confianza. De hecho, cuanto menor sea la significación más improbable es que el azar nos haya jugado una mala pasada y más probable es que las diferencias se den verdaderamente en el conjunto de la población. Esto también nos servirá para todos los apartados siguientes.

En este caso, podemos observar un p-valor muy pequeño, menor de 0,001, por lo que las diferencias observadas entre las dos variables se pueden asegurar con un riesgo de error muy pequeño. Es decir, que la variable *angina* esté relacionada con *death* para la población en general. El resultado de esta etapa es que hay significación y, por tanto, se puede continuar con el análisis. En caso de que no exista significación estadística se pueden sacar conclusiones de la tabla, pero solo serán aplicables a las personas entrevistadas en la muestra (n). En este caso, no se podrá inferir ninguna conclusión a la población (N) a partir de los datos disponibles.

Una vez establecido que existe relación significativa entre ambas variables se suele calcular una medida para evaluar la intensidad de esta relación. Es habitual hacerlo a partir del indicador de asociación bilateral V de Cramer, que varía entre 0 y 1: un valor de 0 equivaldría a decir que no existe ninguna relación (o, con valores cercanos a 0, que es muy baja) mientras que un valor cercano a 1 debería ser leído en términos de asociación fuerte entre las variables. En ciencias de la salud no es habitual encontrar valores muy elevados de este indicador, por lo que superar el valor de 0,3 ya se considera una intensidad elevada. En los resultados del caso de *angina* y *death*, podemos ver que hay un valor cercano a 0 (0,13), intensidad baja. Normalmente, mientras este valor no sea inferior a 0,1, se puede seguir con el análisis.

En casos extremos, cuando trabajamos con muestras muy pequeñas, podemos encontrarnos que incluso agrupando al máximo sigan sin cumplirse las condiciones de aplicación. Así, en tablas 2 filas x 2 columnas, cuando alguna celda esperada es inferior a 5, pero todas son superiores a 3, se aplica una corrección en el contraste de la khi cuadrado de Pearson llamada corrección de Yates, y que Jamovi identifica como la opción corrección de continuidad de $\chi 2$. Si lo incluimos en este caso casi no cambia, debido al número de datos que hay.

Comparación de medias

En esta sección, analizaremos la relación entre una variable independiente cualitativa y una variable dependiente cuantitativa (de la que se puede calcular la media) mediante los procedimientos «T-tests» y «One-way ANOVA», así como para realizar la representación gráfica pertinente e interpretar la información que se obtiene de Jamovi para dar respuesta a las hipótesis de relación entre estas dos variables de acuerdo con un modelo de análisis básico (aspecto descriptivo y de contenido del análisis), y determinar la significación de esta información para extrapolarla al conjunto de la población (aspecto inferencial del análisis).

El objetivo principal es mostrar cómo comparar las medias de una variable cuantitativa entre los distintos grupos o categorías de una variable cualitativa. Nos encontramos ante la posible relación entre dos variables, una métrica cualitativa, que hará el papel de variable independiente (en Jamovi, de tipo NOMINAL u ORDINAL), y otra cuantitativa, la variable dependiente (en Jamovi, de tipo CONTINUA, con datos ENTERA o DECIMAL).

Comparación de dos medias. T de Student

Empezaremos con el caso de muestra independientes, por ejemplo, dos variables tomadas en diferentes momentos (*dyabp* o *imc*) y para diferentes condiciones, por ejemplo, *angina* o *death*.

Figura 10. Tipos de pruebas T, centrada ahora en muestras independientes

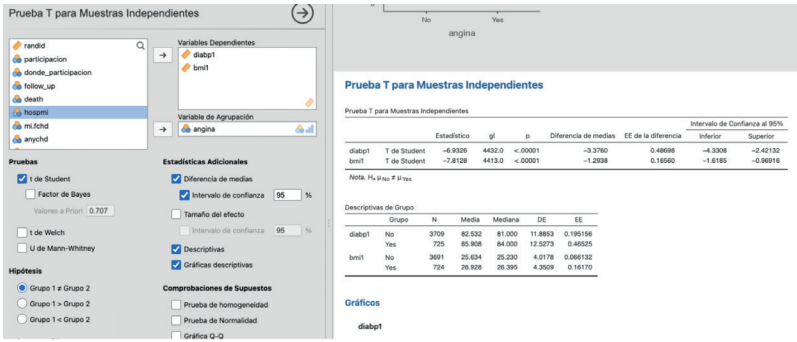


Figura 11. Resultados de la prueba T con datos independientes

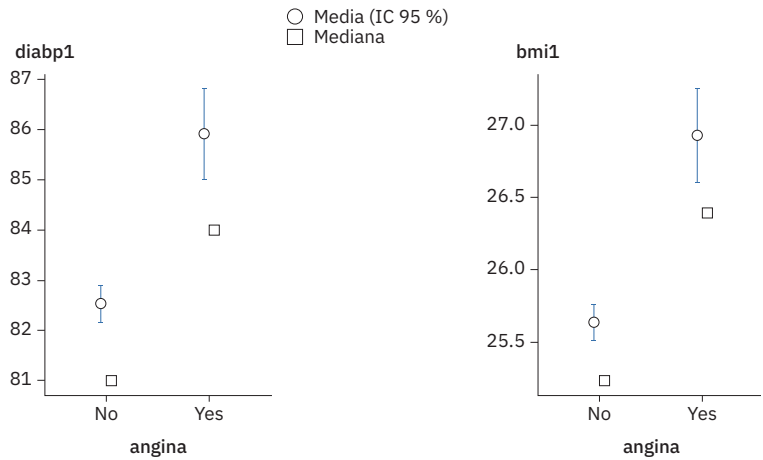


Figura 12. Resultados gráficos de la prueba T con datos independientes

La interpretación de esta visualización gráfica es:

1. En la muestra obtenida, existe evidencia de diferencias entre la media de los dos casos, puesto que las líneas verticales no coinciden.
2. Además, estas diferencias son suficientemente elevadas como para asegurar que existen diferencias entre *diabp* y *bmi* para la separación entre *angina* No y Sí en la población en general porque los niveles de confianza no coinciden.

Además, podemos completar con el contraste de Shapiro-Wilk para la normalidad de la variable. Tal y como se ha mencionado anteriormente,

para poder realizar un contraste de tipo paramétrico (que es el más recomendado) es necesario que las muestras sean «suficientemente grandes», o bien que la distribución «no difiera mucho» de una distribución normal. Cuando no se cumple la condición, es decir cuando en algún grupo poco numeroso no existe normalidad, es recomendable utilizar un contraste no paramétrico. La mayoría de los autores considera que si uno de los grupos comparados supera los 30 casos es suficiente. Así, solo evaluaremos la normalidad en los grupos que no lleguen a ese número de casos.

Para evaluar la normalidad realizaremos la prueba de normalidad de Shapiro-Wilk que ofrece Jamovi. Este contraste evalúa las siguientes hipótesis:

- H0: la distribución es normal.
- H1: la distribución no es normal.

Se interpreta de forma que si el p-valor es menor (o igual) a 0,05 consideramos evidencia de soporte a H1 (no hay normalidad), mientras que si es superior a 0,05 podemos asumir la normalidad (con un 95 % de nivel de confianza). A continuación, podemos ver los resultados obtenidos para nuestros datos en el tema de normalidad (tabla 2).

Tabla 2. Contraste de normalidad de diabp1 y bmi1

Tests of Normality

		statistic	p
diabp1	Shapiro-Wilk	0.97219	<.00001
	Kolmogorov-Smirnov	0.060164	<.00001
	Anderson-Darling	23.510	<.00001
bmi1	Shapiro-Wilk	0.95746	<.00001
	Kolmogorov-Smirnov	0.052538	<.00001
	Anderson-Darling	28.408	<.00001

Comparación de más de dos medias. ANOVA de un factor

La pregunta que nos formularemos en este caso es similar a la anterior. Lo único que cambia es que la variable explicativa cualitativa tiene más de dos categorías o valores y, por tanto, genera más de dos grupos que comparar. En este caso nos preguntamos si la variable continua, por ejemplo, *bm*i o *diapb*, se agrupa según más de un caso, *ag-basal_cat*. Elegiremos según aparece en la imagen.

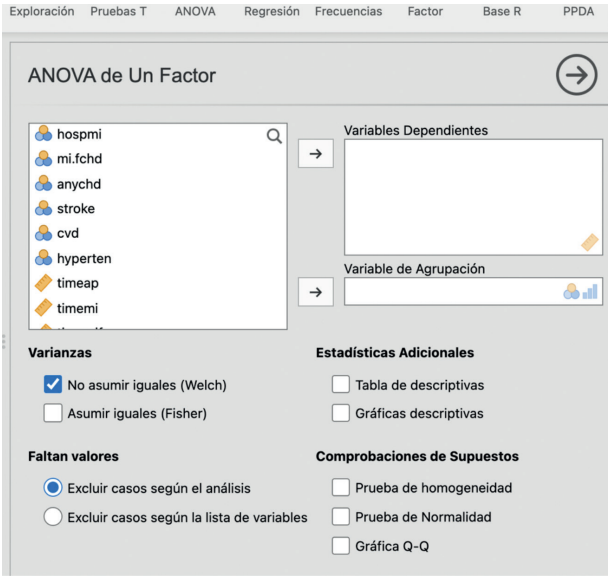


Figura 13. Procedimiento ANOVA de un factor

Los resultados los podemos ver en la figura 14.

El resultado nos lo vuelve a dar en forma de p-valor. En este caso toma un valor menor a 0,001, de modo que existe evidencia suficiente para afirmar que las varianzas no son iguales: es necesario realizar el contraste ANOVA en caso de varianzas diferentes.

En el panel de definición del procedimiento anterior se puede comprobar que existen dos contrastes posibles:

- Contraste paramétrico en caso de varianzas diferentes: *Don't assume equal (Welch's)*.
- Contraste paramétrico en caso de varianzas iguales: *Assume equal (Fisher's)*.

También podemos ver los resultados descriptivos del análisis.

Los resultados gráficos de ANOVA, los podemos ver en la figura 15, donde se muestra de forma clara, la diferencia para la mayoría de los casos.

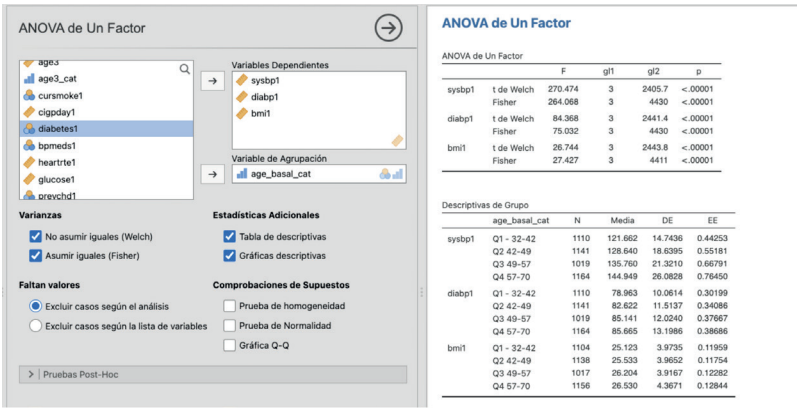


Figura 14. Resultados ANOVA de un factor

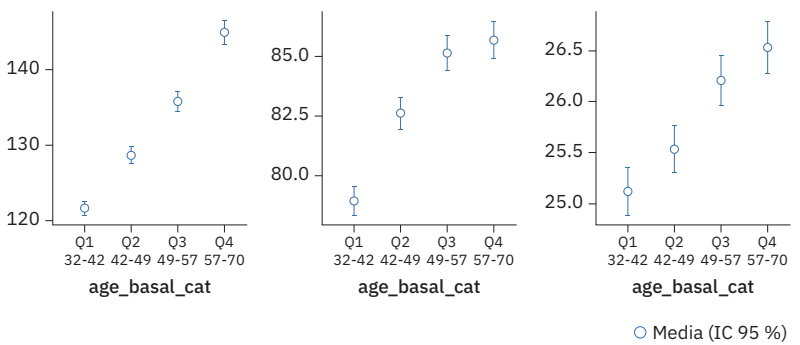


Figura 15. Resultados gráficos ANOVA de un factor

Cuantos más grupos tengamos más comparaciones 2 a 2 deberemos hacer. Para evaluarlas es necesario:

1. En el mismo panel de definición del procedimiento «Análisis de variancias simple», haz clic en «Post-Hoc Test» para activar las comparaciones múltiples a posteriori, o comparaciones 2 a 2.
2. Por defecto no se calcula ninguna comparación. Hay que activar el tipo adecuado, según las conclusiones del contraste de Levene previo: «Games-Howell» en este caso porque existe.

