

# MANUAL JAMOVİ

Aplicacions a  
dades de salut

Pablo Juan Verdoy  
Marc Saez Zafra  
Xavier Perafita Basart  
Laura Serra Saurina





**Manual Jamovi. Aplicacions a dades de salut**



# **Manual Jamovi**

## Aplicacions a dades de salut

Pablo Juan Verdoy  
Marc Saez Zafra  
Xavier Perafita Basart  
Laura Serra Saurina

Dades CIP recomanades per la Biblioteca de la UdG

CIP 519.2 JUA

Juan Verdoy, Pablo, autor

Manual Jamovi : aplicacions a dades de salut / Pablo Juan

Verdoy, Marc Saez Zafra, Xavier Perafita Basart, Laura Serra

Saurina. – Girona : Universitat de Girona, desembre de 2025.

– 1 recurs en línia (174 pàgines) : il·lustracions, gràfics, taules

ISBN 978-84-8458-748-4

I. Saez Zafra, Marc, autor II. Perafita Basart, Xavier, autor

III. Serra, Laura, autor 1. Estadística mèdica – Programari

2. Jamovi (Programari) 3. Llibres electrònics

CIP 519.2 JUA

Amb el suport de:



Generalitat de Catalunya  
**Departament de Recerca  
i Universitats**

Revisió lingüística: Documenta Universitaria

Disseny i maquetació: Documenta Universitaria

© del text, de les figures i de les taules: Pablo Juan Verdoy,

Marc Saez Zafra, Xavier Perafita Basart,

Laura Serra Saurina

© de l'edició: Universitat de Girona - Edicions UdG

DOI: 10.33115/b/9788484587484

ISBN: 978-84-8458-748-4

Girona, desembre 2025



Els textos i les imatges publicats en aquesta obra estan subjectes –  
excepte que s'indiqui el contrari– a una llicència Creative Commons de  
tipus Reconeixement-NoComercial-SenseObraDerivada (BY-NC-ND)  
v.4.0. Es pot copiar, distribuir i transmetre l'obra públicament sempre  
que se citi l'autor i la font, i sempre que no se'n faci un ús comercial ni  
cap obra derivada.

La llicència completa es pot consultar a:

<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.ca>

**Universitat  
de Girona**

# Índex

|                                                     |    |
|-----------------------------------------------------|----|
| Pròleg.....                                         | 9  |
| Capítol 1. Jamovi, instal·lació i mòduls .....      | 11 |
| Introducció .....                                   | 11 |
| Instal·lació.....                                   | 13 |
| Mòduls .....                                        | 18 |
| Mòdul jmv .....                                     | 20 |
| Mòdul jmvBase.....                                  | 22 |
| Referències.....                                    | 35 |
| Capítol 2. JAMОВI Cloud .....                       | 37 |
| Introducció .....                                   | 37 |
| Mòduls .....                                        | 42 |
| Capítol 3. Tractament de les dades .....            | 51 |
| Conceptes bàsics d'estadística.....                 | 51 |
| Explicació de les dades d'exemple: Framingham ..... | 52 |
| Tipus de fitxers.....                               | 56 |
| Tipus de variables.....                             | 58 |
| Tractament de dades.....                            | 61 |
| Exemple de tractament de dades i variables .....    | 65 |
| Referències.....                                    | 66 |
| Capítol 4. Descriptiva d'una variable .....         | 67 |
| Variables quantitatives .....                       | 67 |
| Variables qualitatives .....                        | 80 |
| Referències.....                                    | 85 |

|                                                                                                  |            |
|--------------------------------------------------------------------------------------------------|------------|
| <b>Capítol 5. Descriptiva de dues o més variables .....</b>                                      | <b>87</b>  |
| Altres tractaments gràfics.....                                                                  | 90         |
| Referències.....                                                                                 | 94         |
| <b>Capítol 6. Introducció a la inferència amb una variable .....</b>                             | <b>97</b>  |
| Conceptes bàsics: població, mostra, estimació i contrastos.....                                  | 97         |
| Estadística descriptiva.....                                                                     | 102        |
| Construcció d'un interval de confiança per a una proporció.....                                  | 104        |
| Construcció d'un interval de confiança per a la mitjana.....                                     | 106        |
| Contrastos d'hipòtesis.....                                                                      | 110        |
| Contrast d'una mitjana .....                                                                     | 117        |
| Referències.....                                                                                 | 121        |
| <b>Capítol 7. Introducció a la inferència de relació entre dues variables...123</b>              |            |
| Comparació de proporcions. Taules de contingència .....                                          | 127        |
| Comparació de mitjanes.....                                                                      | 131        |
| Referències.....                                                                                 | 140        |
| <b>Capítol 8. Mesures utilitzades en epidemiologia.....141</b>                                   |            |
| Quocients.....                                                                                   | 141        |
| Mesures de freqüència .....                                                                      | 142        |
| Mesures d'associació.....                                                                        | 144        |
| Mesures d'impacte .....                                                                          | 146        |
| Aplicació amb la base de dades de Framingham.....                                                | 150        |
| Referències.....                                                                                 | 154        |
| <b>Capítol 9. Introducció a la regressió .....</b>                                               | <b>155</b> |
| Regressió lineal.....                                                                            | 155        |
| Referències.....                                                                                 | 164        |
| <b>Capítol 10. Introducció a les regressions<br/>amb variables dependents qualitatives .....</b> | <b>165</b> |
| Regressió logística .....                                                                        | 165        |
| Regressió multinomial.....                                                                       | 168        |
| Regressió ordinal .....                                                                          | 171        |
| Contrastos no paramètrics.....                                                                   | 172        |
| Referències.....                                                                                 | 174        |

## Pròleg

Jamovi és un *software* d'anàlisi de dades gratuït i de codi obert que ofereix una interfície gràfica intuïtiva per dur a terme anàlisis estadístiques. És una excel·lent alternativa a programes com SPSS o R, especialment per a aquells que no tenen experiència en programació. Pel que fa a la relació entre Jamovi i R, per exemple, malgrat que ambdós programes d'anàlisi de dades són gratuïts i de codi obert, hi ha algunes diferències clau entre ells: Jamovi té una interfície gràfica intuïtiva similar a un full de càlcul, ideal per als principiants i, a més, permet un aprenentatge més suau i fàcil, ideal per als qui no tenen experiència en programació.

Jamovi és ideal per a principiants que busquen una eina fàcil d'usar per efectuar anàlisis bàsiques; fins i tot, per a usuaris experimentats que necessiten més flexibilitat i potència per dur a terme anàlisis complexes. Les característiques principals de Jamovi són la seva interfície intuïtiva, la seva àmplia gamma d'anàlisi, el codi obert i les seves funcions addicionals. A més, l'ús de Jamovi és recomanable tant per a estudiants com per a investigadors o professionals. En el capítol 1, **Jamovi, instal·lació i mòduls**, es desenvolupa la interfície i els mòduls necessaris. En el capítol 2, es presenta **Jamovi Cloud**, per a qui vol utilitzar el programari a la xarxa. En el capítol 3, **Tractament de dades**, podem veure, a partir d'un conjunt de dades presentat com a exemple, els conceptes bàsics d'estadística i els tipus de variables. En el capítol 4, **Descriptiva d'una variable**, com el seu nom indica, es mostren tots els passos i les possibilitats de Jamovi quan volem conèixer, treballar o mostrar una variable de forma descriptiva. El capítol 5, **Descriptiva de dues o més variables**, com a continuació de l'anterior, es complementa amb les eines de l'anàlisi de dues o més variables simultàniament. De forma més avançada en recerca estadística, els capítols següents, el capítol 6, **Introducció a la inferència amb una variable**, i el capítol 7, **Introducció a la inferència de relació entre dues variables**, ajuden a l'ús de Jamovi per a investigacions més complexes que la mateixa descriptiva. Després d'aquests dos capítols, es presenta el capítol 8, **Mesures més utilitzades en l'epidemiologia**, seguit del capítol 9, **Introducció a la regressió**, i un darrer, el desè, **Introducció a les**

**regressions amb variables dependents qualitatives**, materials necessaris i complementaris dels capítols anteriors més avançats.

Aquest manual suposa una bona introducció al treball i anàlisi estadística i a la seva comprensió per part de professors, investigadors i, especialment, de l'alumnat, que necessita ajuda en l'ús d'eines estadístiques de forma senzilla i clara. Més enllà de la mateixa descriptiva, Jamovi proporciona una experiència d'aprenentatge accessible i enriquidora per a usuaris de tots els nivells. El seu enfocament intuïtiu i les seves capacitats avançades el converteixen en un aliat indispensable per a l'anàlisi de dades en l'àmbit acadèmic i professional. Tot això ho proporciona Jamovi.

# Capítol 1

## Jamovi, instal·lació i mòduls

### Introducció

Jamovi és un *software* d'anàlisi de dades gratuït i de codi obert que ofereix una interfície gràfica intuïtiva per dur a terme anàlisis estadístiques. És una excel·lent alternativa a programes com SPSS o R, o més complexos com Python, Stata o Epidat, especialment per a aquells que no tenen experiència en programació.

Pel que fa a la relació entre **Jamovi i R** —ambdós programes d'anàlisi de dades gratuïts i de codi obert—, podem destacar algunes diferències clau entre ells.

#### Interfície

- **Jamovi:** interfície gràfica intuïtiva similar a un full de càlcul, ideal per a principiants.
- **R:** interfície basada en codi, requereix coneixements de programació.

#### Flexibilitat

- **Jamovi:** menys flexible que R, amb un conjunt d'anàlisis predefinides.
- **R:** més flexible i potent, permet dur a terme anàlisis més complexes i personalitzades.

#### Facilitat d'aprenentatge

- **Jamovi:** corba d'aprenentatge més suau, ideal per als qui no tenen experiència en programació.
- **R:** corba d'aprenentatge més pronunciada, requereix temps i esforç per dominar el llenguatge.

Mentre Jamovi és ideal per a principiants que busquen una eina fàcil d'usar per fer anàlisis bàsiques, R és ideal per a usuaris experimentats que necessiten més flexibilitat i potència per dur a terme anàlisis complexes.

Es pot executar el codi R a Jamovi amb l'ús de mòduls, en concret el mòdul Rj Editor. Una altra opció és el paquet jmv, que permet fer les anàlisis de Jamovi des d'una sessió interactiva d'R.

Si ens centrem en les característiques principals de Jamovi:

- **Interfície intuïtiva.** S'assembla a un full de càlcul, la qual cosa facilita el seu ús per a principiants.
- **Àmplia gamma d'anàlisis.** Ofereix una àmplia gamma d'anàlisis estadístiques, des de proves t fins a regressió lineal i anàlisi de variància.
- **Codi obert.** Això significa que és gratuït, transparent i pot ser modificat per la comunitat.
- **Funcions addicionals.** Funcions com la creació de gràfics personalitzats, la gestió de dades i l'exportació de resultats.

Qui pot utilitzar Jamovi?

- **Estudiants.** Eina ideal per a estudiants que estan aprenent estadística per primera vegada.
- **Investigadors.** Pot ser utilitzat per investigadors per fer anàlisis estadístiques complexes.
- **Professionals.** Pot ser utilitzat per professionals en diverses àrees com la psicologia, la sociologia, l'economia i l'educació.

Després d'aquesta petita introducció, començarem amb Jamovi.

Per començar amb Jamovi, pots seguir aquests passos:

1. **Descarrega i instal·la Jamovi** des del lloc web oficial.  
<https://www.JAMОВI.org/>
2. **Obre Jamovi** i familiaritza't amb la interfície.
3. **Consulta la documentació** de Jamovi per aprendre a dur a terme anàlisis específiques.  
<https://www.JAMОВI.org/user-manual.html>
4. **Visita la comunitat de Jamovi** per obtenir ajuda i consells.  
<https://forum.JAMОВI.org>

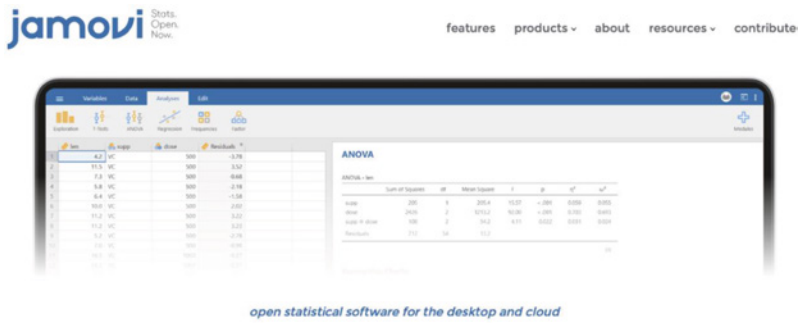


Figura 1. Web de Jamovi

A continuació, es desenvolupa l'apartat de la instal·lació pròpiament dita.

## Instal·lació

Jamovi permet el seu ús en dos formats, *online* o *desktop*, com podem veure a la seva web.

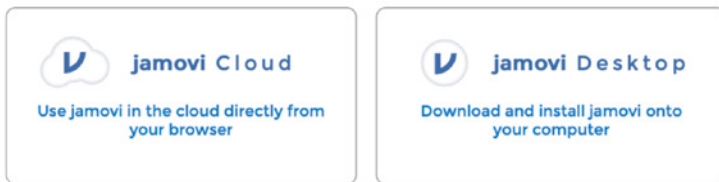


Figura 2. Web de Jamovi

En l'opció Cloud, podem veure l'opció de pagament, mentre que, a Desktop, tenim les diferents versions per descarregar el programa per als diferents sistemes:

- **Windows:** descarrega l'instal·lador .exe
- **Mac:** descarrega l'arxiu .dmg
- **Linux:** descarrega l'arxiu .AppImage

A continuació, executa l'instal·lador, però tingues en compte:

- Assegura't de tenir prou espai lliure en disc abans d'instal·lar Jamovi.
- Si tens problemes per instal·lar Jamovi, consulta la documentació oficial: [https://docs.JAMОВI.org/\\_pages/um\\_1\\_installation.html](https://docs.JAMОВI.org/_pages/um_1_installation.html)
- També pots trobar ajuda a la comunitat de Jamovi: <https://www.JAMОВI.org/library.html>

Els possibles problemes que poden sorgir a la instal·lació de Jamovi.

### Problemes de descàrrega

- **Error de connexió a internet:** assegura't de tenir una connexió a internet estable.
- **Servidor ocupat:** intenta descarregar Jamovi en un altre moment.
- **Arxiu malmès:** descarrega l'arxiu d'instal·lació de nou.

### Problemes d'instal·lació

- **No es compleix amb els requisits del sistema:** verifica que el teu equip compleixi amb els requisits mínims per instal·lar Jamovi.
- **Arxius d'instal·lació corruptes:** descarrega l'arxiu d'instal·lació de nou.
- **Permisos d'administrador:** assegura't de tenir permisos d'administrador per instal·lar Jamovi.
- **Programari antivirus:** desactiva temporalment el *software* antivirus mentre instal·les Jamovi.

### Problemes en iniciar Jamovi

- **Arxius de l'aplicació danyats:** reinstal·la Jamovi.
- **Versions incompatibles de Java:** assegura't de tenir l'última versió de Java instal·lada.
- **Problemes amb la configuració del sistema:** consulta la documentació oficial de Jamovi per obtenir més informació.

## Altres problemes

- **Idioma:** Jamovi està disponible en diversos idiomes. Assegura't de descarregar la versió en l'idioma que desitges fer servir.
- **Actualitzacions:** Jamovi s'actualitza amb freqüència. Assegura't de tenir l'última versió instal·lada.

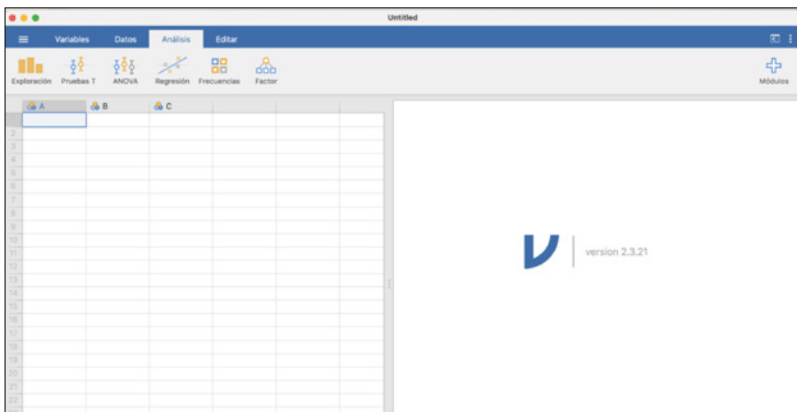


Figura 3. Jamovi en obrir el programa

Abans de veure els diferents mòduls que s'afegiran a Jamovi per al treball desenvolupat en aquest manual per a salut, veurem les diferents parts del *software*.

A la part superior tenim el menú d'opcions, on primerament tenim el menú d'arxius.

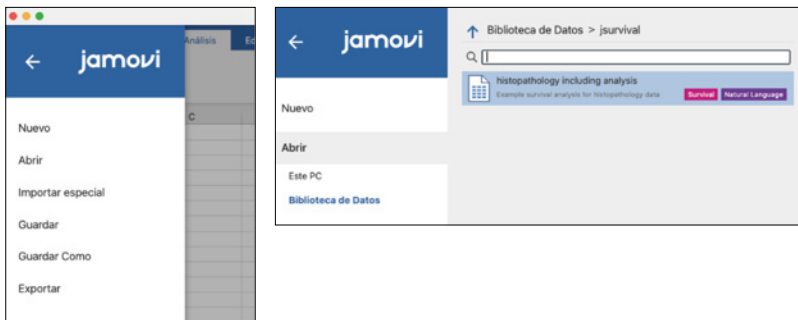
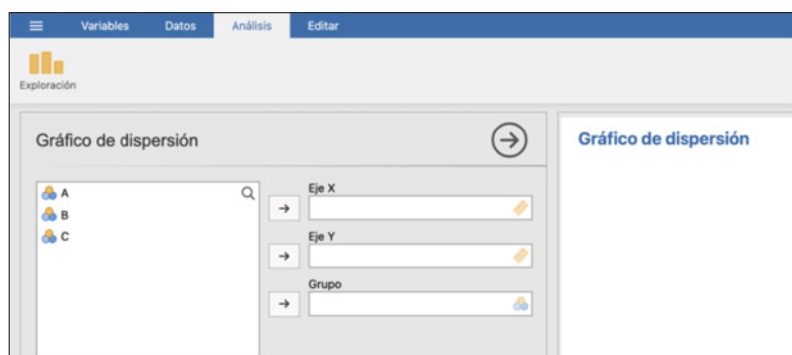
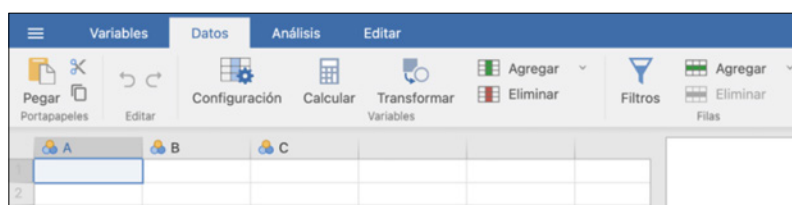


Figura 4. Menú d'arxius

En aquesta part, que desenvoluparem més endavant, podrem veure tot el tractament de dades.

També tenim la gestió de dades, que desenvoluparem en la part de transformació d'aquestes, a més de la finestra d'anàlisi i edició.



**Figura 5. Menús de gestió de dades, anàlisi i edició**

Finalment, abans d'entrar en el desenvolupament dels mòduls, que afectaran les diferents finestres, comentarem el menú de configuració de Jamovi que apareix en el desenvolupament dels tres punts verticals de la part dreta.

Com es pot observar, en aquest menú es podran canviar les configuracions i visualització d'idiomes, dades numèriques, etc. El mateix lector pot desenvolupar aquests punts amb proves concretes.

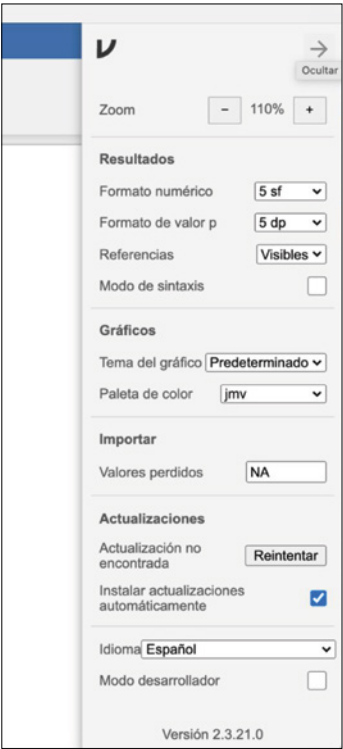


Figura 6. Menú de configuració

## Mòduls

Els mòduls a Jamovi són extensions que es poden instal·lar per ampliar les funcionalitats del programa. Aquests mòduls et permeten fer anàlisis estadístiques més especialitzades que no són possibles amb les eines bàsiques de Jamovi.

### On trobar-los?

Hi ha dues maneres de trobar mòduls:

- **Biblioteca Jamovi.** Hi accedeix des del menú «Mòduls». Aquí trobaràs una llista de mòduls preinstal·lats que pots instal·lar amb un sol clic.
- **Repositoris externs.** Pots trobar mòduls addicionals en repositoris com <https://github.com/> o <https://cran.r-project.org/>

### On trobar-los?

Instal·lar mòduls a Jamovi és un procés bastant senzill que es pot dur a terme dins de la mateixa aplicació. Els mòduls són extensions que agreguen funcionalitats addicionals a Jamovi, que és una aplicació estadística de codi obert dissenyada per ser fàcil d'usar.

Els passos per seguir:

1. **Obrir Jamovi.** Inicia l'aplicació Jamovi en la teva computadora. Si no la tens instal·lada, pots descarregar-la des del seu lloc web oficial <https://www.JAMOV.I.org>
2. **Accés a la biblioteca de mòduls.** Un cop obert Jamovi, busca la icona de la «botiga de mòduls» a la barra de menú principal. La icona sol assemblar-se a una caixa oberta o a una bossa de compres.
3. **Explorar mòduls disponibles.** En fer clic en aquesta icona, s'obrirà la biblioteca de mòduls de Jamovi. Aquí podràs explorar els diferents mòduls disponibles per a la seva instal·lació. Pots navegar per les diferents categories o usar la barra de recerca per trobar un mòdul específic.
4. **Instal·lar un mòdul.** Quan trobis el mòdul que desitges instal·lar, fes-hi clic per veure més detalls. Després, simplement fes clic al botó «Instal·lar» o «Agrega» per iniciar la instal·lació del mòdul.

5. **Esperar la instal·lació.** El mòdul començarà a descarregar-se i instal·lar-se automàticament. Aquest procés pot trigar uns minuts, depenent de la mida del mòdul i de la velocitat de la teva connexió a Internet.
6. **Tancar i reobrir Jamovi (si cal).** Alguns mòduls poden requerir que reiniciïs Jamovi per completar la instal·lació. Si és així, tanca l'aplicació i torna a obrir-la.
7. **Usar el mòdul instal·lat.** Un cop instal·lat el mòdul, aquest hauria d'aparèixer a la barra de menú de Jamovi o al lloc corresponent dins de l'aplicació. Pots fer clic al mòdul per començar a utilitzar les seves funcionalitats addicionals.

Si sorgeix algun problema durant la instal·lació d'un mòdul, com missatges d'error o problemes de compatibilitat, revisa si tens la versió més actualitzada de Jamovi. Els desenvolupadors de mòduls sovint actualitzen les seves extensions per ser compatibles amb les versions més recents de Jamovi, per la qual cosa mantenir la teva aplicació actualitzada pot solucionar molts problemes.

### **Com puc saber quin mòdul necessito?**

Per saber quin mòdul necessites, has de tenir en compte:

- El tipus d'anàlisi que desitges fer.
- El tipus de dades que tens.
- El teu nivell d'experiència en estadística.

Avantatges d'usar mòduls:

- Amplien les capacitats de Jamovi.
- Permeten dur a terme anàlisis més especialitzades.
- Són gratuïts i de codi obert.
- Estan desenvolupats per una comunitat activa d'investigadors.

Desavantatges d'usar mòduls:

- Alguns mòduls poden ser complexos d'usar.
- No tots els mòduls estan ben documentats.
- És possible que alguns mòduls no siguin compatibles amb la versió actual de Jamovi.

A continuació, desenvoluparem de forma més concreta algun dels mòduls que poden ser utilitzats per a l'anàlisi, sobretot en dades de salut.

Per començar a instal·lar-los, podem veure, a la part d'anàlisi, una creu al cantó dret que ens permet tractar els mòduls.



Figura 7. Instal·lació de mòdul

## Mòdul jmv

Aquest mòdul, **jmv** de Jamovi, és un complement per a l'anàlisi de dades **multivariant**. *jmv* ofereix una àmplia gamma d'eines per:

- **Anàlisi de Components Principals (ACP)**. Permet identificar les principals variables que expliquen la variabilitat en un conjunt de dades.
- **Anàlisi Factorial (AF)**. Similar a l'ACP, però amb la capacitat d'identificar factors latents que no són directament observables.
- **Anàlisi Discriminant (AD)**. Permet classificar els individus en diferents grups en funció de les seves característiques.
- **Escalament Multidimensional (MDS)**. Representa les relacions entre les variables en un espai de menor dimensionalitat.
- **Anàlisi de Conglomerats (AC)**. Permet agrupar els individus en funció de les seves característiques.
- **Regressió Múltiple**. Permet modelar la relació entre una variable dependent i un conjunt de variables independents.

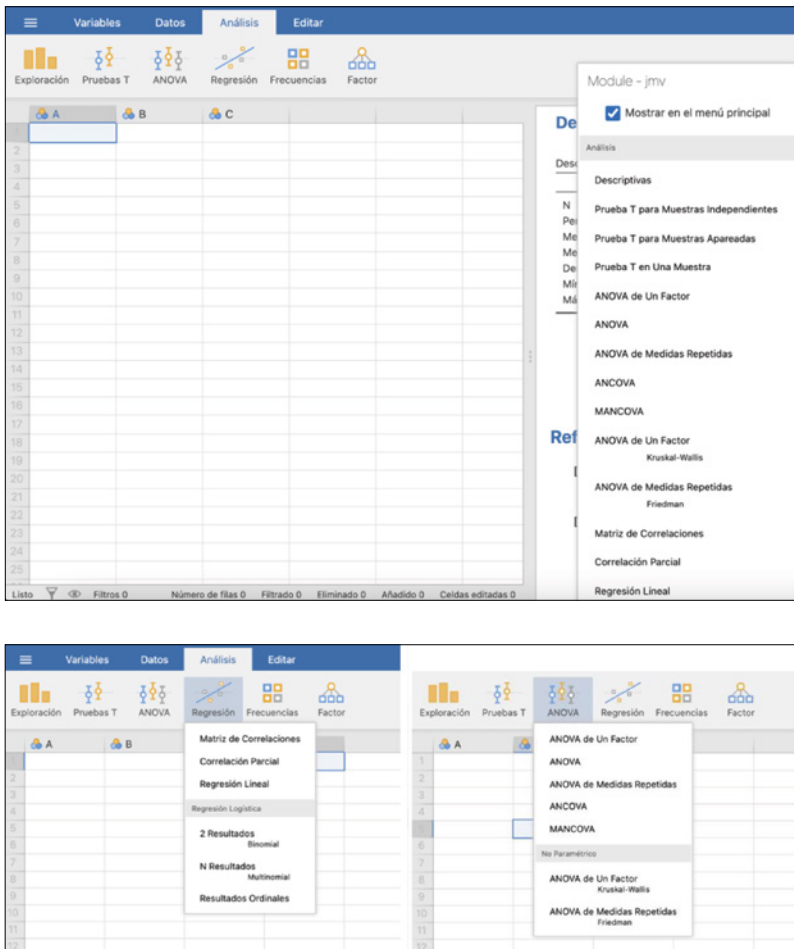


Figura 8. Mòdul jmv

### Funcions clau del mòdul *jmv*

- **Interfície intuïtiva.** *jmv* ofereix una interfície gràfica d'usuari que facilita la selecció de les opcions d'anàlisi i la interpretació dels resultats.
- **Àmplia gamma d'anàlisi.** *jmv* ofereix una àmplia gamma de tècniques d'anàlisi multivariant, la qual cosa el converteix en una eina versàtil per a diferents tipus d'investigacions.