Tabla 3. Games-Howell de varianzas diferentes

Prueba Post-Hoc de Games-Howell – sysbp1

		Q1 – 32-42	Q2 42-49	Q3 49-57	Q4 57-70
Q1 – 32-42	Diferencia de medias	—	−6.9776	−14.0974	−23.2872
	valor p	—	< .00001	< .00001	< .00001
Q2 42-49	Diferencia de medias		—	−7.1198	−16.3095
	valor p		—	< .00001	< .00001
Q3 49-57	Diferencia de medias			—	−9.1897
	valor p			—	< .00001
Q4 57-70	Diferencia de medias				—
	valor p				—

Prueba Post-Hoc de Games-Howell – diabp1

		Q1 – 32-42	Q2 42-49	Q3 49-57	Q4 57-70
Q1 – 32-42	Diferencia de medias	—	−3.6588	−6.1778	−6.70146
	valor p	—	< .00001	< .00001	< .00001
Q2 42-49	Diferencia de medias		—	−2.5190	−3.04270
	valor p		—	< .00001	< .00001
Q3 49-57	Diferencia de medias			—	−0.52369
	valor p			—	0.76662
Q4 57-70	Diferencia de medias				—
	valor p				—

Prueba Post-Hoc de Games-Howell – bmi1

		Q1 – 32-42	Q2 42-49	Q3 49-57	Q4 57-70
Q1 – 32-42	Diferencia de medias	—	−0.41044	−1.08157	−1.40706
	valor p	—	0.06870	< .00001	< .00001
Q2 42-49	Diferencia de medias		—	−0.67112	−0.99662
	valor p		—	0.00047	< .00001
Q3 49-57	Diferencia de medias			—	−0.32549
	valor p			—	0.25871
Q4 57-70	Diferencia de medias				—
	valor p				—

Se muestran los resultados en formato de tabla, donde se sitúan las diferentes categorías de la variable en filas y columnas, de modo que el punto de cruce nos da la significación de la diferencia entre los grupos correspondientes. Así, el valor p<0,001 es la significación en la mayoría de los casos (p-valor < 0,05).

Otros módulos para relaciones entre variables

Como ya vimos con anterioridad en el capítulo de inferencia para una variable, podemos utilizar otros módulos no genéricos para estudiar la relación entre variables. En la figura 16, podemos ver el uso del módulo Jeva y Toster. Los dos nos permiten realizar algunos casos como los anteriormente desarrollados.

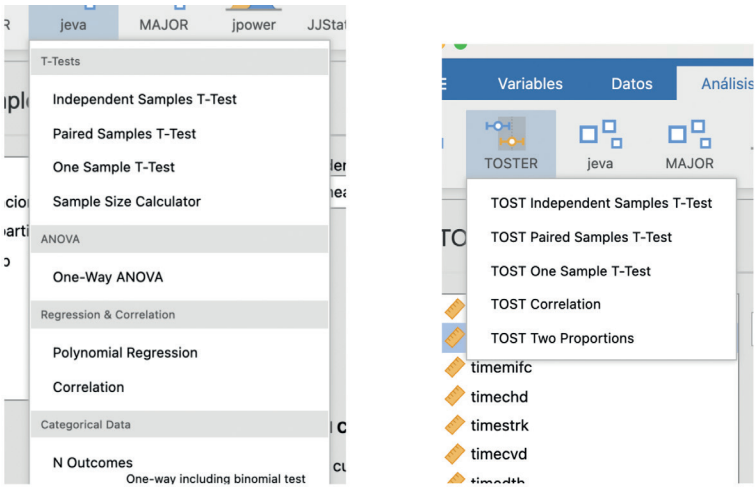


Figura 16. Jamovi de Jeva y Toster

En la siguiente imagen, podemos ver la comparación de medias de dos muestras independientes con Jeva. Los resultados son mucho más complejos y completos, aunque los resultados son los mismos.

Otras dos posibilidades son esci o análisis robusto con Walrus (figura 18).

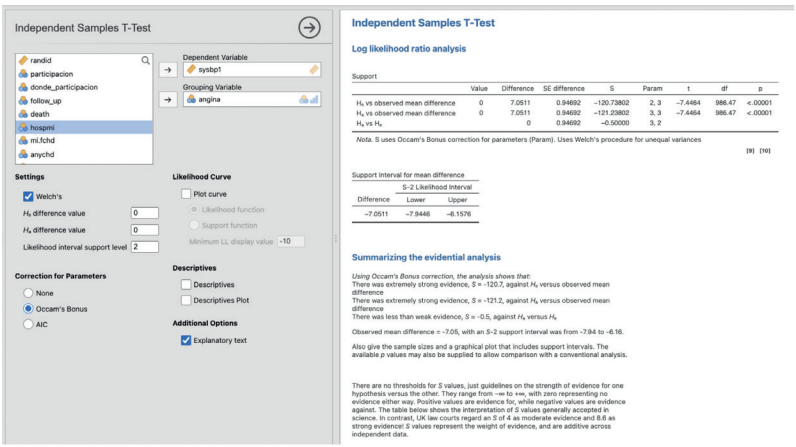


Figura 17. Resultados con JEVA

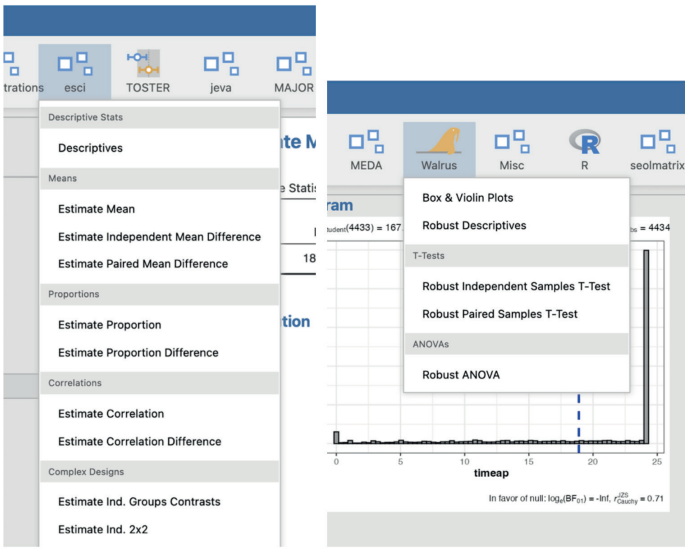
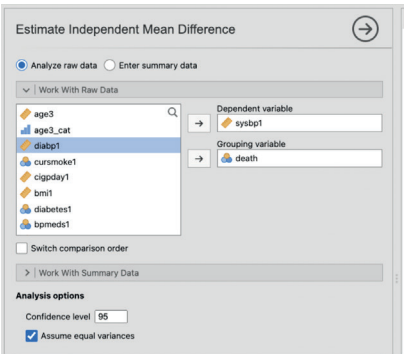


Figura 18. Jamovi de esci y Walrus



Estimate Independent Mean Difference

Compare Two Means

Condition	M	95 % CI Lower	95 % CI Upper	s	N
Yes	142.086	140.799	143.372	25.833	1550
No	127.826	127.149	128.503	18.536	2884
Difference	14.260	12.940	15.579	21.372	4434

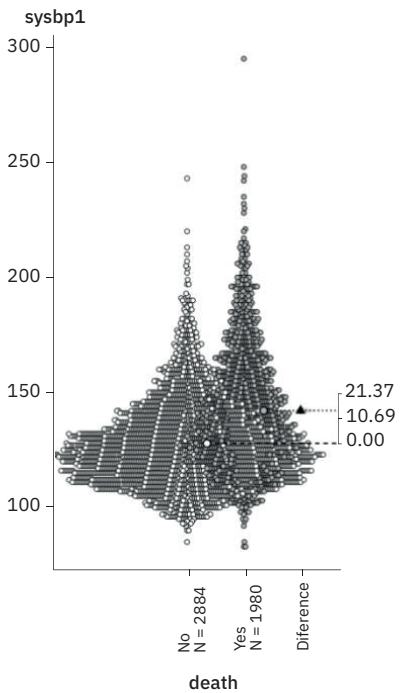
Notes

CIs are at the 95 % level.
This comparison was made on unpaired data.
Equal variance was assumed.
s in the row for the difference is the pooled standard deviation

Standardized Mean Difference

$d_{unbiased}$ = 0.67 95% CI [0.60, 0.73]
Note that the standardized effect size is d_unbiased
because the denominator used was SDpooled
which had a value of 21.4
The standardized effect size has been corrected for bias.
The bias-corrected version of Cohen's d is sometimes also (confusingly) called Hedges' g.

Descriptives plot



Mdiff = 14,260 95 % CI [12,940, 15,579]

Decision making

t-table

t	df	p
21.186	4432.0	<.00001

Figura 19. Resultados de esci

Referencias

The Jamovi project (2022). *Jamovi*. (Versión 2.3) [Computer Software]. Recuperado de: https://www.JAMOVI.org

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

Linlin Yan (2020). *Venn Diagram by ggplot2, with really easy-to-use API*. [R package]. Recuperado de: https://github.com/yanlinlin82/ggvenn.

Serdar Balci (2022). *ClinicoPath Jamovi Module doi:10.5281/zenodo.3997188*. [R package]. Recuperado de: https://github.com/sbalci/ClinicoPathJAMOVIModule https://www.serdarbalci.com/ClinicoPathJamoviModule/

Bjoern Koneswarakantha (2019). *easyalluvial: Generate Alluvial Plots with a Single Line of Code.* [R package]. Recuperado de: https://CRAN.R-project.org/package=easyalluvial

Nick Barrowman (2020). *vtree: Display Information About Nested Subsets of a Data Frame*. [R package]. Recuperado de: https://CRAN.R-project.org/package=vtree

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., y RStudio (2018). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. [R package]. Recuperado de: https://CRAN.R-project.org/package=ggplot2

Patil, I. (2018). *ggstatsplot: 'ggplot2' Based Plots with Statistical Details*. [R package]. Recuperado de: https://CRAN.R-project.org/package=ggstatsplot

Edwards, A. W. F. (2001). *Occam's Bonus*. Cambridge University Press.

Glover, S. (2018). *Likelihood Ratios; A Tutorial*. https://osf.io/preprints/metaarxiv/g3j2k/

Cahusac, P.M.B. (2020). *Evidence-Based Statistics*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119549833

Capítulo 7
Medidas utilizadas en epidemiología

La **epidemiología** es una disciplina científica en el área de la salud pública (no solo la medicina), que estudia la distribución, frecuencia, magnitud y factores determinantes de las enfermedades existentes en poblaciones humanas definidas. Rich la describió en 1979 como la ciencia que estudia la dinámica de salud en las poblaciones; por lo tanto, involucra el análisis e interpretación de las personas que también están sanas.

Cocientes

En epidemiología se utilizan varios tipos de medidas, de entre las cuales señalaremos los **cocientes**, entre los que se encuentran las proporciones, las razones, las tasas y los odds.

Las **proporciones** son cocientes en los que el numerador está incluido en el denominador, no tienen unidades y su rango varía entre 0 y 1 (0 % y 100 % si se expresa en tanto por ciento).

Por ejemplo, en una población con 500 hombres y 700 mujeres, la proporción de hombres es 500 / (500 + 700) = 0,4167 (41,67 %) y la de mujeres es 700 / (500 + 700) = 0,5833 (58,33 %). Se podría interpretar que en esta población un 41,67 % son hombres y un 58,33 % mujeres.

Las **razones** (*ratio* en inglés) son cocientes en los que el numerador no está incluido en el denominador, no tienen unidades, pero su rango no está acotado, variando entre 0 e infinito.

En el ejemplo anterior, la razón de masculinidad es 500 / 700 = 0,7143 y la razón de feminidad 700 / 500 = 1,4. Se podría interpretar que por cada hombre hay 1,4 mujeres.

Las **tasas** (*rates* en inglés) son cocientes que indican el número de individuos de una población (o muestra) con un atributo. El denominador indica la población total, pero incluye el tiempo. Se interpreta como la velocidad de ocurrencia del atributo en la población (o muestra). Las tasas más utilizadas en epidemiología son la tasa de natalidad (nacimientos/población en un período determinado), la tasa de fecundidad (nacimientos/mujeres en edad fértil en un período determinado), la tasa bruta de mortalidad (defunciones/población en un período determinado).

Por convención, en el denominador se utiliza el número de individuos estimado en la mitad del período (usualmente un año). Normalmente se dan en tanto por mil.

Por ejemplo, supongamos una población de 8500 habitantes en los que en un año se producen 80 infartos de miocardio: la tasa de infarto es de 80 / 8500 = 0,0094, es decir, 9,4 infartos por cada 1000 habitantes por año.