- **Visualització de dades.** *jmv* ofereix diferents eines per visualitzar els resultats de les anàlisis, com gràfics de dispersió, diagrames de *scree plot* i dendrogrames.
- **Exportació de resultats.** *jmv* permet exportar els resultats de les anàlisis a diferents formats, com CSV, SPSS i R.

## Mòdul *jmvBase*

Aquest mòdul, *jmvBase*, és un complement del mòdul anterior que ens permet efectuar, com podem veure en la imatge, anàlisis més completes (ANOVA, contrastos d'independència, etc.).

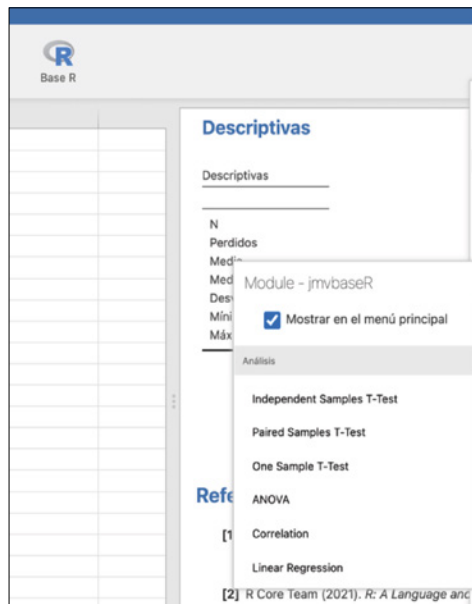


Figura 9. Mòdul *jmvBase*

## Mòdul MAJOR

Un altre dels mòduls és *MAJOR*, que s'utilitza per al desenvolupament de metaanàlisis.

## Mòdul ClinicoPath

El següent mòdul, les opcions del qual podem veure en la imatge següent, ens permet ampliar el tema descriptiu de les nostres dades (a desenvolupar en l'apartat de descripció de variables). Cal notar la gran quantitat de possibilitats que ajudaran en la part de salut.

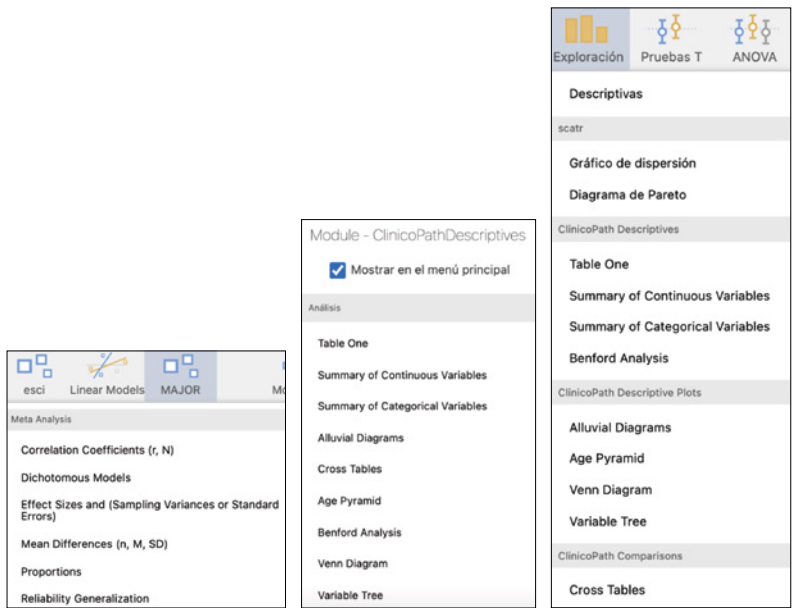


Figura 10. Mòdul major

Figura 11. Mòdul ufs

## Mòdul Statkat

Aquest mòdul ens permet trobar la tècnica estadística adequada per a les dades i preguntes de recerca. Com es mostra a la finestra, ens permet triar tècnica per al cas d'una sola variable o el cas en què hàgim de veure relació entre variables.

## Mòdul ufs

El mòdul *ufs* fa que les funcions del paquet R del mateix nom estiguin disponibles a Jamovi. Aquestes inclouen funcions per calcular intervals de confiança per a mides d'efecte, calcular la mida de mostra requerida per estimar una mida d'efecte amb un interval de confiança d'un ample donat, produir gràfics de diamant, crear una taula de respostes múltiples, i efectuar algunes operacions bàsiques com disminuir o augmentar estimacions de mida d'efecte.

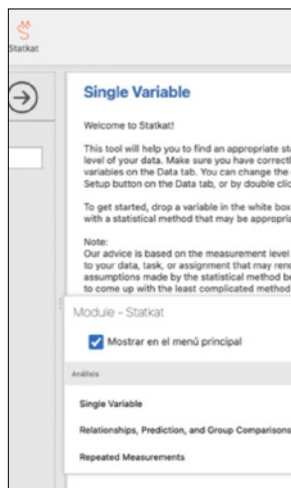


Figura 12. Mòdul Statkat

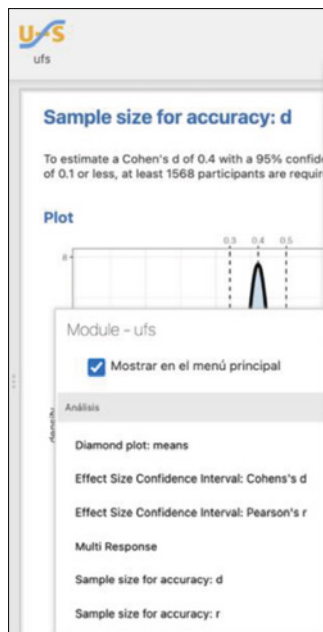
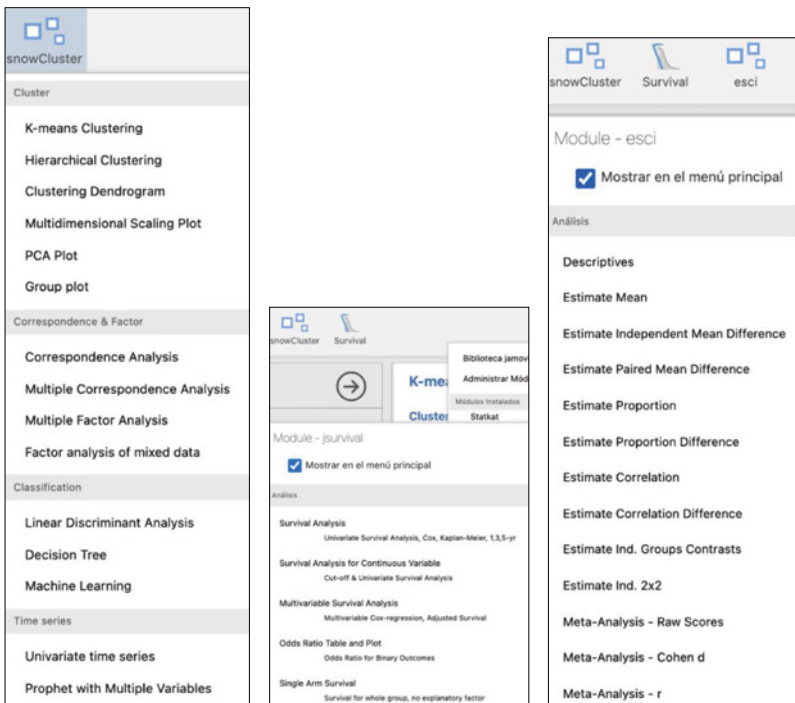


Figura 13. Mòdul ufs

## Mòdul snowcluster, jsurvival i esci

El mòdul *snowcluster*, entre altres funcions, permet als usuaris fer agrupaments K-means, correspondència simple i múltiple, i visualitzar els resultats amb eixos de PCA, mentre que es pot complementar amb el mòdul *jsurvival*, que ens permet fer tota mena d'anàlisis de supervivència.



**Figura 14. Mòdul snowCluster (esquerra), mòdul jsurvival (central) i mòdul esci (dreta)**

*esci* ofereix un pas fàcil cap a les estadístiques d'estimació (també conegudes com les «noves estadístiques»), un enfocament que emfatitza les mides d'efecte, els intervals d'estimació i la metaanàlisi. Aquest mòdul, *esci*, pot proporcionar estimacions i intervals de confiança per a la majoria de les anàlisis que aprendries en un curs d'estadístiques de pregrau i metaanàlisi (la qual cosa realment hauria de ser part d'un bon curs d'estadístiques de pregrau, ja que, encara que el mòdul estigui pensat per a qualsevol usuari, va ser especialment desenvolupat pensant en els estudiants). La majoria de les anàlisis es poden executar a partir

de dades brutes o de dades resumides (cosa que et permet generar estimacions a partir d'articles de revistes que només van informar proves d'hipòtesis). Totes les anàlisis generen visualitzacions agradables que emfatitzen les mides d'efecte i la incertesa. *esci* és per a tothom, però va ser desenvolupat especialment pensant en els estudiants: proporciona instruccions pas a pas, retroalimentació clara i intenta prevenir errors de principiant (com calcular una mitjana en una variable nominal).

## Mòdul *jpower*

Aquest mòdul està desenvolupat a l'article «Power to the people: A beginner's Tutorial to Power Analysis using Jamovi», de James E. Bartlett i Sarah J. Charles, de 2020.

Paquet molt útil per a la presa de decisions referents a la mida de mostra en les dades, més concretament en salut. A més, ens permet triar entre mides de mostra o potència del mostreig (entre altres factors) per a una o més mostres.

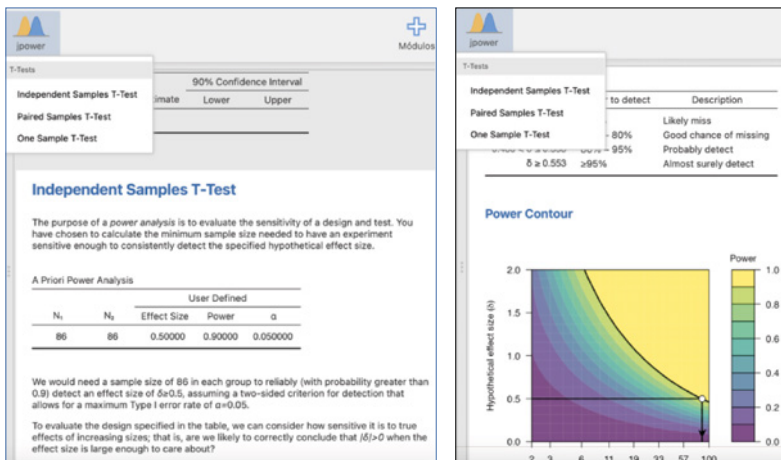


Figura 15. Mòdul *jpower*

## Mòdul jjSTATPLOT i editor de Rj

El mòdul *jjSTATPLOT* serveix per crear gràfics amb detalls de proves estadístiques inclosos en els mateixos gràfics rics en informació. En un flux de treball típic d'anàlisi exploratòria de dades, la visualització de dades i el modelatge estadístic són dues fases diferents: la visualització informa el modelatge, i el modelatge, al seu torn, pot suggerir un mètode de visualització diferent, i així successivament. La idea central de *ggstatsplot* (nom del *package* en base R) és simple: combinar aquestes dues fases en una en forma de gràfics amb detalls estadístics, cosa que facilita i accelera l'exploració de dades.

Per completar la part gràfica i totes les anàlisis que fem a Jamovi, podem utilitzar el mòdul *Rj Editor*, per veure i desenvolupar els codis de R.

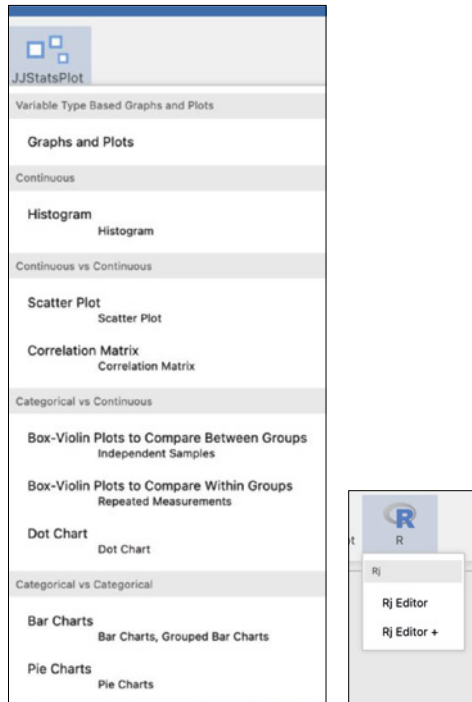


Figura 16. Mòduls jjSTATPLOT i Rj Editor

## Mòduls jsq i seolmatrix

El mòdul *jsq* ens permet dur a terme les diferents anàlisis bayesianes que podem veure en la imatge, mentre que el mòdul *seolmatrix*, per contra, ens serveix per completar altres mòduls en la part de correlació, avaluació o anàlisi jeràrquica.

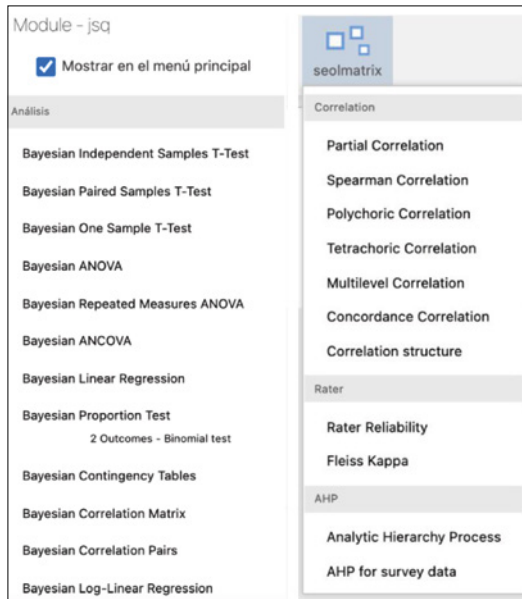


Figura 17. Mòduls jsq.i seolmatrix

## Mòdul distrACTION

El mòdul *distrACTION* ens mostra de forma gràfica la forma i desenvolupament de les diferents distribucions de probabilitat contínues o discretes.

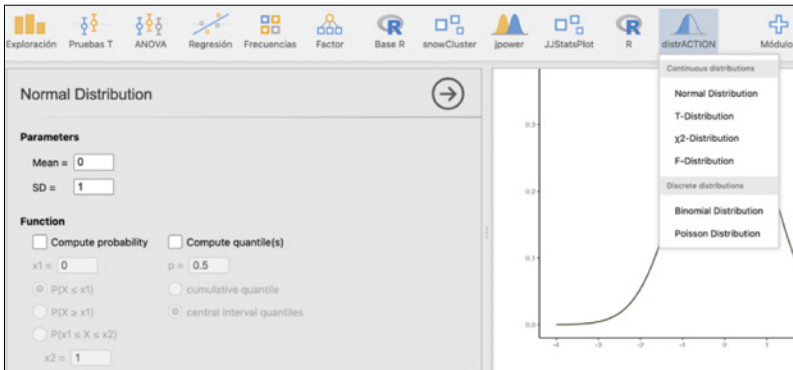


Figura 18. Mòdul distrACTION

## Mòdul Flexplot i mòdul bmttest

Aquests dos mòduls són molt concrets, un per a l'elaboració de models lineals generalitzats (en algunes de les seves possibilitats) i gràfics i, d'altra banda, el mòdul *bmttest*, que només realitza el contrast de Brunner-Munzel, que a més està publicat en la següent adreça: <https://www.mdpi.com/2624-8611/5/2/26>

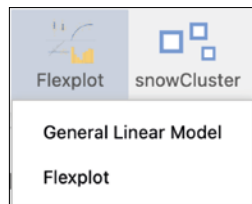


Figura 19. Mòdul Flexplot

## Mòduls Demonstration i el mòdul clt

Altres dos mòduls complementaris per a Jamovi són el que ens mostra demostracions de les dades (*Demonstration*) i el paquet *clt*, indicat per fer anàlisis com correlació o contrast d'hipòtesis.

## Mòdul jeva

Alguna vegada has intentat fer una anàlisi evidencial? És possible que t'hagi resultat difícil trobar una plataforma estadística per a això. Ara hi ha el mòdul jeva de Jamovi, que pot proporcionar raons de versemblança per a una varietat de proves estadístiques comunes.

A més, ens mostra moltes altres possibilitats, com podem veure a l'enllaç: <https://blog.JAMOVI.org/2023/02/22/jeva.html>

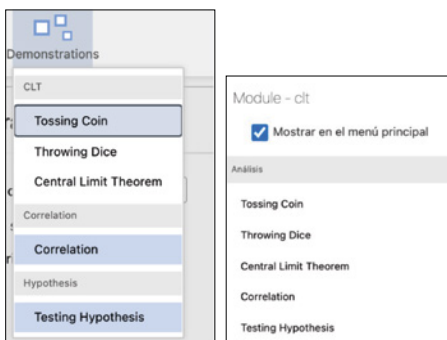


Figura 20. Mòduls Demonstrations i clt

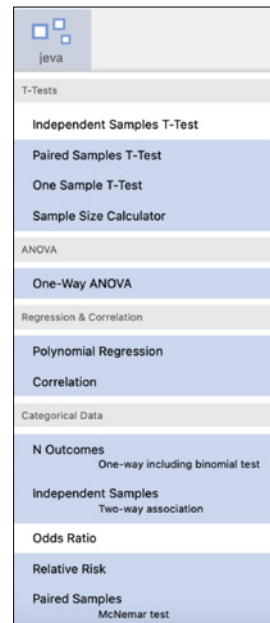


Figura 21. Mòdul jeva

Mòdul ztestvist

Mòdul clar i concís per mostrar gràficament, i amb càlculs, el contrast d'una mostra.

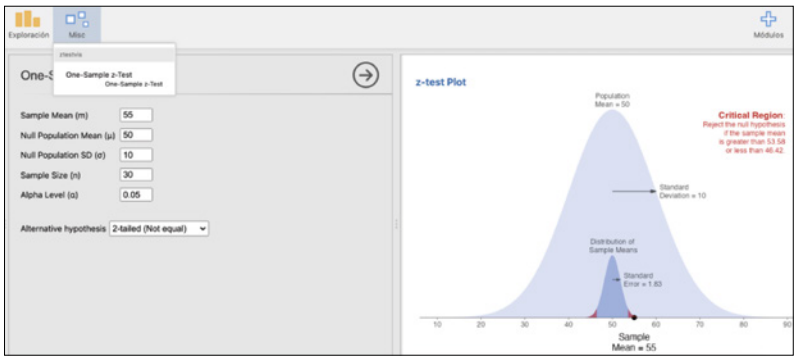


Figura 22. Mòdul ztestvist

Mòdul medul

És un mòdul amb molts usos en l'àmbit de la salut, per al desenvolupament de diferents anàlisis de decisió de les dades, com sensibilitat, intervals de confiança, etc.



Figura 23. Mòdul medul

## Mòduls PPDA i MEDA

Com podem veure en la imatge, el mòdul PPDA ens permet elaborar descripcions automàtiques de models univariants i multivariants. En canvi, el PPDA és molt més concret en aquest tipus d'anàlisi, com, per exemple, per efectuar les corbes ROC.

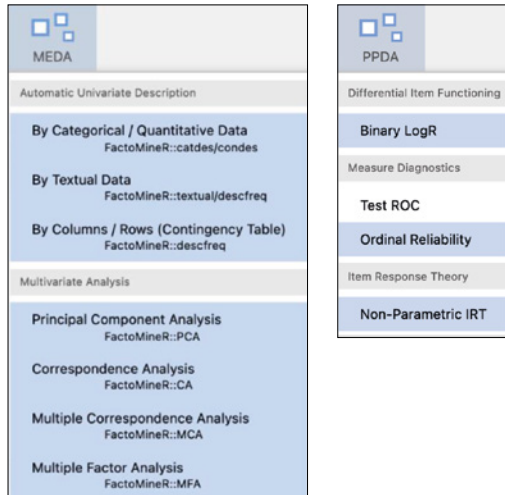


Figura 24. Mòduls PPDA i MEDA

## Mòdul gamlj

El mòdul *GAMlj* ofereix múltiples eines per realitzar tota mena d'anàlisis multivariants.

Tot el seu desenvolupament es pot veure a l'enllaç: <https://gamlj.github.io>

*GAMlj* ofereix eines per estimar, visualitzar i interpretar models lineals generals, models lineals mixtos i models lineals generalitzats amb variables categòriques o contínues, amb opcions per facilitar l'estimació d'interaccions, pendents simples, efectes simples, proves *post hoc*, etc.

- Enfocament d'ANOVA i regressió.
- Variables independents contínues i categòriques.
- Prova F, proves de raó de versemblança i estimacions de paràmetres.

- Intervals de confiança, estàndard, perfil i d'arrencada.
- Moderació facilitada.
- Anàlisi de pendents simples.
- Anàlisi d'efectes simples.
- Anàlisi d'interacció simple.
- Anàlisi *post hoc*.
- Gràfics per a qualsevol ordre d'interaccions.
- Selecció automàtica dels millors mètodes d'estimació i selecció de graus de llibertat.
- Estimació de tipus III.
- Un ampli conjunt d'índexs de mida d'efecte, segons el model que es consideri.

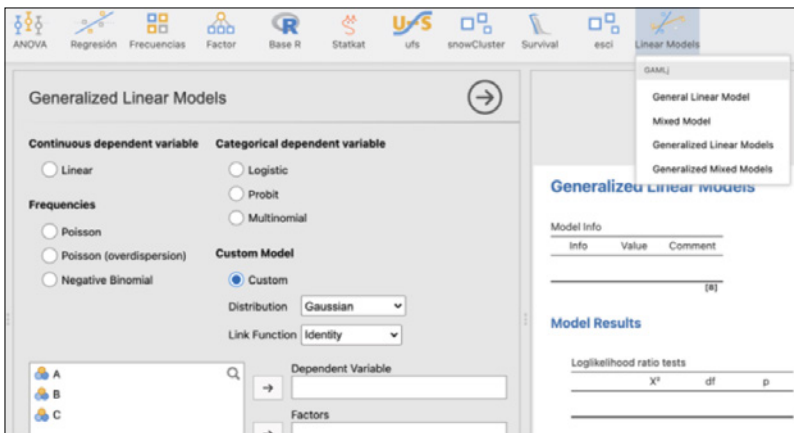


Figura 25. Mòdul GAMLj

Els possibles models:

- 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 binomial ANOVA-like model (GMixed).
- Mixed Ordinal regression (GMixed).
- Mixed Ordinal ANOVA-like model (GMixed).

## Referències

**Introducció a Jamovi, per JD Leongómez:**

<https://jdleongomez.info/es/post/JAMОВI/>

**Guia didàctica: Introducció a l'anàlisi de dades en estudis biomèdics (Jamovi):**

<https://sites.google.com/iacs.es/analisisdatosJAMОВI/inicio>

**Informació sobre com instal·lar i usar els mòduls:**

<https://www.JAMОВI.org/user-manual.html>

**Pàgina web de la biblioteca Jamovi:**

<https://www.JAMОВI.org/library.html>

**Canal de YouTube de RecuEdu:**

[https://www.youtube.com/watch?v=JMslTnk1\\_gU](https://www.youtube.com/watch?v=JMslTnk1_gU)

**Blog de Matt C. Howard:**

<https://mattchoward.com/instalacion-de-mòduls-en-JAMОВI/>



## Capítol 2

# JAMOVİ Cloud

### Introducció

Com s'ha esmentat en el capítol 1, des del web de Jamovi (<https://www.jamovi.org/>) s'ofereix la possibilitat d'utilitzar el programa en dos formats, online i Desktop:

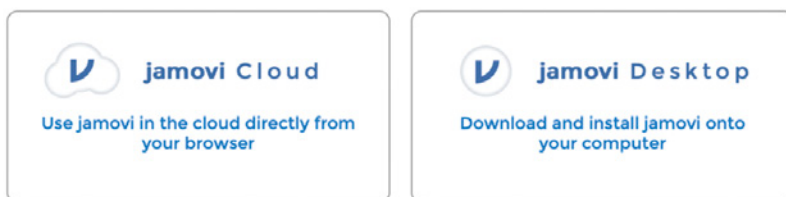


Figura 1. Web de Jamovi

L'opció Cloud permet fer servir Jamovi al núvol directament des del nostre navegador sense necessitat d'instal·lar-nos el programa. Això facilita tenir sempre actualitzada la versió, que hi hagi compatibilitat amb tots els dispositius (incloent-hi iPads i Chromebooks). Així mateix, facilita als usuaris guardar el seu treball a OneDrive (i, pròximament, en altres serveis en línia també) i les anàlisis s'executen en CPUs d'alt rendiment en els nostres centres de dades. També cal tenir en compte que la versió online requereix un suport tècnic mínim.

### Com s'emmagatzemen les dades en l'opció Cloud?

En relació amb l'emmagatzematge de les dades al núvol de Jamovi cal tenir en compte que aquesta versió és probablement més segura que la versió d'escriptori o que el software d'escriptori en general.

Quan les persones obren un arxiu de dades al núvol de Jamovi, sigui pujant-lo sigui accedint des d'una ubicació en línia com OneDrive, l'arxiu es carrega als servidors de Jamovi. Aquest arxiu es transmet utilitzant xifrat de grau industrial, el mateix que es fa servir per protegir comptes bancaris. Les dades romanen en els servidors durant la sessió de l'usuari, en una ubicació segura a la qual altres usuaris no poden accedir. Quan la sessió acaba, ja sigui perquè l'usuari tanca la sessió, tanca totes les seves pestanyes o finestres, o deixa la connexió inactiva per un temps, les dades s'eliminen d'aquests servidors. Si l'usuari ja no utilitza les dades, aquestes són eliminades.

1. A més, tots els serveis al núvol de Jamovi operen en centres de dades ubicats en països amb lleis de privacitat de dades raonables, com Austràlia, Països Baixos, Singapur, Estats Units, Canadà, Alemanya i el Regne Unit.

Algunes institucions insisteixen que totes les dades dels usuaris romanguin en els dispositius individuals, sota la impressió que això és més segur que usar un servei en línia com el núvol de Jamovi. L'argument és que, si el núvol de Jamovi és atacat per hackers, les dades dels usuaris podrien ser extretes i filtrades a internet. El perfil de risc de cada organització és diferent, reflectint el valor de la informació que manegen i els tipus d'actors d'amenaques que enfronten. Tanmateix, en la majoria dels casos, un servei web com el núvol de Jamovi —un servei desenvolupat i mantingut per professionals de software amb un profund coneixement en seguretat— representa un objectiu més difícil que una computadora d'escriptori individual. S'ha demostrat una vegada i una altra que els usuaris finals, subjectes a atacs de phishing o obertura d'arxius adjunts maliciosos, representen un esglaó molt més feble que els serveis en línia que podrien allotjar les seves dades.

De fet, si considerem les dades d'un usuari allotjat a OneDrive i accedits a través del núvol de Jamovi, aquestes dades mai són presents ni són vulnerables en la computadora local de l'usuari.

## Plans disponibles del Cloud

Dit això, hi ha dos plans disponibles: un pla gratuït amb algunes limitacions i un pla prioritari econòmic. Finalment, hi ha una tercera opció per a grups (més de 5 usuaris).

Després d'escollir l'opció Jamovi Cloud, apareix la pantalla següent:



Figura 2. Web de Jamovi

A partir d'aquí, si fas clic al botó «Start» i després al botó «Get Started», se t'obren les opcions que s'ofereixen:

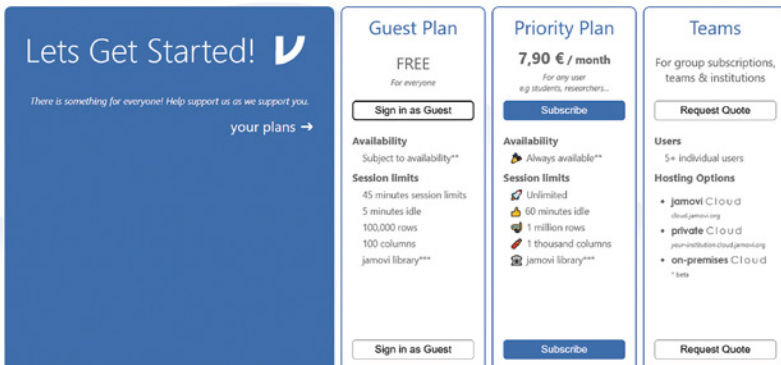


Figura 3. Web de Jamovi

Per a l'opció gratuïta (Guest Plan) cal tenir en compte que:

- Hi ha restriccions de l'ús segons disponibilitat. Si hi ha massa usuaris connectats no podem obrir la versió gratuïta online.
- Permet poder treballar amb el programa un màxim de 45 minuts. Quan la sessió està a punt d'expirar apareix una nova informant sobre els minuts restants.
- La sessió finalitzarà després de 5 minuts d'inactivitat.
- La capacitat que ofereix és de 100.000 files i 100 columnes.
- De moment no s'inclou la llibreria de Rj Editor.