Los **odds** (a veces se traducen como 'ventajas') son un caso particular de razón y son el cociente entre individuos que tienen un atributo y otros que no lo tienen. Tampoco tienen unidad.

Por ejemplo, de 90 pacientes dados de alta en el hospital A, tenemos 60 reingresos. El odds (de reingresar) se calcularía como odds = 60 / (90 – 60) = 2. Es decir, los reingresos son 2 veces más frecuentes que los no reingresos.

Medidas de frecuencia

Las **medidas de frecuencia** cuantifican cuán frecuente o común resulta determinado fenómeno de interés (enfermedad, lesión, muerte).

Entre las medidas de frecuencia señalaremos la prevalencia y la incidencia.

Las medidas de **prevalencia** informan de la proporción de personas de una población con una característica determinada (por ejemplo, una enfermedad o un factor de riesgo, etc.) en un momento o período de tiempo determinado. Puede interpretarse como la probabilidad individual de una persona de la población de tener la característica. Es una medida propia de los estudios transversales. Existen dos medidas de prevalencia, la **prevalencia puntual** (la más utilizada) y la **prevalencia de período**.

Por ejemplo, en una población de 10000 personas, se informa de que 500 personas sufren determinada enfermedad. La prevalencia (o prevalencia puntual) es de 500 / 10000 = 0,05, es decir del 5 %. También se puede interpretar que la probabilidad de que una persona de esa población tenga la enfermedad es del 5 %. Si en esa misma población, 2430 personas desarrollan la enfermedad en un período de 5 años, la prevalencia de período es 2430 / 10000 = 0,243, 24,3 %. Es decir, que la probabilidad de que una persona de esa población tenga la enfermedad en un período de 5 años es del 23,4 %.

Las medidas de **incidencia** se fijan en el número de casos nuevos de ocurrencia de un evento (por ejemplo, enfermedad) que aparecen entre las personas susceptibles (por ejemplo, de enfermar) de una población

a lo largo de un período de tiempo. Es necesario un seguimiento. Se calculan sobre cohortes.

Igualmente existen dos medidas de incidencia, la **incidencia acumulada** y la **tasa o densidad de incidencia**.

En la **incidencia acumulada**, las personas controladas son fijas y deben estar libres del evento al inicio del seguimiento. Puede interpretarse como la probabilidad individual de que una persona de la población desarrolle la característica.

Por ejemplo, en una población de 1230 personas sin COVID-19, en una semana se infectaron 28. La **incidencia acumulada** se calculará como:

(número de casos nuevos durante el período) /
/ (número de casos susceptibles al principio del período) =
= 28 / 1230 = 0,0227

También se suele expresar por mil. En este caso la incidencia acumulada en 7 días es de 22,7 por cada 1000 habitantes.

La **densidad de incidencia** es el cociente entre los nuevos casos durante el período y la suma de los tiempos de seguimiento de cada individuo. Más adecuado para poblaciones dinámicas. Informa sobre la velocidad del acontecimiento. A diferencia de la incidencia acumulada no tiene interpretación individual.

Por ejemplo, se sigue durante 12 meses a un grupo de 6 individuos que han sufrido un accidente cerebrovascular (ACV), para evaluar la incidencia de recidiva. La situación puede esquematizarse en la imagen:

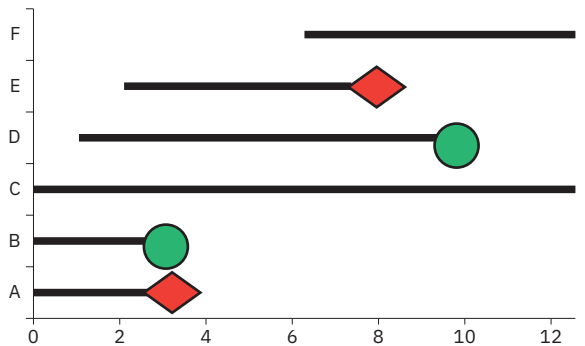


Figura 1. Seguimiento de 6 individuos con ACV durante 12 meses

Es una cohorte dinámica en la que se permite incorporaciones. Por tanto, el seguimiento no empezó al mismo tiempo para todos los individuos.

Los rombos rojos señalan recidivas (pacientes A y E), mientras que los círculos verdes señalan pérdidas. Así, el paciente B desapareció del estudio a los 2,5 meses, sin que hasta ese momento hubiera sufrido una recidiva; el paciente D falleció por otra causa no relacionada con el ACV; los pacientes C y F terminaron el período de estudio sin recidiva.

La densidad de incidencia se calcularía como:

(número de casos nuevos durante el período) /
/ (suma del tiempo de seguimiento de los casos susceptibles,
es decir sin la enfermedad)

El número de casos nuevos durante el período de seguimiento es igual a 2.

El tiempo de seguimiento de los casos susceptibles:

- Individuo A: 3 meses (hasta que tuvo la recidiva)
- Individuo B: 2,5 meses
- Individuo C: 12 meses
- Individuo D: 8 meses
- Individuo E: 6 meses (hasta que tuvo la recidiva).
- Individuo F: 6 meses
- Total tiempo de seguimiento: 3 + 2,5 + 12 + 8 + 6 + 6 = 37,5 meses.

Por tanto,

densidad de incidencia = 2 / 37,5 = 0,053 por cada mes

Medidas de asociación

Las **medidas de asociación** miden la asociación entre dos variables dicotómicas, el factor de riesgo y la enfermedad o muerte (en ambos casos, presencia o ausencia).

En general, las medidas de asociación cuantifican el riesgo. El término *riesgo* implica que la presencia de una característica o factor aumenta la probabilidad de consecuencias adversas. El riesgo se puede medir mediante:

- **Riesgo absoluto**. Mide la incidencia del daño en la población total.
- **Riesgo relativo**. Compara la frecuencia con que ocurre el daño entre los que tienen el factor de riesgo y los que no lo tienen.

Por ejemplo, mostramos la tabla de un estudio de seguimiento donde 853 mujeres estuvieron pasivamente expuestas al humo del tabaco durante la gestación y 1620 no lo estuvieron, y su asociación con el bajo peso de la criatura al nacer.

Tabla 1. Distribución de gestantes según exposición pasiva al humo de tabaco y recién nacidos según bajo peso o peso normal

Exposición pasiva al tabaco	Recién nacido bajo peso		Total
	Sí	No	
Sí	20	833	853
No	14	1606	1620
Total	34	2439	2473

Riesgo (absoluto) de bajo peso en las gestantes expuestas al tabaco:

$$20 / 853 = 0{,}0234$$

Riesgo (absoluto) de bajo peso en las gestantes no expuestas al tabaco:

$$14 / 1620 = 0{,}0086$$

Es decir, el riesgo (o probabilidad, puesto que se trata de una proporción) de que una gestante expuesta al tabaco tenga un recién nacido con bajo peso es del 2,34 %, mientras que en una gestante no expuesta al tabaco es del 0,86 %.

El riesgo relativo mide la fuerza de la asociación entre la exposición y la enfermedad. Indica la probabilidad de que se desarrolle la enfermedad en los expuestos a un factor de riesgo en relación con el grupo de los no expuestos.

Siguiendo con nuestro ejemplo, calcularemos el **riesgo relativo**.

Riesgo relativo (RR) = (Riesgo de bajo peso en las gestantes expuestas al tabaco) / (Riesgo de bajo peso en las gestantes no expuestas al tabaco)

$$RR = 0{,}0234/0{,}0086 = 2{,}7209$$

El riesgo de que una gestante expuesta al tabaco tenga un recién nacido con bajo peso es 2,72 mayor que en una gestante no expuesta al tabaco. Como es un cociente de dos proporciones también se puede expresar como probabilidad. Por otra parte, se puede interpretar de la siguiente manera:

$$(RR - 1) * 100 \; (2{,}7209 - 1) * 100 = 172{,}09\,\%$$

Una gestante expuesta al tabaco tiene un 172,09 % mayor riesgo (o probabilidad) que una gestante no expuesta, de que el recién nacido tenga bajo peso.

Por otra parte, si los eventos son infrecuentes (lo más habitual en ciencias de la salud) el riesgo también se puede medir mediante el **odds ratio** (OR) (también denominado *razón de ventajas*, entre muchos otros términos). En los estudios de **casos y controles**, dado que la incidencia es desconocida, el método de estimación del riesgo relativo es diferente y se estima calculando el OR.

En el ejemplo que vimos más arriba, de 90 pacientes dados de alta en el hospital A, tenemos 60 reingresos, mientras que en el hospital B, de 50 pacientes dados de alta, tenemos 30 reingresos.

- Odds (de reingresar) en el hospital A = 60 / (90 – 60) = 2.
- Odds (de reingresar) en el hospital B = 30 / (50 – 30) = 1,5.

Así, el odds ratio (razón de odds) es igual a 2 / 1,5 = 1,33

El riesgo de reingresar en el hospital A es 1,33 mayor que en el hospital B. También

$$(OR - 1) * 100 = (1{,}33 - 1) * 100$$

El hospital A tiene un 33 % mayor de riesgo de reingreso que el hospital B.

Como los odds no son proporciones, el OR no se puede interpretar como probabilidad.

Tanto en el RR como en el OR, cuando son mayores que 1 el factor es un **factor de riesgo** de la enfermedad, cuando es menor que 1 es un **factor protector** y cuando es igual a 1, no existe relación entre el factor y la enfermedad.

Medidas de impacto

Las **medidas de impacto** describen el posible impacto potencial en la población que pudiera tener la eliminación del factor de riesgo en cuestión sobre el desarrollo de la enfermedad; son las medidas de impacto potenciales, de las que solo es posible su cálculo en diseños de prevención de enfermedad, como los estudios de cohortes y casos y controles. Las medidas de impacto potenciales se dividen en absolutas, cuando miden el exceso de riesgo en los expuestos en la población que se

toma de referencia, y en relativas, cuando miden el porcentaje de riesgo que se debe a exposición.

Medidas de impacto absoluto

El **riesgo atribuible en los expuestos** (RAE) o, también, **reducción absoluta de riesgo** (RAR), se calcula como la diferencia entre la incidencia de expuestos y no expuestos (RAE = RAR = $I_{expuestos} - I_{no\ expuestos}$). La diferencia entre ambos valores da el valor del riesgo de enfermedad en la cohorte expuesta, que se debe exclusivamente a la exposición.

En el ejemplo de la tabla 1:

- Incidencia de bajo peso en las gestantes expuestas al tabaco: 20 / 853 = 0,0234
- Incidencia de bajo peso de las gestantes no expuestas al tabaco: 14 / 1620 = 0,0086

$$RAE = RAR = 0{,}0234 - 0{,}0086 = 0{,}0148$$

Se podría interpretar que suprimiendo la exposición al tabaco en las gestantes evitaríamos por término medio 1,48 (redondeando 1,5) recién nacidos de bajo peso por cada 100 nacimientos en un período determinado.

Utilizando el RAE se puede calcular el **número necesario de pacientes a tratar** para reducir un efecto o **NNT** como

$$NNT = 1 / RAE = 1 / (I_{expuestos} - I_{no\ expuestos})$$

Si el RAE es negativo se utiliza el **número necesario para dañar, NNH**, que se obtiene simplemente cambiando el orden del denominador,

$$NNH = 1 / RAE = 1 / (I_{no\ expuestos} - I_{expuestos})$$

El NNT y el NNH son unas medidas de asociación utilizadas en ensayos clínicos.

Cuanto menor sea el NNT (mayor la diferencia entre la incidencia en los expuestos) mayor efectividad tendrá el tratamiento del ensayo clínico. Y cuanto mayor sea el NNH más seguro.

Por ejemplo, en la tabla 2, vemos que los regímenes intensivos de insulina en pacientes diabéticos para prevenir las retinopatías son muy efectivos. El NNT es igual a 4. Es decir, necesitaríamos tratar (con un régimen

intensivo de insulina) a 4 pacientes diabéticos para evitar 1 caso de retinopatía. Sin embargo, para prevenir la neuropatía diabética es poco efectivo. El NNT es igual a 15. Necesitaríamos tratar (con un régimen intensivo de insulina) a 15 pacientes diabéticos para evitar 1 caso de neuropatía diabética. Mucho menos efectiva es la exploración de mamas, además de la mamografía, en mujeres sanas de 50 a 69 años, para prevenir la muerte por cáncer de mama. Se necesitarían 1075 mujeres para evitar 1 caso.

Medidas de impacto relativo

La **fracción atribuible** estima la proporción de la enfermedad entre los expuestos que puede ser atribuible al hecho de estar expuestos. La fracción atribuible se puede calcular en los expuestos y en la población.