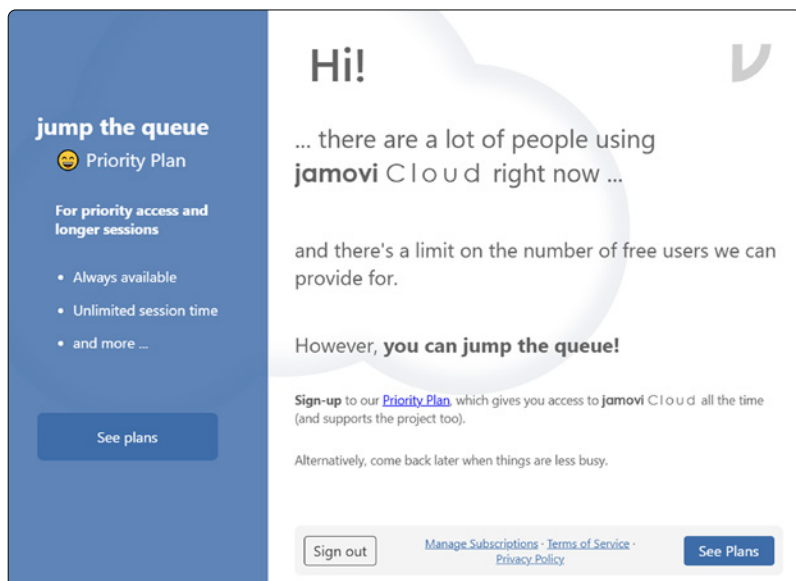


Figura 4. Missatge de Jamovi sobre la no disponibilitat



Figura 5. Missatge de Jamovi sobre la finalització del temps de treball

Per a l'opció de pagament cal tenir en compte:

- No hi ha límit de temps per treballar amb el programa.
- La sessió finalitzarà després de 60 minuts d'inactivitat.
- La capacitat és d'1 milió de files i 1000 columnes.
- Per al moment no s'inclou la llibreria de Rj Editor.

## Versió gratuïta

Seleccionem la versió gratuïta i accedim mitjançant Google. Es pot iniciar la sessió amb l'email o també amb Microsoft, però mostrem l'opció de fer-ho amb Google.

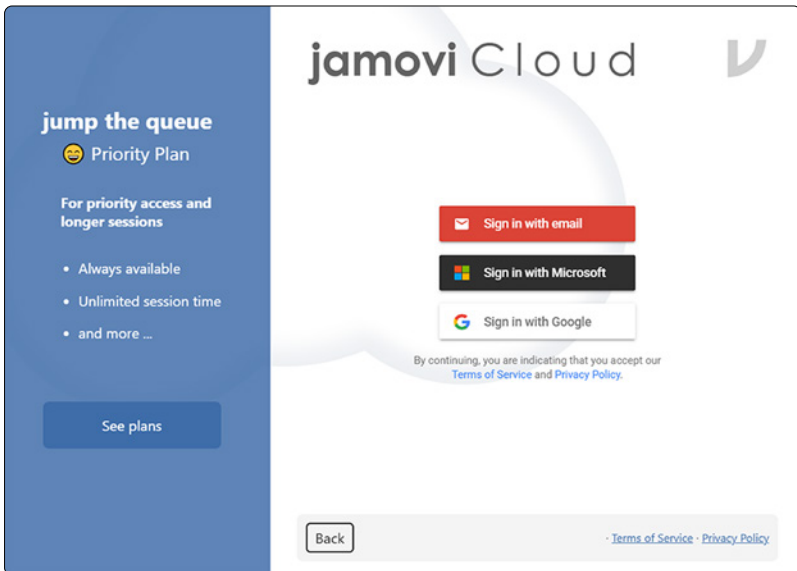


Figura 6. Web de Jamovi

La versió que s'obre en el mode online és la 2.6.19 i la interfície no difereix de la versió Desktop:

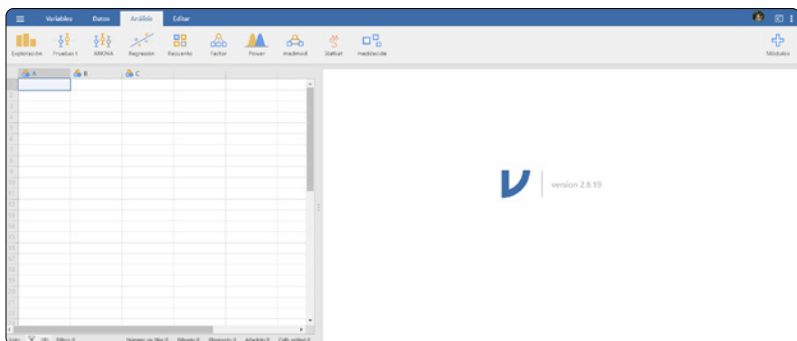


Figura 7. Jamovi en obrir l'inici del programa

## Mòduls

Com s'explicava en el capítol anterior, Jamovi compta amb extensions que es poden instal·lar per ampliar les funcionalitats del programa. En la versió Cloud també es poden instal·lar, a excepció del Rj Editor que, de moment, no està habilitat. La resta dels mòduls estan disponibles i la seva instal·lació és immediata, a diferència de la versió Desktop que requereix com a mínim uns minuts, depenent de la mida del mòdul i de la velocitat de la connexió a Internet.

També és important tenir en compte que si has carregat la versió Cloud des d'un compte de Google, en tornar a obrir el Jamovi des d'aquest mateix compte, els mòduls que ja vas instal·lar una vegada estaran disponibles la resta de les vegades que obris el programa.

A partir d'ara revisarem els mòduls comentats en el capítol anterior i esmentarem els avantatges i els inconvenients de cadascun en relació amb la versió Desktop.

### Mòdul jmv

Aquest mòdul no presenta diferències respecte a la versió Desktop.

### Mòdul jmvbaseR – Base R<sub>1.1.2</sub>

Aquest mòdul, jmvbaseR, no presenta diferències respecte a la versió Desktop.

### Mòdul MAJOR

La icona d'aquest mòdul difereix del que apareix en la versió Desktop, però les funcionalitats que ofereix són exactament les mateixes.

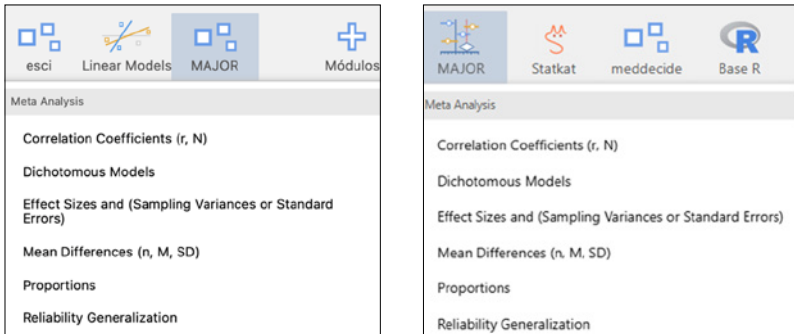


Figura 8. Mòdul MAJOR: versió Desktop (esquerra); versió Cloud (dreta)

## Mòdul ClinicoPath

Aquest mòdul no presenta diferències respecte a la versió Desktop.

## Mòdul Statkat

Aquest mòdul no presenta diferències respecte a la versió Desktop.

## Mòdul ufs

Aquest mòdul no presenta diferències respecte a la versió Desktop.

## Mòdul snowcluster, jsurvival i esci

En la versió Cloud sí que s'aprecien certes diferències en aquests tres mòduls en comparació amb la versió Desktop. En primer lloc, el logo del mòdul esci és diferent. Aquests serien els de la versió online:

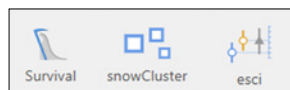


Figura 9. Icones de la versió Cloud

I amb relació a les funcionalitats, no hi ha diferències per a la llibreria jsurvival però sí per a les altres dues.

En el cas del mòdul snowcluster veiem, d'una banda, que la presentació de les opcions del menú és diferent i que, a més, la versió Cloud ofereix dues opcions que no apareixen en la versió Desktop: «Time series clustering» i «ROC Analysis».

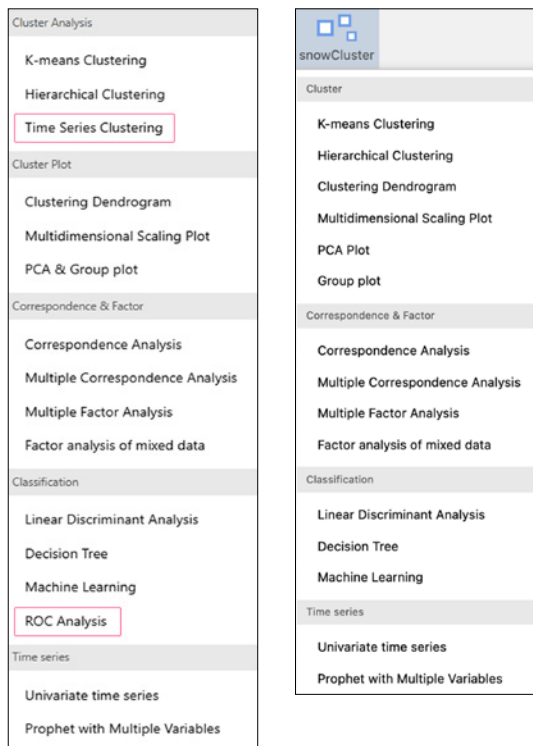


Figura 10. Mòdul snowcluster: versió Cloud (esquerra); versió Desktop (dreta)

En el cas del mòdul esci, la presentació de les funcions també difereix de la versió Desktop i, a banda, hi ha variacions

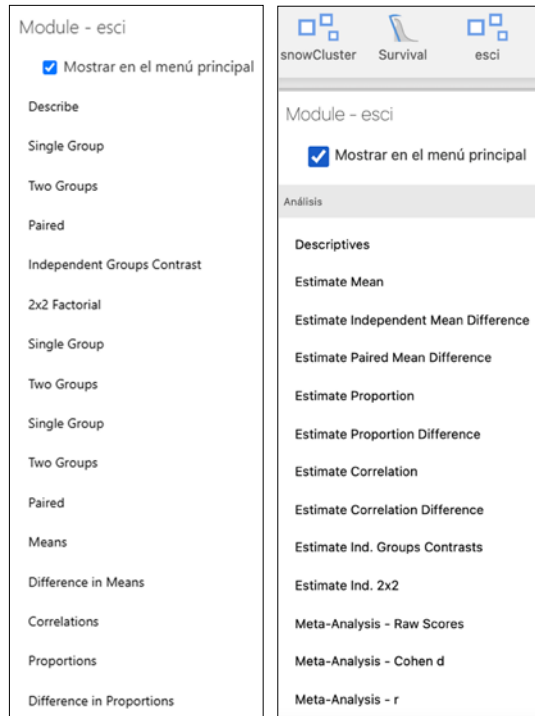


Figura 11. Mòdul esci: versió Cloud (esquerra); versió Desktop (dreta)

## Mòdul jpower

L'única diferència que té aquest mòdul és que en el menú apareix «Power» en comptes de «jPower», però tant la icona com les opcions que ofereix són idèntiques que la versió Desktop.



## Mòdul JJSTATPLOT i editor de Rj

El mòdul JJSTATPLOT, en el desplegable, té les mateixes funcions que la versió Desktop a excepció de «Graphs and Plots» dins de la categoria Type Based Graphs and Plots. Aquesta no apareix en la versió Cloud. En relació amb l'editor de Rj, com ja hem comentat, de moment, no hi ha opció d'instal·lar-lo en la versió online.

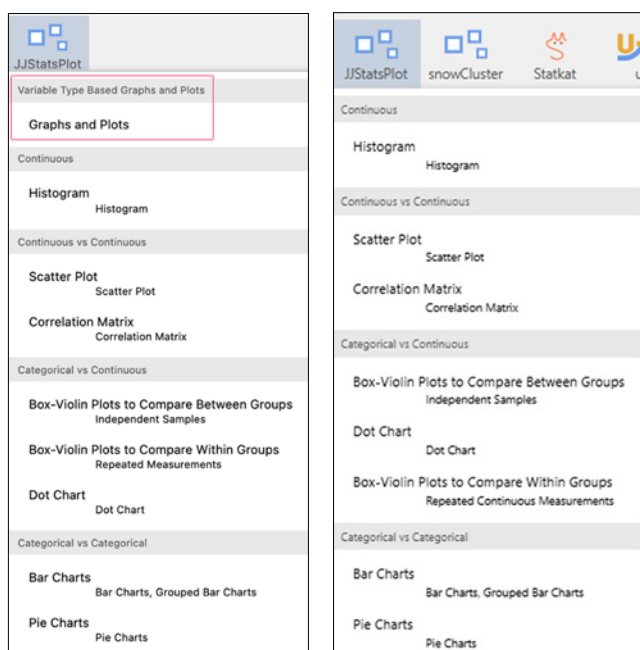


Figura 12. Mòdul JJStatsPlot: versió Cloud (esquerra); versió Desktop (dreta)

## Mòduls jsq i seolmatrix

El mòdul jsq té les mateixes funcions que la versió Desktop, però cal tenir en compte que un cop instal·lat no apareix cap icona al menú principal. No obstant això, les opcions es poden veure anant a Mòduls.