La **fracción atribuible en el grupo expuesto (fracción etiológica** o **porcentaje de riesgo atribuible en los expuestos)** (FAE), establece el grado de influencia que tiene la exposición en la presencia de enfermedad entre los expuestos. Su cálculo se realiza como:

$$FAE = (I_{expuestos} - I_{no\ expuestos}) / I_{expuestos}$$

Con los datos de la tabla 1,

$$FAE = (0{,}0234 - 0{,}0086) / 0{,}0234 = 0{,}6325$$

que se interpretaría como que el 63,25 % del bajo peso en los recién nacidos de gestantes expuestas al tabaco se debe a la exposición.

La **fracción atribuible en la población** (FAP), muestra la proporción en que el daño podría ser reducido si los factores de riesgo causales desapareciesen de la población total. Se calcularía como:

$$FAP = (I_{población} - I_{no\ expuestos}) / I_{población}$$

Si disponemos de la incidencia de la exposición en la población, el cálculo también se puede realizar del siguiente modo con esta fórmula alternativa,

$$FAP = Pt\ (RR - 1) / (Pt\ (RR - 1) + 1)$$

donde Pt denota la relevancia de la exposición (o factor de riesgo) en la población. Así pues, la fracción atribuible en la población total es una

Tabla 2. NNT para tratamientos diferentes

Enfermedad	Intervención	Episodios que se previenen	Tasa en el grupo control	Tasa en el grupo experimental	Duración del seguimiento	NNT para evitar un episodio adicional
Diabetes (DMID)[1]	Regímenes intensivos de insulina	Neuropatía diabética	0.096	0.028	6.5 años	15
Diabetes (DMNID)[2]	Regímenes intensivos de insulina	Retinopatías	0.38	0.13	6 años	4
		Nefropatía	0.30	0.10	6 años	5
Infartio de miocardio[3]	Estreptoquinasa y aspirina	Muerte a las 5 semanas	0.134	0.081	5 semanas	19
		Muerte a los 2 años	0.216	0.174	2 años	24
Presión arterial diastólica 115-129 mmHg[4]	Fármacos antihipertensivos	Muerte, apoplejía o infarto de miocardio	0.0545	0.0467	5.5 años	128
Personas mayores independientes[5]	Estudio geriátrico exhaustivo	Permanencia en residencias por un largo período de tiempo	0.10	0.04	3 años	17
Mujeres embarazadas con eclampsia[6]	MgSO4 iv (vs diacepan)	Convulsiones recurrentes	0.279	0.132	Horas	7
Mujeres sanas de edad 50-69 años[7]	Exploración de mamas además de mamografía	Muerte por cáncer de mama	0.00345	0.00252	9 años	1075
Estenosis grave sintomática de la arteria[8]	Endarterectomía	Aplplejía total o muerte	0.181	0.08	2 años	10
Niños prematuros[9]	Corticosteroides prenatales	Síndrome de distrés respiratorio	0.23	0.13	Días	11

Fuente: Pita Fernández y López de Ullibarri Galparsoro (2001).

medida de asociación influenciada por la prevalencia del factor de riesgo en la población total.

Para calcularla en nuestro ejemplo necesitaremos la prevalencia del bajo peso al nacer que, en España, se estima en torno al 7,8 %. Así pues,

$$FAP = 0{,}078 * (2{,}7209 - 1) / (0{,}078 * (2{,}7209 - 1) + 1) =$$
$$= 0{,}1342302 / 1{,}1342302 = 0{,}1183$$

El riesgo de bajo peso en la población atribuible a la exposición de las gestantes al tabaco es del 11,83 %.

Aplicación con la base de datos de Framingham

Para ejemplificar las medidas de asociación, nos centraremos en las variables de la base de datos Framingham:

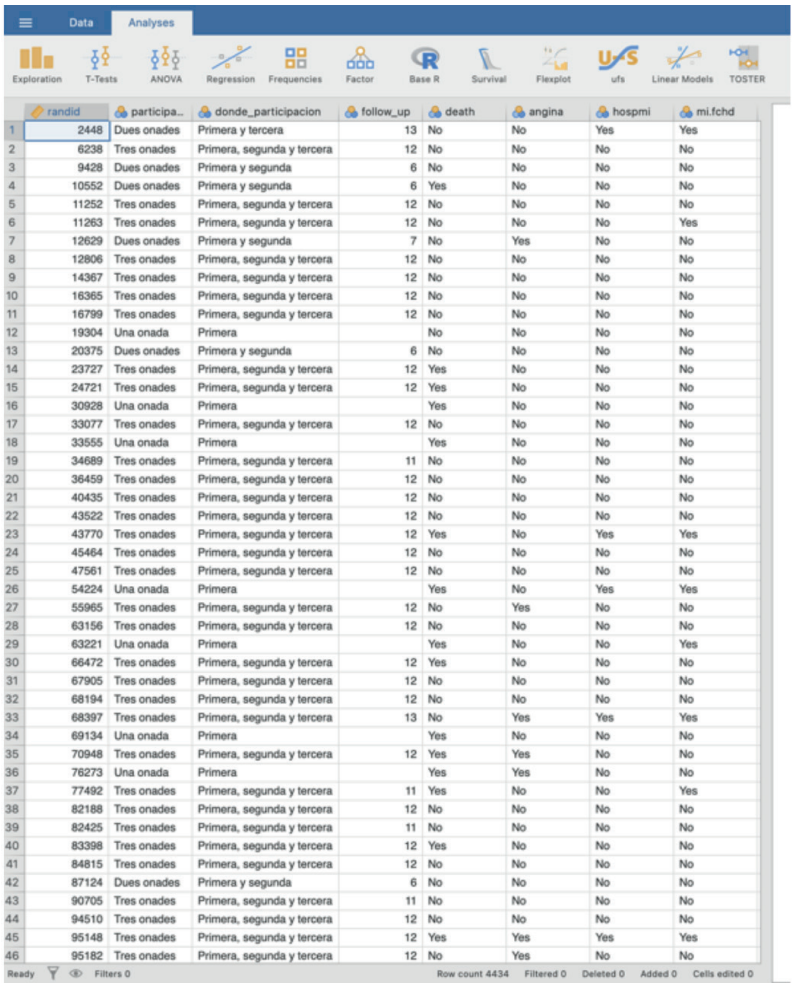
- **death**. Muerte por cualquier causa.
- **sexo.** Sexo del participante.
- **obese.** 1 Obesidad (IMC ≥ 30 kg/m^2), 0 otro caso.

Utilizaremos Jamovi para facilitar el análisis que vamos a realizar. En primer lugar, cargamos los datos del fichero Framingham (figura 2).

Mediremos la asociación entre la obesidad en el momento de entrada en la cohorte (*obese1*) y la muerte (*death*). En la pestaña «Análisis» elegimos el menú «Frecuencias» y en el apartado «Tablas de contingencia» ubicamos la variable *death* por columnas y la variable *obese* por filas. En «Celdas» marcamos porcentajes por filas. Además, en la barra «Estadísticas» desmarcamos todas las opciones.

Como vemos, la incidencia de muerte entre los sujetos que entraron obesos en la cohorte es del 44,9 %, mientras que entre los sujetos que no eran obesos al entrar fue del 33,3 %.

Estimaremos el riesgo relativo, señalando en «Estadísticos», riesgo relativo e intervalos de confianza.



	randid	participa...	donde_participacion	follow_up	death	angina	hospmi	mi.fchd
1	2448	Dues onades	Primera y tercera	13	No	No	Yes	Yes
2	6238	Tres onades	Primera, segunda y tercera	12	No	No	No	No
3	9428	Dues onades	Primera y segunda	6	No	No	No	No
4	10552	Dues onades	Primera y segunda	6	Yes	No	No	No
5	11252	Tres onades	Primera, segunda y tercera	12	No	No	No	No
6	11263	Tres onades	Primera, segunda y tercera	12	No	No	No	Yes
7	12629	Dues onades	Primera y segunda	7	No	Yes	No	No
8	12806	Tres onades	Primera, segunda y tercera	12	No	No	No	No
9	14367	Tres onades	Primera, segunda y tercera	12	No	No	No	No
10	16365	Tres onades	Primera, segunda y tercera	12	No	No	No	No
11	16799	Tres onades	Primera, segunda y tercera	12	No	No	No	No
12	19304	Una onada	Primera		No	No	No	No
13	20375	Dues onades	Primera y segunda	6	No	No	No	No
14	23727	Tres onades	Primera, segunda y tercera	12	Yes	No	No	No
15	24721	Tres onades	Primera, segunda y tercera	12	Yes	No	No	No
16	30928	Una onada	Primera		Yes	No	No	No
17	33077	Tres onades	Primera, segunda y tercera	12	No	No	No	No
18	33555	Una onada	Primera		Yes	No	No	No
19	34689	Tres onades	Primera, segunda y tercera	11	No	No	No	No
20	36459	Tres onades	Primera, segunda y tercera	12	No	No	No	No
21	40435	Tres onades	Primera, segunda y tercera	12	No	No	No	No
22	43522	Tres onades	Primera, segunda y tercera	12	No	No	No	No
23	43770	Tres onades	Primera, segunda y tercera	12	Yes	No	Yes	Yes
24	45464	Tres onades	Primera, segunda y tercera	12	No	No	No	No
25	47561	Tres onades	Primera, segunda y tercera	12	No	No	No	No
26	54224	Una onada	Primera		Yes	No	Yes	Yes
27	55965	Tres onades	Primera, segunda y tercera	12	No	Yes	No	No
28	63156	Tres onades	Primera, segunda y tercera	12	No	No	No	No
29	63221	Una onada	Primera		Yes	No	No	Yes
30	66472	Tres onades	Primera, segunda y tercera	12	Yes	No	No	No
31	67905	Tres onades	Primera, segunda y tercera	12	No	No	No	No
32	68194	Tres onades	Primera, segunda y tercera	12	No	No	No	No
33	68397	Tres onades	Primera, segunda y tercera	13	No	Yes	Yes	Yes
34	69134	Una onada	Primera		Yes	No	No	No
35	70948	Tres onades	Primera, segunda y tercera	12	Yes	Yes	No	No
36	76273	Una onada	Primera		Yes	Yes	No	No
37	77492	Tres onades	Primera, segunda y tercera	11	Yes	No	No	Yes
38	82188	Tres onades	Primera, segunda y tercera	12	No	No	No	No
39	82425	Tres onades	Primera, segunda y tercera	11	No	No	No	No
40	83398	Tres onades	Primera, segunda y tercera	12	Yes	No	No	No
41	84815	Tres onades	Primera, segunda y tercera	12	No	No	No	No
42	87124	Dues onades	Primera y segunda	6	No	No	No	No
43	90705	Tres onades	Primera, segunda y tercera	11	No	No	No	No
44	94510	Tres onades	Primera, segunda y tercera	12	No	No	No	No
45	95148	Tres onades	Primera, segunda y tercera	12	Yes	Yes	Yes	Yes
46	95182	Tres onades	Primera, segunda y tercera	12	No	Yes	No	No

Figura 2. Ventana de los datos Framingham

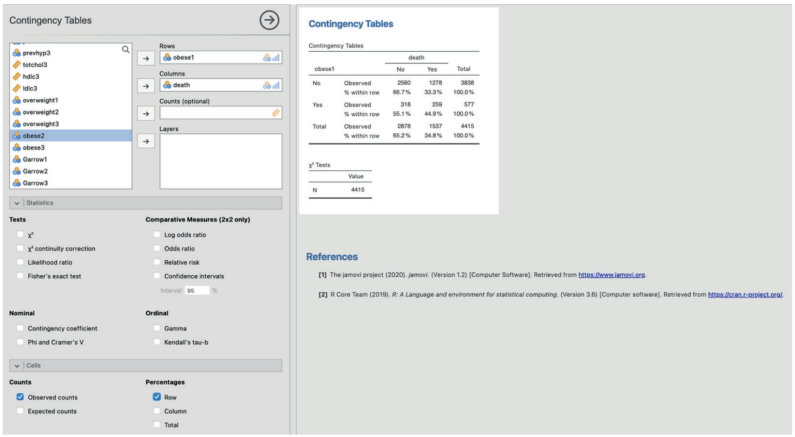


Figura 3. Ventanas que aparecen en «Tablas de contingencia»

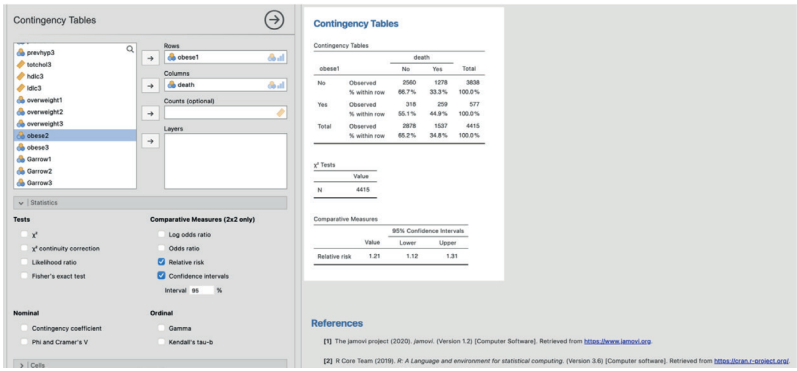


Figura 4. Ventanas que aparecen en «Tablas de contingencia»

El riesgo relativo es igual a 1,21. Nótese que es estadísticamente significativo (al 95 %), pues su intervalo de confianza no contiene el 1. Téngase en cuenta que la misma información la obtenemos utilizando el contraste chi-cuadrado de independencia (o, también, diferencia de proporciones).