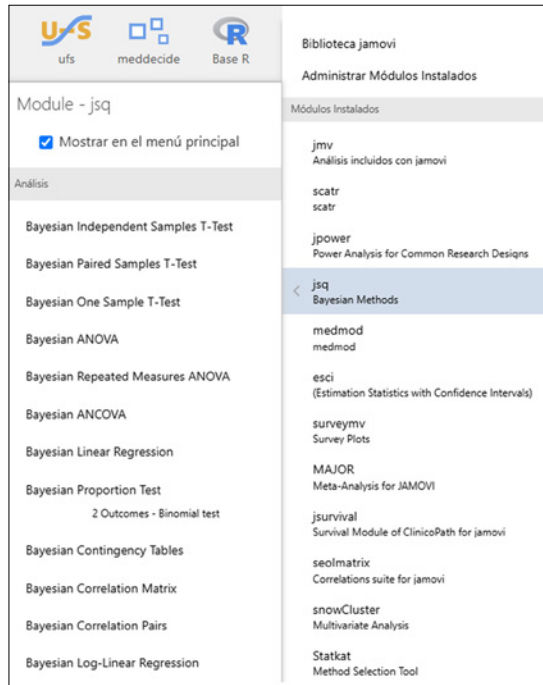


Figura 13. Mòdul jsq: visualització de les funcions del mòdul

En relació amb el mòdul seolmatrix, aquí sí que veiem diferències en les opcions que ens ofereix el desplegable, aquesta vegada sí, des de la icona del menú principal. En la versió Cloud, les opcions ressaltades en vermell són exclusives i no estan disponibles en la versió Desktop. D'altra banda, en la versió Desktop, les opcions assenyalades en blau són úniques i no es troben en la versió Cloud (figura 14).

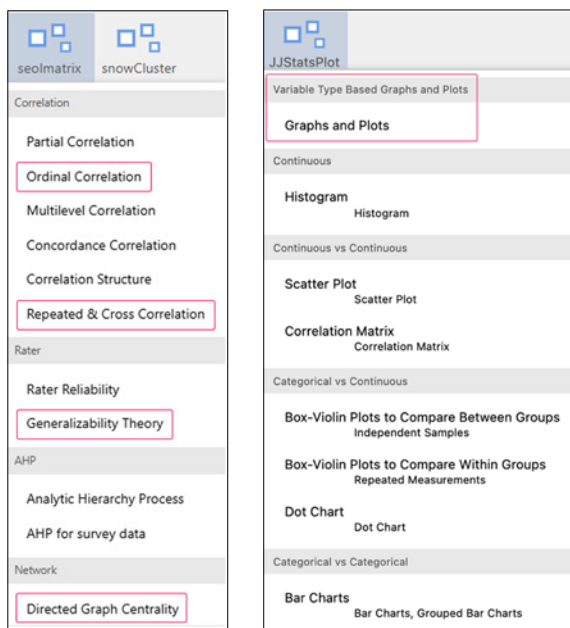


Figura 14. Mòdul seolmatrix: versió Cloud (esquerra); versió Desktop (dreta)

## Mòdul distrACTION

Aquest mòdul no presenta diferències respecte a la versió Desktop.

## Mòdul Flexplot i mòdul bmttest

Aquests dos mòduls no presenten diferències respecte a la versió Desktop.

## Mòdul Demonstration i mòdul clt

En la versió Cloud aquests dos mòduls estan unificats en un de sol que s'anomena clt - Demonstrations i no presenta diferències respecte a les funcionalitats de la versió Desktop.

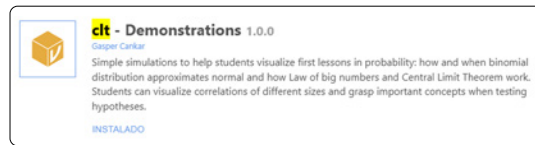


Figura 15. Mòdul clt - Demonstrations

## Mòdul jeva

S'observen diferències en les opcions que ens ofereix el desplegable:

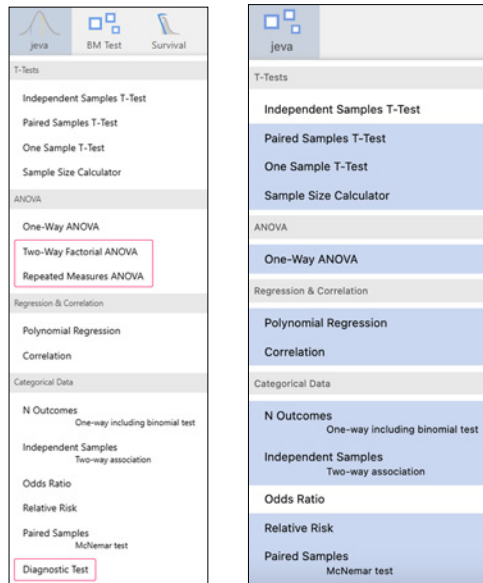


Figura 16. Mòdul jeva: versió Cloud (esquerra); versió Desktop (dreta)

## Mòdul ztestvis

Aquest mòdul no presenta diferències respecte a la versió Desktop.

## **Mòdul meddecide**

Aquest mòdul no presenta diferències respecte a la versió Desktop.

## **Mòduls PPDA i MEDA**

El mòdul MEDA no presenta diferències respecte a la versió Desktop.

D'altra banda, el mòdul PPDA no apareix a la biblioteca de Jamovi. No obstant això, veiem que el Test ROC es pot calcular des del mòdul snowcluster, en el qual existeix la funció «ROC Analysis». Les altres funcions «Binary LogR», «Ordinal Reliability» i «Non-Parametric IRT» s'hauria de comprovar que es poden utilitzar des d'altres mòduls.

## **Mòdul gamlj**

Aquest mòdul no presenta diferències respecte a la versió Desktop.

## Capítol 3

# Tractament de les dades

### Conceptes bàsics d'estadística

Entre les diferents definicions de l'estadística n'escollim dues: i) disciplina que estudia la variabilitat; ii) ciència que estudia com s'ha d'utilitzar la informació i com donar una guia d'acció en situacions pràctiques que presenten incertesa. El procés de recerca estadística requereix la recol·lecció, organització, anàlisi, interpretació i presentació de les dades.

La paraula *estadística*, tot i que té l'arrel llatina *status*, que significa 'xifres relacionades amb l'Estat', és en realitat la forma femenina del terme alemany *statistik*, derivat al seu torn de l'italià *statista*, que significa 'home d'estat'.

L'estadística es pot dividir en:

- **Estadística descriptiva:** l'estadística descriptiva és una branca de l'estadística que s'encarrega d'organitzar, resumir i presentar de manera sistemàtica i comprensible un conjunt de dades per extreure conclusions útils sobre les seves característiques. El seu objectiu principal és descriure i analitzar les dades mitjançant tècniques i mètodes que permeten entendre la seva distribució, tendència central, dispersió, forma i altres propietats rellevants.
- **Inferència estadística** o, també, **estadística inductiva:** és la part de l'estadística que tracta de les condicions i procediments sota els quals podem extreure conclusions o inferències de la població a partir de la informació mostral.

En aquest capítol 3 i en el següent, utilitzarem l'estadística descriptiva per a l'anàlisi de la base de dades Framingham.

## Explicació de les dades d'exemple: Framingham

Per al tractament de dades, com és lògic, calen dades. Per a l'explicació de totes les possibles tècniques que s'utilitzen a Jamovi relacionades amb la salut, utilitzarem una base de dades coneguda, la de Framingham.

La malaltia cardiovascular (MCV) és la principal causa de mort i malaltia greu als Estats Units. El 1948, es va iniciar l'Estudi Framingham sobre el Cor (d'ara endavant, estudi Framingham) sota la direcció de l'Institut Nacional del Cor (ara conegut com a Institut Nacional del Cor, Pulmó i Sang o NHLBI, per les seves sigles en anglès). En aquell moment, se sabia poc sobre les causes generals de les malalties cardíques i els accidents cerebrovasculars, tot i que les taxes de mortalitat per malaltia cardiovascular estaven augmentant constantment des de principis del segle xx, convertint-se en una epidèmia.

Més tard, l'estudi Framingham es va convertir en un projecte conjunt de l'NHLBI i la Universitat de Boston. L'objectiu de l'estudi Framingham era la identificació dels factors o característiques comunes que contribueixen al desenvolupament de malalties cardiovasculars en un gran grup de participants. Aquests participants encara no havien desenvolupat símptomes evidents de malaltia cardiovascular ni havien patit un atac cardíac o un accident cerebrovascular. Els investigadors van reclutar 5 209 homes i dones d'entre 30 i 62 anys, de la ciutat de Framingham, Massachusetts, i van començar la primera ronda d'exàmens físics i entrevistes sobre l'estil de vida, que posteriorment analitzarien a la recerca de patrons comuns relacionats amb el desenvolupament d'MCV. Des de 1948, els subjectes van ser controlats cada dos anys, recopilant el seu historial clínic detallat, fent-los un examen físic, així com proves de laboratori.

El 1971, l'estudi va incorporar una segona generació, que va consistir en 5 124 fills adults dels participants originals i els seus cònjuges. L'abril de 2002, l'estudi va entrar en una nova fase en registrar una tercera generació de participants, els nets de la cohort original, en concret, 4 095 participants. Aquest pas és de vital importància per ampliar la nostra comprensió de les malalties cardíques i els accidents cerebrovasculars, i com aquestes condicions afecten les famílies.

Al llarg dels anys, el seguiment precís de la població de l'estudi Framingham ha permès identificar els principals factors de risc de malaltia cardiovascular: pressió arterial alta, colesterol alt, tabaquisme, obesitat, diabetis i la inactivitat física, així com una gran quantitat d'informació valuosa sobre els efectes de factors relacionats, com els nivells de triglicèrids i colesterol HDL, d'edat, sexe i problemes psicosocials.

Amb l'ajut d'una altra generació de participants, l'estudi pot apropar-se a les causes profundes de les malalties cardiovasculars i ajudar en el

desenvolupament de noves i millors formes de prevenir, diagnosticar i tractar les malalties cardiovasculars.

L'estudi Framingham ha tingut un impacte significatiu en la pràctica mèdica. Previ a la dècada de 1950, es considerava normal que les persones amb alteracions en les constants vitals o resultats de laboratori no manifestessin signes clínics evidents de malaltia, qualificant aquestes anomalies com a benignes si no causaven molèsties clíniques.

Una troballa rellevant de l'estudi va ser la demostració estadística de la relació entre la diabetis i les malalties cardiocirculatòries, reconeixent la hiperglicèmia com un factor de risc. L'estudi Framingham es va iniciar amb el propòsit d'avaluar i demostrar els factors personals i ambientals que influïen en l'aparició primerenca de l'arterioesclerosi. Les conclusions de l'estudi, en confirmar estadísticament la importància dels factors de risc, han estat àmpliament satisfactòries i la seva implementació és essencial en la pràctica clínica.

Aquest estudi exemplifica el treball en equip, i els desafiaments enfrontats durant la seva continuació i expansió ressalten l'entusiasme i la dedicació dels professionals involucrats. Es va aconseguir l'objectiu proposat, validant estadísticament les percepcions disperses sobre el risc associat amb les alteracions en les constants físiques i metabòliques. L'estudi Framingham estableix de manera contundent el concepte de *factors de risc*.

L'arxiu Framingham.sav és un subconjunt anonimitzat de les dades corresponents a quaranta anys de seguiment. Conté mesures de variables sobre 4 434 pacients lliures de malaltia coronària en l'inici del seguiment (moment basal).

Aquestes dades, sempre per a una finalitat docent, es poden sol·licitar a <https://biolincc.nhlbi.nih.gov/teaching/> (requereix registre previ).

A continuació, veurem el conjunt de les variables que tenim en aquest fitxer, per familiaritzar-nos amb ell.

### Variables identificadores

- **Randid.** Número d'identificació únic per a cada participant
- **Participació.** 1 una onada, 2 dues onades, 3 tres onades
- **On\_participació.** 1 primera (onada), 2 segona, 3 tercera, 4 primera i segona (onades), 5 segona i tercera, 6 primera i tercera, 7 totes (les tres onades)
- **Seguiment.** Temps de seguiment (en anys).

### Variables resposta (variables dependents)

- **death.** Mort per qualsevol causa.
- **angina.** Angina de pit.
- **hospmi.** Infart de miocardi hospitalitzat.
- **mi\_fchd.** Infart de miocardi hospitalitzat o malaltia coronària fatal.
- **anychd.** Angina de pit, infart de miocardi (hospitalitzat i silenciós o no reconegut), insuficiència coronària (angina inestable) o malaltia coronària fatal.
- **stroke.** Infart aterotrombòtic, embòlia cerebral, hemorràgia intracerebral, hemorràgia subaracnoidal o malaltia cerebrovascular fatal.
- **cvd.** Infart de miocardi (hospitalitzat i silenciós o no reconegut), malaltia coronària fatal, infart aterotrombòtic, embòlia cerebral, hemorràgia intracerebral, hemorràgia subaracnoidal o malaltia cerebrovascular fatal.
- **hypertens.** Hipertensió. Definida com el primer examen tractat per pressió arterial alta o el segon examen en què la pressió sistòlica és major o igual a 140 mmHg o la pressió diastòlica és major o igual a 90 mmHg.
- **timeap.** Nombre de dies des de l'examen inicial fins al primer episodi d'angina durant el seguiment o nombre de dies des de l'examen inicial fins a la data de censura. La data de censura pot ser el final del seguiment, la mort o l'última data de contacte coneguda si el subjecte es perd durant el seguiment.
- **timemi.** Definit com a dalt per al primer esdeveniment d'HOSPMI durant el seguiment.
- **timemifc.** Definit com a dalt per al primer esdeveniment de MI\_FCHD durant el seguiment.
- **timechd.** Definit com a dalt per al primer esdeveniment d'ANYCHD durant el seguiment.
- **timestrk.** Definit com a dalt per al primer esdeveniment de STROKE durant el seguiment.
- **timecvd.** Definit com a dalt per al primer esdeveniment de CVD durant el seguiment.
- **timedth.** Nombre de dies des de l'examen inicial fins a la mort si s'escau durant el seguiment o nombre de dies des de l'examen inicial fins a la data de censura. La data de censura pot ser el final del

seguiment o l'última data de contacte coneguda si el subjecte es perd durant el seguiment.

- **timehyp.** Definit com a dalt per al primer esdeveniment de HYPERTEN durant el seguiment.