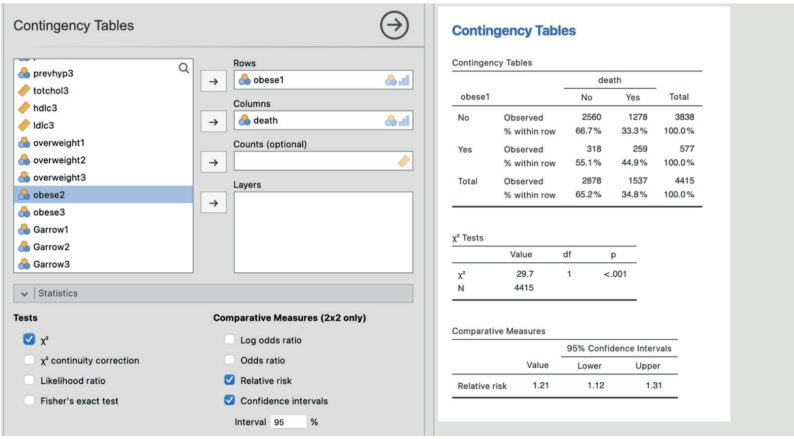


Figura 5. Resultados del comando de «Tablas de contingencia»

Rechazamos la hipótesis nula (de independencia o de igualdad de proporciones), puesto que el p-valor del contraste es menor que el riesgo (< 0,001 es menor que 0,05).

Así pues, ser obeso es un factor de riesgo de morir. En concreto, si el sujeto entró obeso en la cohorte tuvo un 21 % más de riesgo de morir que si no era obeso cuando entró.

Aunque siendo una cohorte no sería apropiado medir el riesgo utilizando el odds ratio, lo estimaremos también.

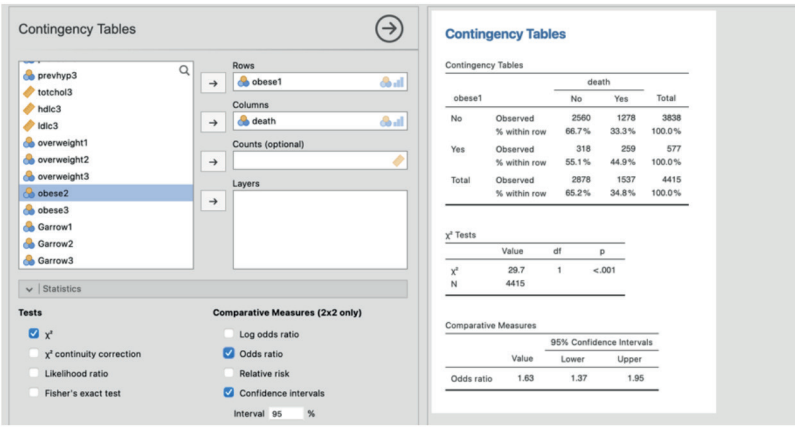


Figura 6. Resultados del comando de «Tablas de contingencia»

El odds ratio, también estadísticamente significativo, es igual a 1,63.

Calcularemos ahora las medidas de impacto.

El riesgo atribuible en los expuestos (RAE) se calcularía como

$$RAE = I_{expuestos} - I_{no\ expuestos} = 0{,}449 - 0{,}333 = 0{,}116$$

Si ninguno de los sujetos hubiera sido obeso habríamos evitado, por término medio, 11,6 (redondeando, 12) muertes por cada 100 sujetos vivos.

La fracción atribuible en el grupo expuesto (FAE) se calcularía como

$$FAE = (I_{expuestos} - I_{no\ expuestos}) / I_{expuestos} = \\ = (0{,}449 - 0{,}333) / 0{,}449 = 0{,}2584$$

El 25,84 % de las muertes se debe a la obesidad.

Referencias

The Jamovi project (2022). *Jamovi*. (Versión 2.3) [Computer Software]. Recuperado de: https://www.JAMOVI.org

Pita Fernández, S., Vila Alonso, M. T. y Carpente Montero, J. (1997). Determinación de los factores de riesgo. *Cad Aten Primaria*; 4:75-78. Disponible en (último acceso el 20 de mayo de 2024): https://www.fisterra.com/formacion/metodologia-investigacion/determinacion-factores-riesgo/

Pita Fernández, S. y López de Uribarri Galparsoro, I. (1998, revisado en 2001). Número necesario a tratar para reducir un evento. *Cad Aten Primaria*; 5:96-98. Disponible en (último acceso el 20 de mayo de 2024): https://www.fisterra.com/formacion/metodologia-investigacion/numero-necesario-pacientes-tratar-para-reducir-evento/

Porta, M. (editor), Greenland, S., Hernán, M., dos Santos Silva, I., Last, J. M. (associate editors) (2014). *A dictionary of epidemiology* [6.ª edición]. Nueva York: Oxford University Press.

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

Capítulo 8
Introducción a la regresión

En este capítulo, vamos a ver de forma introductoria distintos tipos de regresión, como son la lineal y la logística. La regresión estadística es una herramienta fundamental para comprender las relaciones entre variables y tomar decisiones informadas en diversos ámbitos. Su versatilidad y capacidad para modelar relaciones complejas la convierten en un método indispensable para el análisis de datos.

Regresión lineal

La **regresión lineal** es una técnica estadística ampliamente utilizada para modelar y analizar la relación entre una **variable dependiente** y una o más **variables independientes**. La regresión lineal busca encontrar la mejor **línea recta** que se ajuste a los datos observados. En su forma más simple, asume una relación **lineal** entre las variables. Esto nos lleva más allá de la simple relación lineal; se aplica en diversos campos, como economía, ciencias sociales, medicina e ingeniería.

Tipos de regresión

- **Regresión lineal simple**. Analiza la relación entre una variable dependiente y una variable independiente.
- **Regresión lineal múltiple**. Estudia la relación entre una variable dependiente y dos o más variables independientes, considerando múltiples factores.

Empezaremos con la regresión lineal simple, donde utilizaremos variables continuas para ver su relación. En general, utilizaremos las mismas variables que en los capítulos 5 y 6.

Inicialmente, podemos ver el diagrama de dispersión, que, en la imagen siguiente, nos muestra Jamovi (figura 1). También podemos ver en la imagen siguiente, el diagrama de dispersión, pero separando por otra variable (figura 2).

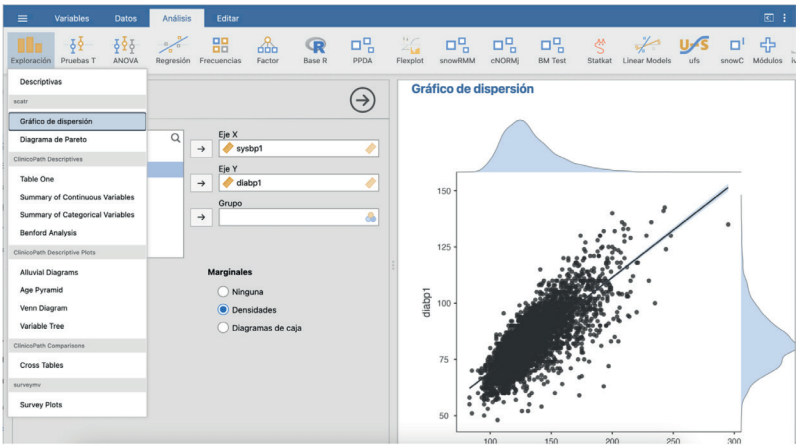


Figura 1. Diagrama de dispersión

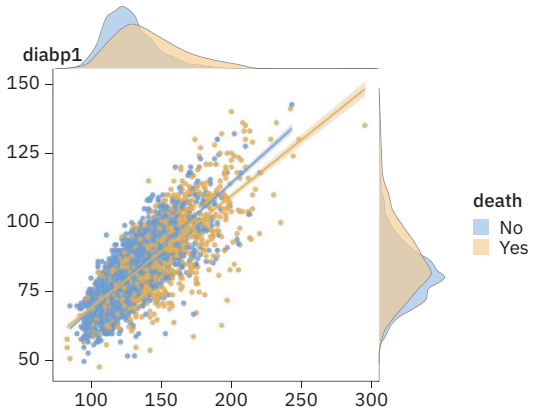


Figura 2. Diagrama de dispersión, separando por clases

Después de ver gráficamente si hay relación entre las variables, podemos realizar el análisis propiamente dicho (figura 3).

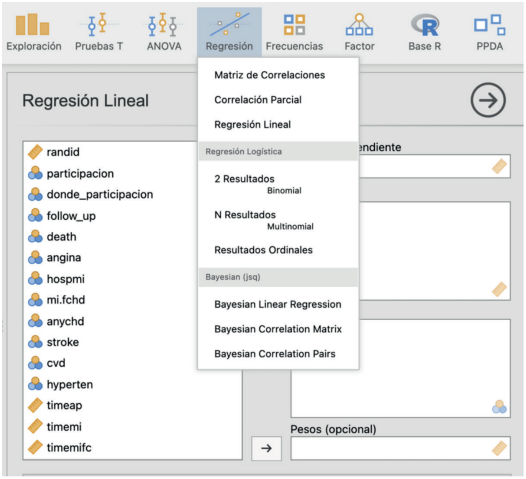


Figura 3. Imagen de Jamovi para realizar la regresión lineal

Con todo ello, obtenemos los siguientes resultados como ejemplo de aplicación de regresión lineal:

Tabla 1. Resultados de la regresión lineal entre sysbp1 y diabp1

Medidas de ajuste del modelo

Modelo	R	R^2
1	0.78418	0.61493

Coeficientes del modelo - sysbp1

Predictor	Estimador	EE	t	p
Constante	11.6226	1.455594	7.9848	<.00001
diabp1	1.4586	0.017338	84.1286	<.00001

La correlación entre dos variables cuantitativas —el coeficiente r de Pearson— se puede evaluar siempre, no comporta que haya una variable dependiente y una independiente, simplemente nos muestra si ambas variables cambian (correlacionan) al mismo tiempo. En cambio, el modelo de regresión lineal —como ocurre también con el diagrama de dispersión— conlleva ya pensar en una variable dependiente (Y = efecto o respuesta) y una independiente (X = causa o tratamiento). En estos casos, la intensidad de la relación nos la da el coeficiente de determinación, un estadístico que nos dice qué parte de la variabilidad total es explicada por la variable independiente. Se trata de un coeficiente que varía entre 0 y 1, y que en porcentaje nos indica precisamente esto, la proporción

de la variabilidad total de Y explicada por X. Además, suele ser habitual que nos interese cuál es la función de la recta de regresión (la recta que Jamovi ha dibujado en el gráfico de dispersión). Esta función nos permitirá predecir, para cada valor de la variable independiente, el valor que toma, en promedio, la variable dependiente. Como podemos ver en los resultados de la tabla anterior, el coeficiente de correlación es 0,7841 y el de determinación 0,6149.

A partir de la segunda parte de la tabla, podemos obtener la recta de regresión entre las dos variables (Y = a + bX), siendo en este caso a=11,62 y b = 1,46. La pendiente es el aumento de la variable Y al incrementar en una unidad la variable X. En este caso diremos que, para cada diabp1 de más, *sysbp1* incrementa, en media, en 1,45. Es decir, con cada punto de más de *diabp1*, aumenta *sysbp1* en 1,45. El signo de la pendiente será siempre el mismo que el del coeficiente de correlación de Pearson, por lo que también indica si se trata de una relación directa o inversa. Obviamente, el signo debe ser el mismo que la pendiente de la recta dibujada en el diagrama de dispersión. Además, el hecho de tener la recta nos permite también hacer predicción de datos.

En Jamovi también podemos obtener mucha más información sobre la regresión realizada, como podemos ver en las siguientes imágenes (figura 4).

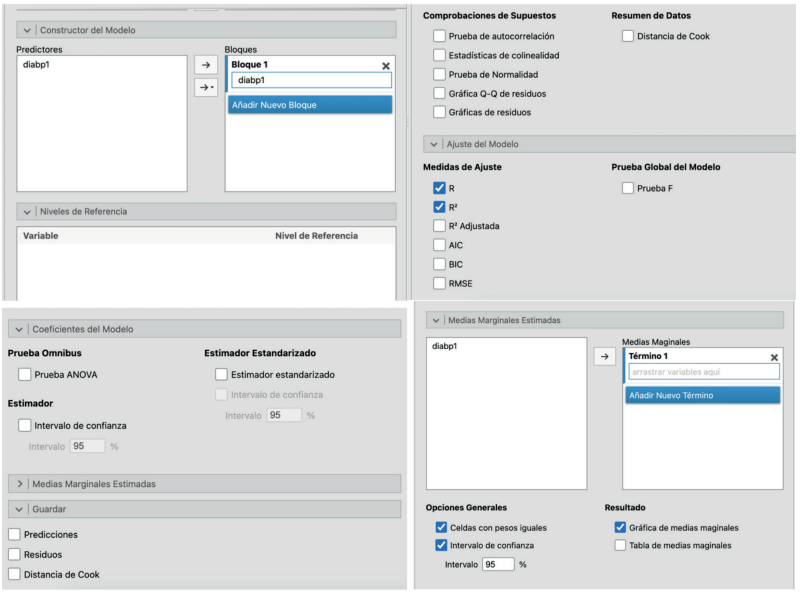


Figura 4. Distintas opciones de la regresión lineal

También vamos a analizar los datos para ver normalidad y homoscedasticidad de los residuos.

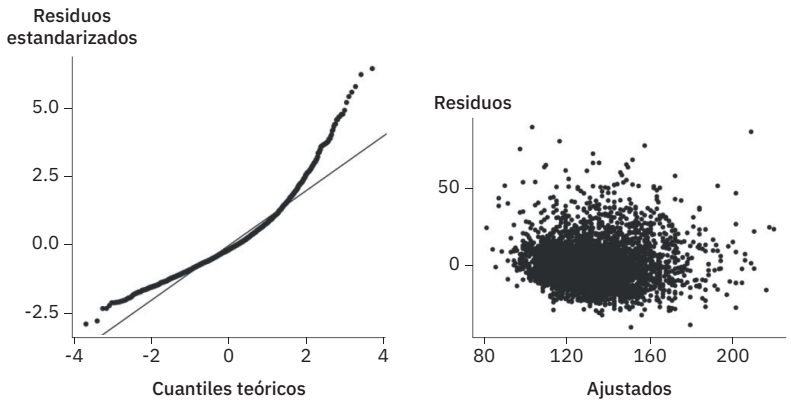


Figura 5. Distintas opciones para estudiar normalidad y homocesdasticidad

Para acabar con este apartado, vamos a introducir la matriz de correlaciones, junto a la inferencia asociada a ese propio análisis (figura 6). En este caso, podemos ver las correlaciones entre los distintos valores de *diabp* y *sysbp* en diferentes momentos temporales.

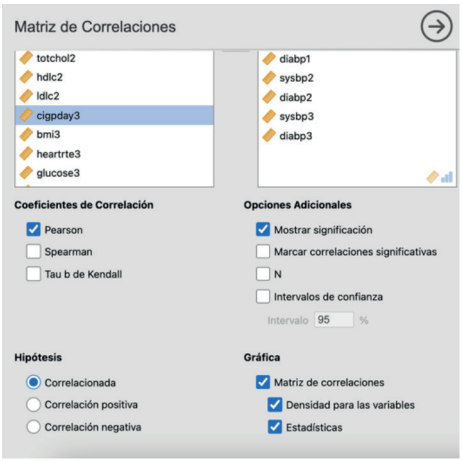


Figura 6. Distintas opciones para estudiar la matriz de correlaciones

En la tabla siguiente podemos ver las correlaciones entre cada par de valores y, además, su p-valor asociado, que en todos los casos coincide con < 0,0001, por lo que hay relación entre las variables. A continuación, ofrecemos también la imagen del gráfico de los diagramas de dispersión de todos los casos con sus valores de regresión, junto a las densidades de cada caso.

Tabla 2. Matriz de correlaciones entre sysbp y diabp en distintos momentos

Matriz de correlaciones

		sysbp1	diabp1	sysbp2	diabp2	sysbp3	diabp3
sysbp1	R de Pearson	—					
	valor p	—					
diabp1	R de Pearson	0.78418	—				
	valor p	< .00001	—				
sysbp2	R de Pearson	0.71368	0.56271	—			
	valor p	< .00001	< .00001	—			
diabp2	R de Pearson	0.50070	0.62216	0.71199	—		
	valor p	< .00001	< .00001	< .00001	—		
sysbp3	R de Pearson	0.62011	0.49124	0.67757	0.49212	—	
	valor p	< .00001	< .00001	< .00001	< .00001	—	
diabp3	R de Pearson	0.37774	0.49979	0.43028	0.56779	0.65613	—
	valor p	< .00001	< .00001	< .00001	< .00001	< .00001	—

Otras opciones, de las que solo comentaremos en este punto, pero que pueden ser desarrolladas, son las distintas relaciones que podemos obtener con las variables de estudio. En la imagen siguiente, vemos que con el paquete seolmatrix podemos desarrollar todo tipo de correlaciones, desde Sperman, incluso multinivel (figura 8). Además, nos permite también obtener la matriz de correlaciones y su relación de forma gráfica.

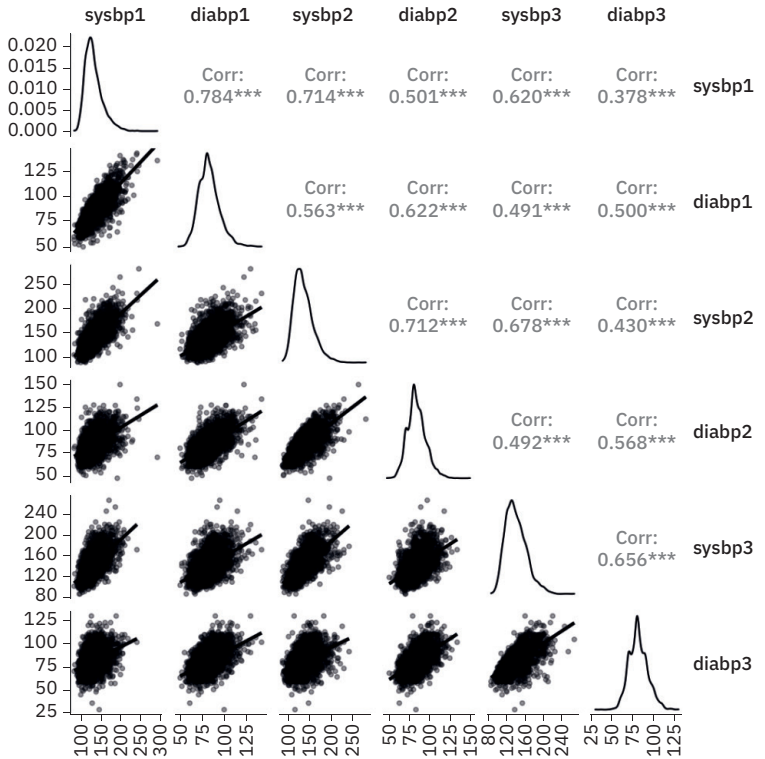
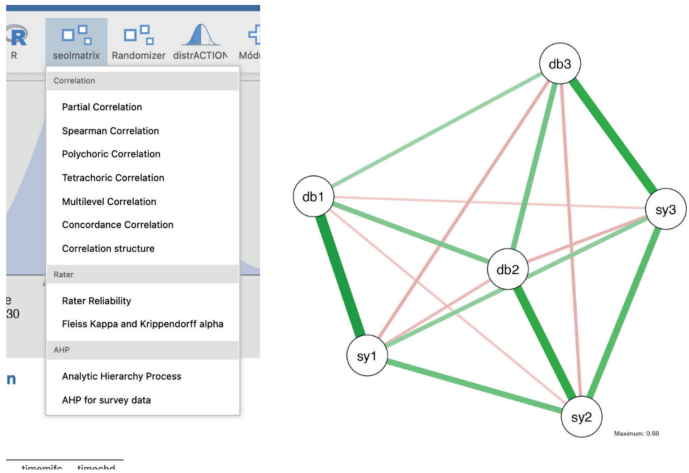


Figura 7. Distintas correlaciones



Spearman correlation

	sysbp1	diabp1	sysbp2	diabp2	sysbp3	diabp3
sysbp1	1.00000	0.78128	0.69727	0.50887	0.61120	0.37883
diabp1	0.78128	1.00000	0.56407	0.60645	0.48840	0.48818
sysbp2	0.69727	0.56407	1.00000	0.72125	0.69471	0.45463
diabp2	0.50887	0.60645	0.72125	1.00000	0.50174	0.56679
sysbp3	0.61120	0.48840	0.69471	0.50174	1.00000	0.66685
diabp3	0.37883	0.48818	0.45463	0.56679	0.66685	1.00000

Figura 8. Distintas opciones de seolmatrix (izquierda) y relaciones de forma gráfica entre variables (derecha)

Regresión lineal múltiple

Si ahora queremos relacionar una variable con varias a la vez, también lo podemos hacer. Por ejemplo, en la siguiente tabla, obtenemos los resultados de introducir más de una variable independiente, que mejora el ajuste por el aumento del valor de R.

Tabla 3. Resultados de regresión lineal múltiple

Medidas de ajuste del modelo

Modelo	R	R^2
1	0.87312	0.76234

Coeficientes del modelo - sysbp1

Predictor	Estimador	EE	t	p
Constante	6.81945	1.486615	4.5872	<.00001
diabp1	1.17198	0.019715	59.4459	<.00001
bmi1	0.63635	0.099680	6.3839	<.00001
sysbp2	0.53132	0.010810	49.1521	<.00001
diabp2	−0.54375	0.023041	−23.5989	<.00001
bmi2	−0.60074	0.096405	−6.2314	<.00001

Para acabar este apartado, vamos a realizar la regresión lineal múltiple con el paquete Lineal Models, que nos dará los mismos resultados para los coeficientes, pero nos ofrecerá más opciones (figura 9).

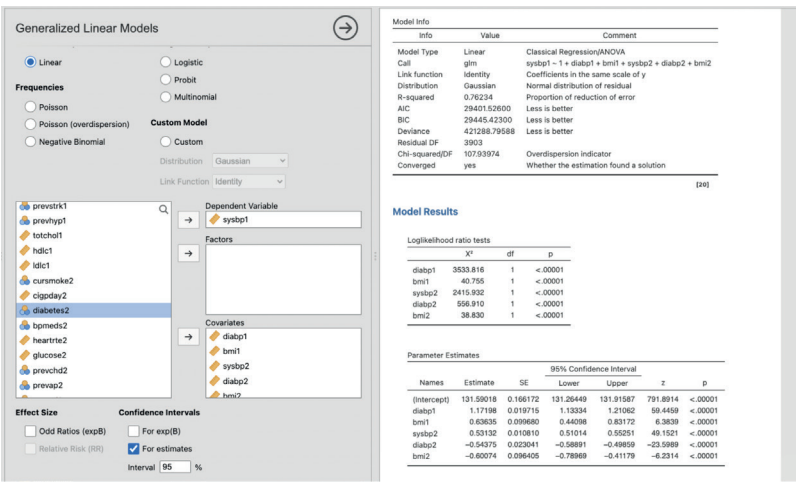


Figura 9. Regresión lineal múltiple con el uso del paquete Lineal Model

Vemos que la información que nos da (figura 9) es mucho más completa, con la significación de las variables, coeficientes e incluso tenemos la opción de añadir los odd ratio y comparar varios modelos a la vez.

Regresión logística

La regresión logística es una técnica de análisis estadístico utilizada para modelar la relación entre una variable dependiente binaria y una o más variables independientes. Es particularmente útil cuando la variable dependiente es categórica y dicotómica, es decir, puede tomar uno de dos valores posibles (como «sí» o «no», «verdadero» o «falso»).

Las características principales de la regresión logística son:

- **Variable dependiente binaria**. Toma valores de 0 o 1.

  Ejemplos:

  Diagnóstico médico (enfermo = 1, sano = 0).

  Resultado de una campaña publicitaria (compra = 1, no compra = 0).

- **Probabilidad como resultado**. La regresión logística estima la probabilidad de que el evento de interés ocurra. La salida del modelo es una probabilidad, que luego puede ser transformada en una predicción binaria usando un umbral (usualmente 0.5).