### **Característiques i factors de risc (variables independents)**

- **sexe.** Sexe del participant.
- **edat basal – edat\_basal\_cat.** Edat en el moment basal (anys i categories, quartils).
- **edat – edat\_cat.** Edat en l'examen (anys i categories, quartils).
- **sysbp.** Pressió arterial sistòlica (mitjana dels dos últims de tres mesuraments) (mmHg).
- **diabp.** Pressió arterial diastòlica (mitjana dels dos últims de tres mesuraments) (mmHg).
- **cursmoke.** Consum actual de cigarrets en l'examen.
- **cigpday.** Nombre de cigarrets fumats per dia.
- **imc.** Índex de massa corporal, pes en quilograms/altura en metres quadrats.
- **diabetis.** Diabètic segons criteris del primer examen tractat o primer examen amb glucosa casual de 200 mg/dL o més.
- **bpmeds.** Tractament per a la hipertensió.
- **glucosa.** Glucosa.
- **prevap:** Angina de pit prevalent en l'examen.
- **prevchd.** Malaltia coronària prevalent definida com a angina de pit preexistent, infart de miocardi (hospitalitzat, silenciós o no reconegut) o insuficiència coronària (angina inestable).
- **prevmi.** Infart de miocardi prevalent.
- **prevstrk.** Accident cerebrovascular prevalent.
- **prevhyp.** Hipertensió prevalent. El subjecte va ser definit com a hipertens si era tractat o si en el segon examen la pressió sistòlica mitjana era  $\geq 140$  mmHg o la pressió diastòlica mitjana era  $\geq 90$  mmHg.
- **hearttrte.** Freqüència cardíaca.
- **totchol.** Colesterol total en sèrum (mg/dL).
- **hdlc.** Colesterol de lipoproteïnes d'alta densitat (HDL) (mg/dL).

- **ldlc**. Colesterol de lipoproteïnes de baixa densitat (LDL) (mg/dL).
- **overweight**. 1 sobrepès ( $25 \text{ kg/m}^2 \leq \text{IMC} \leq 30 \text{ kg/m}^2$ ), 0 un altre cas.
- **obessi**. 1 obesitat ( $\text{IMC} \geq 30 \text{ kg/m}^2$ ), 0 un altre cas.
- **Garrow**. 0 normopès, 1 sobrepès, 2 obès.

A partir del coneixement de les diferents variables, passem a treballar-les, després de conèixer una mica els tipus de fitxers que podem utilitzar a Jamovi.

## Tipus de fitxers

Jamovi és compatible amb una gran varietat de tipus d'arxius, cosa que el fa molt versàtil per treballar amb diferents tipus de dades estadístiques.

Al menú «Arxius», podem veure la forma d'obrir-los (figura 1).

Les opcions que tenim:

- Nou. Obrir una nova instància de Jamovi. A més, ho fa sense tancar l'actual.
- Obrir. Mostra una finestra per seleccionar un arxiu de dades per a carregar-les dins de Jamovi.
  - Funciona amb una varietat d'arxius d'aplicacions comunament utilitzades per a l'anàlisi de dades quantitatives (SPSS, Stata, SAS, JASP) i fulls de càlcul (LibreOffice Calc o Excel), així com arxius de R o el format genèric per a intercanvi de dades CSV.
  - A més, ens permet obrir dades que nosaltres tinguem o dades que ens proporciona el mateix Jamovi (biblioteca de dades).

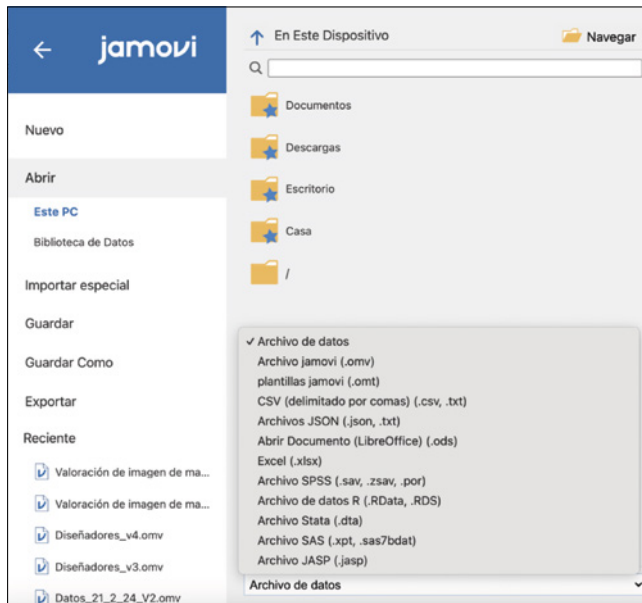


Figura 1. Imatge de Jamovi, per obrir arxius

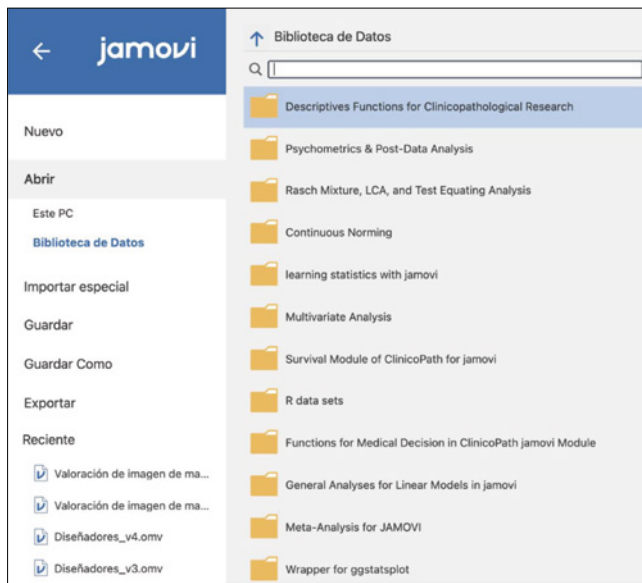


Figura 2. Obrir arxius de la biblioteca de dades

En la figura 2, podem veure els diferents exemples que, per defecte, ens proporciona Jamovi per poder aplicar les diferents tècniques estadístiques d'anàlisi i descriptives.

- Importar especial. Mostra una finestra per seleccionar un o més arxius de dades per agregar-los dins de Jamovi. Funciona amb arxius del tipus OMV.
- Desar i Desar com. Guarda en el format propi de Jamovi (OMV) tota la informació de la sessió actual d'anàlisi, no només la matriu de dades, amb les identifications de cada variable original i cada variable generada o transformada, així com tot el codi R que genera els resultats sol·licitats, així com tot el text agregat als resultats.
- Exportar. Genera diferents tipus d'arxius amb contingut específic, segons el format seleccionat per a l'exportació.
- Resultats. Es poden generar pàgines amb totes les taules i gràfics dels resultats i amb els textos adjunts, en el cas d'exportacions cap a PDF, HTML o LaTeX (paquet zip).
- Dades. En el cas d'exportacions cap a formats d'aplicacions d'anàlisi, el que es genera són arxius únicament amb les dades, sense els resultats.

Ara que coneixem els tipus d'arxius que podem utilitzar a Jamovi, analitzarem els diferents tipus de variables.

## Tipus de variables

Abans de conèixer els tipus de variables, coneixerem els passos previs per preparar les dades. El full de treball de Jamovi està organitzat en columnes, cadascuna de les quals es refereix a una variable. El tipus d'anàlisi a realitzar està lligat al tipus de variables. Podem veure, a continuació, la finestra que se'ns obrirà a la pestanya de dades (figura 3).

Pel que fa al tipus de variable a definir o utilitzar, podem veure que tenim quatre tipus (figura 4): nominal, ordinal, contínua i ID.

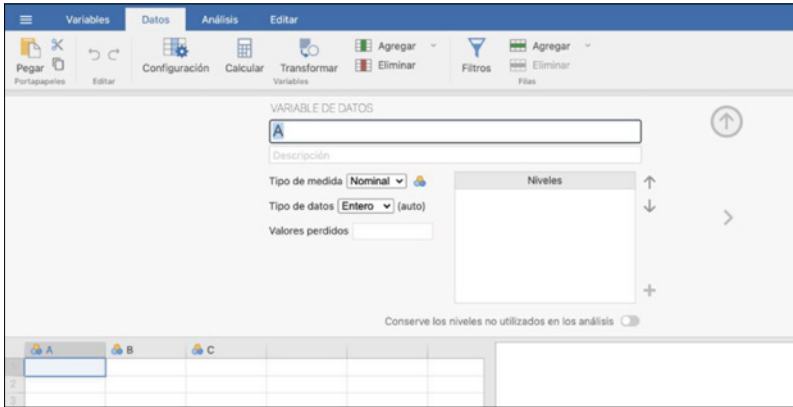


Figura 3. Configuració de dades

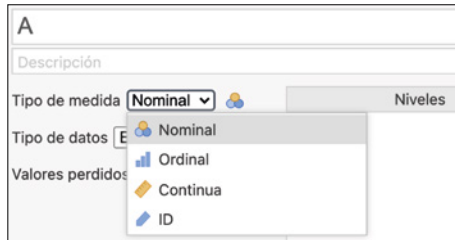


Figura 4. Tipus de dades

Podem definir, de forma resumida, cadascuna d'elles:

- Variables identificadores (ID). Són variables que identifiquen una observació (subjecte, mesura, data, etc.). Per exemple, les variables identificadores de la base de dades Framingham.
- Variables qualitatives, categòriques o atributs. Són variables les modalitats de les quals no són mesurables, expressant únicament qualitats o categories. No es poden expressar mitjançant valors numèrics, encara que puguem utilitzar números per identificar i simplificar cada categoria. Per exemple, la variable dependent de Framingham *death*, que pren els valors 'viu' i 'mort', que codifiquem com a 0 si el subjecte està 'viu' al final del seguiment de la cohort i 1 si 'va morir' abans del final del seguiment. Un altre exemple, en aquest cas variable independent (característiques i factors de risc), Garrow, que pren els valors 'normopès', 'sobrepès' i 'obès', i codifiquem 0 'normopès', 1 'sobrepès' i 2 'obès'.

Les variables qualitatives es classifiquen en:

- Nominals. Només relacions d'identitat. Permeten obtenir informació de la igualtat o desigualtat d'aquests individus per a una determinada característica.
- Les variables nominals poden ser tant dicotòmiques, que pren dos valors, per exemple, la variable *death* a la base Framingham, com politòmiques, més de dues categories. Una variable política nominal serien els tipus de *stroke*: infart aterotrombòtic, embòlia cerebral, hemorràgia intracerebral, hemorràgia subaracnoidal o malaltia cerebrovascular fatal.
- Ordinals. Es pot valorar el seu ordre. Ens permeten obtenir informació de la igualtat o desigualtat d'aquests individus per a una determinada característica. Les variables ordinals només són politòmiques (més de dos valors), per exemple, la variable *Garrow* a la base Framingham.
- Variables quantitatives. Les seves modalitats són mesurables i es poden expressar amb valors numèrics. Aquestes variables es classifiquen en:
  - Discretes. Són aquelles que prenen valors sencers no negatius i amb un nombre finit o infinit numerable de valors diferents. Per exemple, la variable *timemi* a la base de dades Framingham, el nombre de dies des de l'examen inicial fins al primer episodi d'infart de miocardi o, si el subjecte no ha patit un infart, fins al final del seguiment. Aquesta variable pren valors sencers no negatius (0, 1, 2, etc.) i amb un rang no gaire gran, sent un nombre finit de valors (com a màxim el nombre de dies de seguiment).
  - Contínues. Són aquelles que poden prendre qualsevol valor dels infinits que hi ha en un interval de valors. Per exemple, les variables independents de la base Framingham *sysbp*, pressió arterial sistòlica, *diabp*, pressió arterial diastòlica i *glucosa*, nivell de glucosa.

Vegem un exemple (figura 5) on la mateixa variable pot ser nominal, on conservarà les classes o contínua, on aquestes desapareixeran.

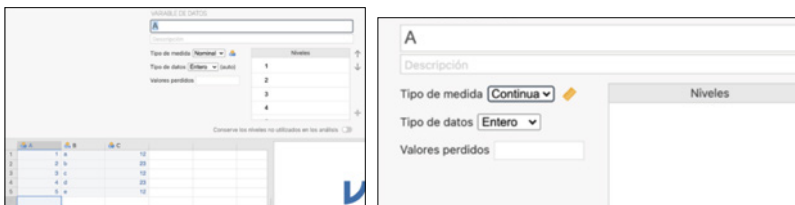


Figura 5. Canvi de variable

## Tractament de dades

### Creació

La primera de les possibilitats que ofereix Jamovi quant a tractament de dades és la seva creació. Vegem, doncs, quins són els passos a seguir en el supòsit d'haver d'introduir dades.

Dades (exemple: nombre de membres d'una família, que seria una variable quantitativa discreta):

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

A les següents finestres (figura 6), podem veure els passos simples per a la introducció de les dades.

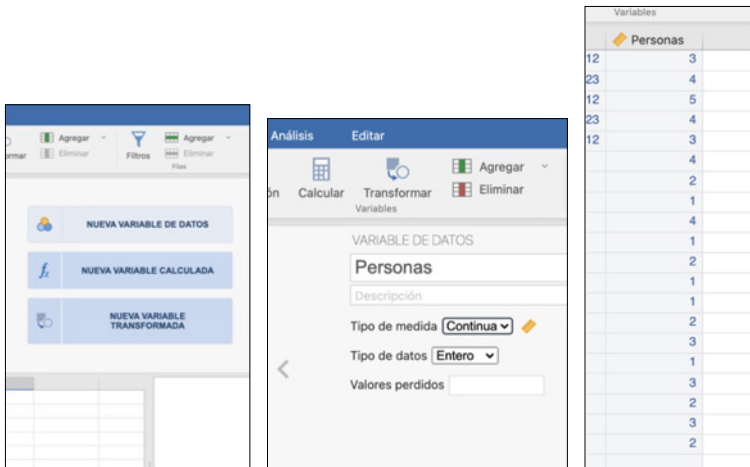


Figura 6. Introducció de dades

Es pot també crear el fitxer de les dades, a partir d'un altre que disposem en un altre *programari* amb format de full de càlcul, copiant-los directament. Això ens permet treballar o crear dades que tinguem en files i les vulguem en columnes. Per això, s'aconsella la creació del fitxer de dades mitjançant el procés de còpia a Jamovi.

## Transformació

Per transformar una variable, cal tenir clar que en crearem una de nova a partir d'una original.

Per exemple, si agafem les dades de l'apartat anterior, el primer pas serà polsar la icona de «Transformar» (figura 7, esquerra).