  Como ejemplo de regresión logística, supongamos que estamos interesados en predecir si un paciente tiene una enfermedad cardíaca (1 = sí, 0 = no) en función de la edad y el nivel de colesterol. Y como aplicaciones de la regresión logística, **Medicina** (predicción de enfermedades, riesgo de complicaciones, etc.), **Marketing** (probabilidad de que un cliente compre un producto), **Finanzas** (predicción de la probabilidad de incumplimiento de crédito) o **Sociología** (estudio de comportamientos y características demográficas). La regresión logística es una herramienta poderosa en estadística y aprendizaje automático para la clasificación binaria, ofreciendo una forma clara e interpretable de entender las relaciones entre las variables y los resultados binarios.

En la siguiente imagen, tenemos las posibilidades de Jamovi para la regresión logística, de forma genérica con la opción de regresión, o con el paquete Linear Models. Aunque hay que decir que tenemos otros paquetes, que no desarrollaremos aquí, que también nos dan esta opción.

Regresión logística binomial

La regresión logística binomial es un método estadístico utilizado para modelar el riesgo de ocurrencia de un evento binario en función de una o más variables independientes. A diferencia de la regresión lineal, que

predice valores continuos, la regresión logística predice el riesgo de que ocurra un evento específico.

Las características que lo definen:

- **Variable dependiente.** Binaria (dos categorías, por ejemplo, 0/1, sí/no).
- **Variables independientes.** Pueden ser cuantitativas o categóricas.
- **Objetivo.** Estimar el riesgo de que ocurra el evento de interés (por ejemplo, la probabilidad de aprobar un examen).
- **Interpretación de coeficientes.** Cambio en el riesgo de que ocurra el evento por una unidad de cambio en la variable independiente, manteniendo constantes las demás variables.

Como hemos visto, los pasos a realizar serán:

1. **Recopilación de datos.** Se debe contar con un conjunto de datos que incluya la variable dependiente binaria y las variables independientes.
2. **Preparación de datos.** Los datos deben estar limpios y preprocesados para eliminar valores atípicos y lidiar con datos faltantes.
3. **Selección de modelo.** Se debe elegir un algoritmo de regresión logística adecuado, como regresión logística de máxima verosimilitud o regresión logística regularizada.
4. **Entrenamiento del modelo.** El modelo se ajusta a los datos de entrenamiento para aprender la relación entre las variables independientes y la variable dependiente.
5. **Evaluación del modelo.** Se evalúa el rendimiento del modelo en un conjunto de datos de prueba independiente para medir su capacidad de generalización.
6. **Interpretación de resultados.** Se interpretan los coeficientes del modelo para comprender la influencia de cada variable independiente en el riesgo de ocurrencia del evento.

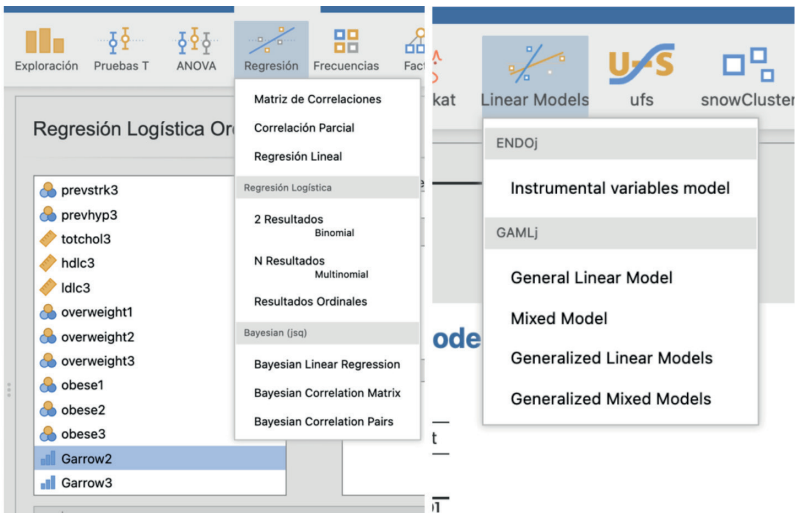


Figura 10. Distintas opciones para realizar regresión logística

Empezaremos con el caso binomial. En la imagen siguiente, podemos ver cómo realizar el análisis de la variable binomial *death* frente a tres variables cualitativas (*sex*, *age_basal_cat* y *obese1*), las cuales se introducen en «Factors» y una variable cuantitativa (*glucose1*), que se introduce en «Covariates» (figura 11).

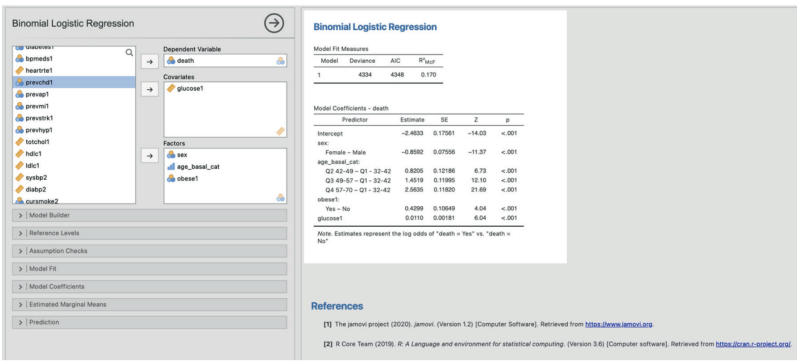


Figura 11. Resultados de regresión logística

Observamos que, de la misma forma que teníamos en la regresión lineal, se presentan los coeficientes y si las variables son o no significativas. Pero, como sabemos, más que los coeficientes, nos interesan los odds ratios, que en Jamovi se expresan como exp(B). Para ello pulsaremos

«Model coefficients» y marcaremos los odds ratio y sus intervalos de confianza (figura 12).

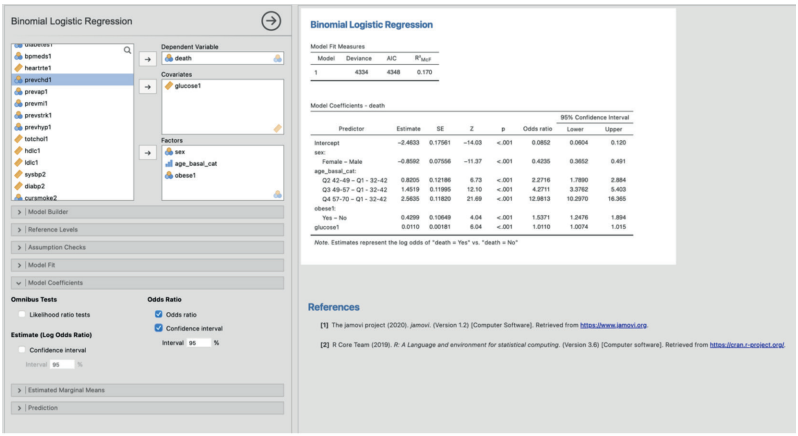


Figura 12. Resultados de regresión logística pidiéndole los odds ratio y sus intervalos de confianza

En este sencillo ejemplo, se interpretaría como sigue. En primer lugar, vemos que todos los coeficientes y los odds ratio, que en Jamovi se expresan como exp(B), son estadísticamente significativos, ya que los valores p son menores que 0,05 (si se trabaja con un nivel de significación del 5 %) y los intervalos de los odds ratio no contienen el valor 1. Interpretaremos a continuación los odds ratio de las variables cualitativas. Las mujeres tuvieron un 57,65 % menos riesgo de morir que los hombres, categoría de referencia

$$(OR - 1) * 100 = (0{,}4235 - 1) * 100? - 57{,}65\,\%$$

Observad que las categorías de referencia Jamovi las coloca a la derecha de cada categoría en las variables cualitativas. Si el sujeto era obeso en la primera medida tuvo un 53,71 % más riesgo de morir que los que no lo eran

$$(1{,}5371 - 1) * 100$$

Observad que a más edad más riesgo de morir, siempre respecto a los sujetos de 32 a 42 años (primer cuartil). Así, los sujetos de 42 a 49 tuvieron un 127,16 % más riesgo de morir que los sujetos de 32 a 42 años; los sujetos de 49 a 57 años un 327,11 % más riesgo; y los de 57 a 70 años un 1198,13 % más riesgo.

Los odds ratio de las variables cuantitativas se interpretan de forma diferente. Así, por cada unidad de la variable *glucosa* el riesgo de morir aumentó un 1,1 %. Recomendamos, no obstante, categorizar todas las variables cuantitativas, por cuanto son más fáciles de interpretar y son más realistas. Por ejemplo, si queremos ver que el riesgo por unidad de glucosa es el mismo con un nivel bajo que con un nivel alto.

Regresión multinomial

Continuamos con el caso en el que la variable dependiente tiene más posibilidades, no solo dos. Tenemos la regresión multinomial, también conocida como regresión logística multinomial. Es una extensión de la regresión logística binaria que se utiliza para modelar la probabilidad de **más de dos categorías** en una variable dependiente. Al igual que la regresión logística binaria, la regresión multinomial predice probabilidades en lugar de valores absolutos.

Sus características principales son:

- **Variable dependiente.** Categórica con **más de dos categorías** (por ejemplo, tipo de transporte: automóvil, autobús, bicicleta).
- **Variables independientes.** Pueden ser cuantitativas o categóricas.
- **Objetivo.** Estimar la probabilidad de que ocurra cada una de las categorías dependientes en función de las variables independientes.
- **Función de salida.** Probabilidades entre 0 y 1 para cada categoría.
- **Interpretación de coeficientes.** Cambio en la probabilidad de una categoría específica por una unidad de cambio en la variable independiente, manteniendo constantes las demás variables.

En qué se parece a la binomial:

- Utiliza la función logística para transformar las probabilidades en valores entre 0 y 1.
- Se basa en el método de máxima verosimilitud para estimar los parámetros del modelo.
- Se interpreta la influencia de las variables independientes a través de los coeficientes.

En qué se diferencia:

- Maneja más de dos categorías en la variable dependiente.

- La función de salida es un vector de probabilidades, una para cada categoría.
- La interpretación de los coeficientes se refiere a cambios en la probabilidad relativa entre categorías, no a cambios absolutos en probabilidades.

Los pasos para realizarla:

1. **Recopilación de datos.** Se debe contar con un conjunto de datos que incluya la variable dependiente categórica con más de dos categorías y las variables independientes.
2. **Preparación de datos.** Los datos deben estar limpios y preprocesados para eliminar valores atípicos y lidiar con datos faltantes.
3. **Selección de modelo.** Se debe elegir un algoritmo de regresión multinomial adecuado, como regresión multinomial de máxima verosimilitud o regresión multinomial regularizada.
4. **Entrenamiento del modelo.** El modelo se ajusta a los datos de entrenamiento para aprender la relación entre las variables independientes y la variable dependiente categórica.
5. **Evaluación del modelo.** Se evalúa el rendimiento del modelo en un conjunto de datos de prueba independiente para medir su capacidad de generalización.
6. **Interpretación de resultados.** Se interpretan los coeficientes del modelo para comprender la influencia de cada variable independiente en la probabilidad de cada categoría dependiente.

Ahora, podemos ver la tabla de los resultados si utilizamos de forma general la regresión multinomial y, con posterioridad, las imágenes (figuras 13 y 14) de los resultados con el paquete Lineal Model.

Tabla 4. Resultados de regresión multinomial

Medidas de ajuste del modelo

Modelo	Desvianza	AIC	R^2_{McF}
1	8215.5	8227.5	0.059829

Coeficientes del Modelo - Garrow1

Garrow1	Predictor	Estimador	EE	Z	p
Overweight - Normoweight	Constante	–4.2302921	0.2667635	–15.8578	<.00001
	sysbp1	0.0059530	0.0024725	2.4077	0.01605
	diabp1	0.0412064	0.0046931	8.7803	<.00001

Coeficientes del Modelo - Garrow1

Garrow1	Predictor	Estimador	EE	Z	p
Obese - Normoweight	Constante	–8.5380157	0.3760676	–22.7034	< .00001
	sysbp1	0.0076365	0.0034580	2.2084	0.02722
	diabp1	0.0743786	0.0066900	11.1179	< .00001

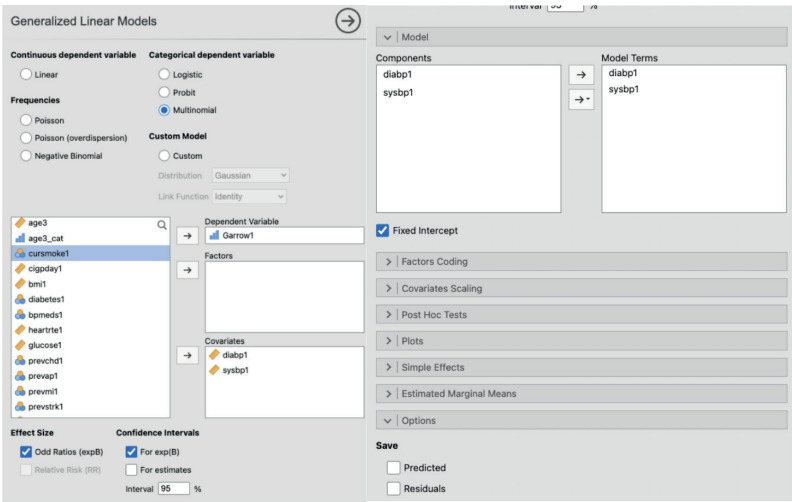


Figura 13. Posibilidades de Lineal Models para regresión múltiple

Generalized Linear Model

Model Info

Info	Value	Comment
Model Type	Multinomial	Model for categorical y
Call	multinom	Garrow1 ~ 1 + diabp1 + sysbp1
Link function	Logit	Log of the odd of each level of y over y=0
Direction	P(y=x)/P(x=0)	P(Garrow1=Overweight)/P(Garrow1=Normoweight) , P(Garrow1=Obese)/P(Garrow1=Normoweight)
Distribution	Multinomial	Multi-event distribution of y
R-squared	0.059829	Proportion of reduction of error
AIC	8227.460000	Less is better
BIC	8265.816000	Less is better
Deviance	8215.459670	Less is better
Residual DF	6.000000	
Chi-squared/DF	.	Overdispersion indicator
Converged	yes	Whether the estimation found a solution

[20]

Model Results

Loglikelihood ratio tests

	X^2	df	p
diabp1	151.0451	2	<.00001
sysbp1	7.4806	2	0.02375

Parameter Estimates

					95% Exp(B) Confidence Interval			
Response Contrasts	Names	Estimate	SE	exp(B)	Lower	Upper	z	p
Obese - Normoweight	(Intercept)	−1.3434800	0.0535847	0.26094	0.92135	1.05175	−25.07206	<.00001
	diabp1	0.0743828	0.0066900	1.07722	1.03253	1.05170	11.11844	<.00001
	sysbp1	0.0076345	0.0034580	1.00766	1.00111	1.01086	2.20778	0.02726
Overweight - Normoweight	(Intercept)	−0.0157337	0.0337688	0.98439	0.23492	0.28983	−0.46593	0.64127
	diabp1	0.0412067	0.0046931	1.04207	1.06319	1.09144	8.78033	<.00001
	sysbp1	0.0059527	0.0024725	1.00597	1.00086	1.01452	2.40757	0.01606

Figura 14. Resultados de Lineal Models para regresión múltiple

Regresión ordinal

En algunas ocasiones, aunque el procedimiento y resultado sea similar, podemos hablar de regresión ordinal. Es un método estadístico utilizado para modelar la **relación entre una variable dependiente ordinal y una o más variables independientes**. A diferencia de la regresión lineal y la regresión logística, que se utilizan para variables dependientes continuas o binarias, respectivamente, la regresión ordinal se emplea cuando la variable dependiente presenta un orden natural, pero las diferencias entre las categorías no son necesariamente iguales.

Sus características son:

- **Variable dependiente.** Ordinal (con categorías ordenadas, por ejemplo, nivel de satisfacción: insatisfecho, neutral, satisfecho).
- **Variables independientes.** Pueden ser cuantitativas o categóricas.

- **Objetivo.** Estimar la influencia de las variables independientes en la probabilidad de que un caso se clasifique en cada categoría de la variable dependiente ordinal.
- **Función de salida.** Probabilidades acumuladas para cada categoría.
- **Interpretación de coeficientes.** Cambio en las probabilidades acumuladas por una unidad de cambio en la variable independiente, manteniendo constantes las demás variables.

Para realizarla:

1. **Recopilación de datos.** Se debe contar con un conjunto de datos que incluya la variable dependiente ordinal y las variables independientes.
2. **Preparación de datos.** Los datos deben estar limpios y preprocesados para eliminar valores atípicos y lidiar con datos faltantes.
3. **Selección de modelo.** Se debe elegir un modelo de regresión ordinal adecuado, como el modelo de probabilidad proporcional o el modelo de efectos lineales acumulados.
4. **Entrenamiento del modelo.** El modelo se ajusta a los datos de entrenamiento para aprender la relación entre las variables independientes y la variable dependiente ordinal.
5. **Evaluación del modelo.** Se evalúa el rendimiento del modelo en un conjunto de datos de prueba independiente para medir su capacidad de generalización.
6. **Interpretación de resultados.** Se interpretan los coeficientes del modelo para comprender la influencia de cada variable independiente en la probabilidad de que un caso se clasifique en cada categoría de la variable dependiente ordinal.

Como, por ejemplo, para el caso de *Garrow1* frente a *sysbp* y *diabp* que hemos realizado con anterioridad. Los resultados los podemos ver en la tabla siguiente.

Tabla 5. Resultados de regresión ordinal

Medidas de ajuste del modelo

Modelo	Desvianza	AIC	R^2_{McF}
1	8213.0	8221.0	0.060114

Nota. La variable dependiente 'Garrow1' tiene el siguiente orden: Normoweight | Overweight | Obese

Coeficientes del modelo - Garrow1

Predictor	Estimador	EE	Z	p

sysbp1	0.0057758	0.0021349	2.7054	0.00682
diabp1	0.0487134	0.0040414	12.0537	<.00001

Contrastes no paramétricos

Finalmente, vamos a comentar de forma muy breve los contrastes de regresión no paramétrica, que son pruebas estadísticas utilizadas para comparar modelos o evaluar la significancia de predictores en un contexto no paramétrico. A diferencia de los métodos paramétricos, que asumen una forma específica para la relación entre las variables (por ejemplo, lineal), los métodos no paramétricos no suponen una forma particular para esta relación, permitiendo una mayor flexibilidad al modelar datos complejos y no lineales.

Las características principales de los contrastes de regresión no paramétrica son:

- Flexibilidad. No se asume una forma funcional específica (como lineal o cuadrática) para la relación entre las variables.

  Se utilizan técnicas como el suavizamiento de *kernel*, *splines* o árboles de decisión para modelar la relación.

- Pruebas de hipótesis. Las pruebas no paramétricas suelen ser más robustas ante la violación de supuestos comunes en métodos paramétricos, como la normalidad de los residuos o la homocedasticidad.

Un ejemplo es la prueba de permutación, donde se calculan todas las posibles reordenaciones de los datos para evaluar la significancia del efecto observado.

Las ventajas que tienes son que no requiere suposiciones estrictas sobre la forma funcional de la relación, suele ser más robusto ante *outliers* y datos no lineales. Sus desventajas son que puede ser computacionalmente intensivo, a veces menos interpretable que los métodos paramétricos.

En resumen, los contrastes de regresión no paramétrica proporcionan una forma flexible y robusta de evaluar la significancia de predictores en modelos que no imponen suposiciones estrictas sobre la relación entre las variables, siendo especialmente útiles en situaciones donde las relaciones son complejas o no lineales.

Para realizar un ejemplo, podemos ver el caso de ANOVA de un factor no paramétrico (figura 15).

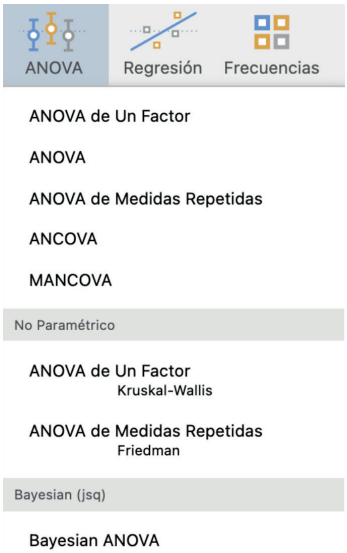


Figura 15. Posibilidades de ANOVA

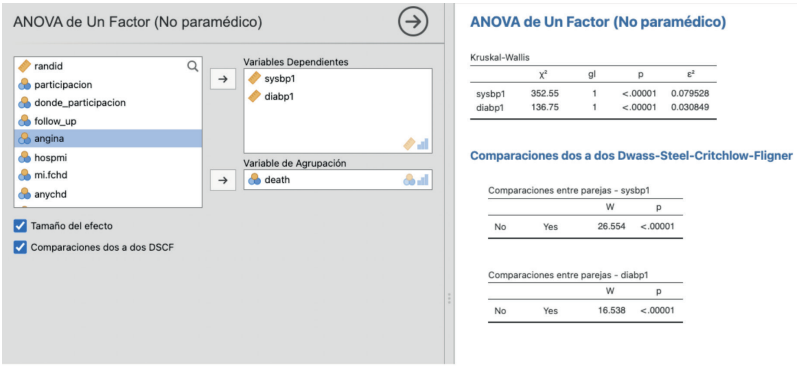


Figura 16. Resultados de ANOVA no paramétrico

Referencias

The Jamovi project (2022). *Jamovi*. (Versión 2.3) [Computer Software]. Recuperado de: https://www.JAMOVI.org

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

Linlin Yan (2020). *Venn Diagram by ggplot2, with really easy-to-use API*. [R package]. Recuperado de https://github.com/yanlinlin82/ggvenn

Serdar Balci (2022). *ClinicoPath Jamovi Module doi: 10.5281/zenodo.3997188*. [R package]. Recuperado de: https://github.com/sbalci/ClinicoPathJAMOVIModule https://www.serdarbalci.com/ClinicoPathJamoviModule/

Bjoern Koneswarakantha (2019). *easyalluvial: Generate Alluvial Plots with a Single Line of Code.*. [R package]. Recuperado de: https://CRAN.R-project.org/package=easyalluvial.

Nick Barrowman (2020). *vtree: Display Information About Nested Subsets of a Data Frame*. [R package]. Recuperado de: https://CRAN.R-project.org/package=vtree

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., y RStudio (2018). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. [R package]. Recuperado de: https://CRAN.R-project.org/package=ggplot2.

Patil, I. (2018). *ggstatsplot: 'ggplot2' Based Plots with Statistical Details*. [R package]. Recuperado de: https://CRAN.R-project.org/package=ggstatsplot

Edwards, A. W. F. (2001). *Occam's Bonus*. Cambridge University Press. https://books.google.com.sa/books?hl=en&lr=&id=-YdbBN-O-JAC&oi=fnd&pg=PA128&dq=edwards+occams+bonus&ots=xqofxHovyX&sig=KYhC8R1_rf2Ky-0ZNWu8WgYnxuY&redir_esc=y

Glover, S. (2018). *Likelihood Ratios; A Tutorial* https://osf.io/preprints/metaarxiv/g3j2k/

Cahusac, P.M.B. (2020). *Evidence-Based Statistics*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119549833

Revelle,W (2019). *psych: Procedures for Personality and Psychological Research*. [R package]. Recuperado de: https://CRAN.R-project.org/package=psych

Epskamp,S., et al. (2012). *qgraph: Network Visualizations of Relationships in Psychometric Data*. [R package]. Recuperado de: https://CRAN.Rproject.org/package=qgraph

Seol, H. (2023). *seolmatrix: Correlations suite for Jamovi*. (Versión 3.7.1) [Jamovi module]. https://github.com/hyunsooseol/seolmatrix

Ripley, B. y Venables, W. (2016). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. [R package]. Recuperado de: https://cran.r-project.org/package=nnet

JASP Team (2018). *JASP*. [Computer software]. Recuperado de: https://jasp-stats.org.

Clyde, M. A. (2017). *BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging*. [R package]. Recuperado de: https://cran.r-project.org/package=BAS

Clyde, M. A., Ghosh, J. y Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics, 20*, 80-101.

Ripley, B., Venables, W., Bates, D. M., Hornik, K., Gebhardt, A., y Firth, D. (2018). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. [R package]. Recuperado de: https://cran.r-project.org/package=MASS

Gallucci, M. (2019). *GAMLj: General analyses for linear models*. [Jamovi module]. Recuperado de: https://gamlj.github.io/

Este manual supone una buena introducción al trabajo y análisis estadístico y a su comprensión por parte de profesores, investigadores y especialmente del alumnado, que necesita ayuda en el uso de herramientas estadísticas de forma sencilla y clara. Más allá de la misma descriptiva, Jamovi proporciona una experiencia de aprendizaje accesible y enriquecedora para usuarios de todos los niveles. Su enfoque intuitivo y sus capacidades avanzadas lo convierten en un aliado indispensable para el análisis de datos en el ámbito académico y profesional. Todo esto nos lo da Jamovi.

Documenta
Universitaria

@DocUniv
documentauniversitaria.com