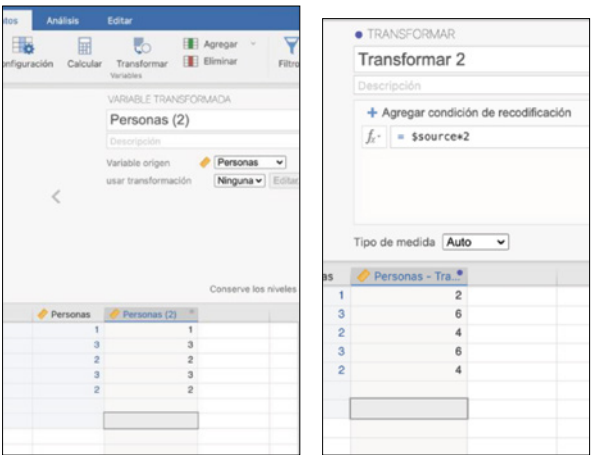


Figura 7. Transformació de variables

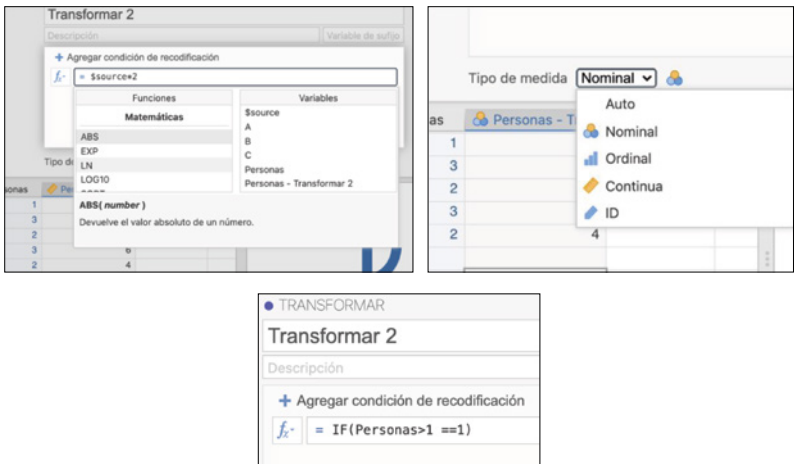


Figura 8. Transformació de variables amb la funció IF

A continuació, es crea i s'aplica la transformació necessària (figura 7, dreta). Per exemple, el que s'ha fet és multiplicar la variable per 2.

També podem sol·licitar que compleixi unes condicions amb la funció IF (figura 8). En aquest cas, el que s'ha aconseguit és la creació d'una variable que ens dona el valor 1 quan tenim més d'una persona per família.

Veiem que, a més, es poden utilitzar moltes funcions tant matemàtiques com estadístiques o lògiques per a la creació de la transformació.

## Recodificació de variables

En l'àmbit de la codificació de dades, la recodificació de variables és una tècnica àmpliament utilitzada. Si bé la seva funció principal és agrupar un rang de valors existents en noves categories, també permet assignar una nova codificació als valors originals, facilitant així l'anàlisi posterior.

En qualsevol cas, la *recodificació* sempre implica un canvi en els valors de les variables, un canvi que pot o no implicar una modificació en la mètrica de la variable. En aquest sentit, es poden donar quatre situacions diferents, en funció del tipus de les variables original i final: passar de qualitativa a quantitativa o a una mateixa qualitativa o a l'inrevés, de quantitativa a qualitativa o quantitativa. En tots els casos, s'ha d'utilitzar el procediment d'aplicació d'una regla de transformació (variable origen + regla de transformació = variable final). Es tracta de crear una nova variable a través d'altres prèvies.

En els dos casos, tant les transformacions com la recodificació de variables, és possible que siguin utilitzades per a casos més complexos. S'aniran desenvolupant durant el document quan aquestes siguin necessàries.

## Filtres

En alguns casos, cal que seleccionem només alguns casos de les dades que tenim. Per exemple, en les dades exemple, figura 9, seleccionar només els casos que siguin, de la variable *Persones*, major o igual a 3. Observem que s'ha creat una columna on posa filtre i amb els casos seleccionats.

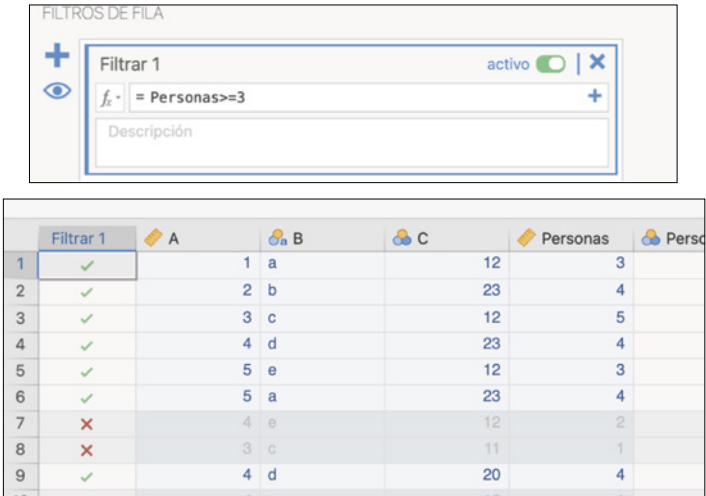


Figura 9. Filtre de variables

Això ens permetrà seleccionar dades de casos concrets, com, per exemple, en bases de dades on vulguem separar per alguna característica: sexe, edat o altres que compleixin una condició.

Càlculs

Finalment, també podem fer, de forma simple, càlculs amb les variables. A la imatge (figura 10), podem veure com crear una nova variable que sigui l'LN de la columna A.

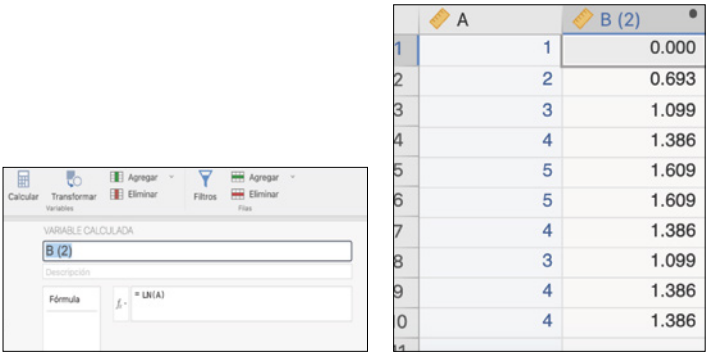


Figura 10. Càlcul de variables

## Exemple de tractament de dades i variables

Amb el fitxer Framingham, veurem algun exemple dels apartats anteriors, aplicat a aquestes dades.

Primerament, podem veure en el fitxer diferents tipus de variables. Per exemple, com a variables contínues o variables qualitatives (figura 11).

| timechd | timestrk | timecvd | timedth | timehyp | donde_participacion        | follow_up | death | angina |
|---------|----------|---------|---------|---------|----------------------------|-----------|-------|--------|
| 17.626  | 24.000   | 17.626  | 24.000  | 24.000  | Primera y tercera          | 13        | No    | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 24.000  | Primera, segunda y tercera | 12        | No    | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 24.000  | Primera y segunda          | 6         | No    | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 24.000  | Primera y segunda          | 6         | Yes   | No     |
| 8.093   | 5.719    | 5.719   | 8.093   | 0.000   | Primera, segunda y tercera | 12        | No    | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 11.732  | Primera y segunda          | 7         | No    | Yes    |
| 15.658  | 24.000   | 15.658  | 24.000  | 0.000   | Primera, segunda y tercera | 12        | No    | No     |
| 1.021   | 24.000   | 24.000  | 24.000  | 6.056   | Primera, segunda y tercera | 12        | No    | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 23.762  | Primera, segunda y tercera | 12        | No    | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 0.000   | Primera                    |           | No    | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 0.000   | Primera y segunda          | 6         | No    | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 0.000   | Primera, segunda y tercera | 12        | Yes   | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 24.000  | Primera, segunda y tercera | 12        | Yes   | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 24.000  | Primera                    |           | Yes   | No     |
| 24.000  | 24.000   | 24.000  | 24.000  | 24.000  | Primera, segunda y tercera | 12        | No    | No     |

Figura 11. Variables contínues (esquerra) i variables qualitatives (dreta)

Crearem un filtre de les dones (figura 12), seleccionarem només els que són homes (Male) (funció sex=="Male").

| ecvd   | timedth | timehyp | sex    |
|--------|---------|---------|--------|
| 17.626 | 24.000  | 24.000  | Male   |
| 24.000 | 24.000  | 24.000  | Female |
| 24.000 | 24.000  | 24.000  | Male   |
| 5.719  | 8.093   | 0.000   | Female |
| 24.000 | 24.000  | 11.732  | Female |
| 15.658 | 24.000  | 0.000   | Female |
| 24.000 | 24.000  | 6.056   | Female |
| 24.000 | 24.000  | 23.762  | Female |
| 24.000 | 24.000  | 0.000   | Male   |

Filtrar 1

✓

2448

Dues onades

Primera y tercera

✗

6238

Tres onades

Primera, segunda y tercera

✓

9428

Dues onades

Primera y segunda

✗

10552

Dues onades

Primera y segunda

✗

11252

Tres onades

Primera, segunda y tercera

✗

11263

Tres onades

Primera, segunda y tercera

✗

12629

Dues onades

Primera y segunda

✗

12806

Tres onades

Primera, segunda y tercera

✓

14367

Tres onades

Primera, segunda y tercera

✓

16365

Tres onades

Primera, segunda y tercera

✗

18793

Tres onades

Primera, segunda y tercera

activo

f<sub>x</sub>

= sex=="Male"

Descripción

Figura 12. Filtre de les dades per sexe

Un altre exemple és transformar les dades d’una variable, i que passin a ser 0 i 1, si compleix o no una condició (figura 13).

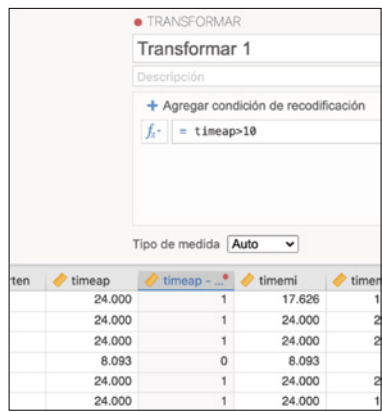


Figura 13. Transformació variable timeap

Aquests són només dos exemples simples en què s'utilitzen les dades del fitxer Framingham. Durant el document, es realitzaran les transformacions, càlculs o filtres que calguin.

## Referències

Balcells, M. (2016). L'estudi Framingham. *Neurosciences and History*. 4(1), 43-46.

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

## Capítol 4

# Descriptiva d'una variable

En aquest capítol desenvoluparem la descriptiva d'una variable. Per a això hem de saber què és l'estadística descriptiva i quins elements la componen. En aquest cas ens centrarem en una variable; en el següent capítol ho farem amb més d'una variable i fent la separació, per exemple, per classes, prenent-ho com una segona variable.

Les tècniques comunes utilitzades en estadística descriptiva inclouen:

- **Mesures de tendència central.** Com la mitjana, la mediana i la moda, que representen valors típics o centrals del conjunt de dades.
- **Mesures de dispersió.** Com la variància i la desviació estàndard, que indiquen la dispersió o la variabilitat de les dades al voltant de la mesura central.
- **Mesures de posició.** Com els percentils i els quartils, que divideixen el conjunt de dades en parts iguals o proporcionals.
- **Representacions gràfiques.** Com histogrames, diagrames de barres, diagrames de caixa (*boxplots*), diagrames de dispersió, entre d'altres, que ofereixen una visualització intuïtiva de la distribució de les dades i les seves característiques.

L'estadística descriptiva proporciona una primera vista a les dades i és fonamental en l'etapa inicial de qualsevol anàlisi estadística o estudi, ja que ajuda a entendre la naturalesa i l'estructura de les dades abans de realitzar inferències més profundes mitjançant tècniques d'estadística inferencial.

A continuació, desenvoluparem el tema de la descriptiva, però separant en funció del tipus de variable, la qual cosa farà que els resultats i la forma de mostrar-los siguin diferents. A més, a partir d'aquí, el desenvoluparem a través de les dades del fitxer que prenem com a exemple, les dades Framingham.

## Variables quantitatives

Per entendre el cas de descriure variables contínues, utilitzarem algunes de les variables que hem descrit en el capítol anterior (*age\_basal*, *sysbp1*,

*diabp1*, *bmi1*, *timecvd*). A més, utilitzarem una descripció que manegi gran quantitat d'elements mostrats en els diferents paquets de Jamovi perquè sigui més completa.

Començarem per les mesures de **tendència central**.

La **mitjana** (o mitjana aritmètica) s'obté amb la suma d'un conjunt de valors dividida entre el nombre total de sumands. Sol representar-se pel símbol  $\bar{x}$ . Així, es calcularia com a

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

on  $\Sigma$  denota la suma,  $i$  és l'ordre de l'observació ( $i = 1, 2, \dots, n$ ),  $n$  és el nombre total de sumands i  $x_i$  és el valor de la variable  $x$  per a l'observació  $i$ .

Per exemple, considerem l'edat, a l'inici del seguiment, dels primers 11 subjectes de la base de dades Framingham.

**Taula 1. Edat d'11 subjectes de la base de dades Framingha**

| 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 mitjana seria 48,73 anys.

$$\bar{x} = \frac{\sum_{i=1}^n \text{age\_basal}_i}{n} = \frac{39 + 46 + 48 + 61 + 46 + 43 + 63 + 45 + 52 + 43 + 50}{11} = 48,73$$

La **mediana** és la mesura de tendència central que, ordenades les dades de menor a major, deixa igual nombre de valors a la seva esquerra que a la seva dreta.

Seguint l'exemple anterior, en primer lloc, ordenarem l'edat de menor a major (també podria ser al revés).

Taula 2. Edat d'11 subjectes de la base de dades Framingham. Dades ordenades

| 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 aquest cas, la mediana seria 46, ja que deixa el mateix nombre de valors (5 valors) per sota (o esquerra) que per dalt (o dreta). Si tinguéssim un nombre de valors parell, per exemple, només els deu primers casos.

Taula 3. Edat de 10 subjectes de la base de dades Framingham

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

La mediana es calcularia com la mitjana aritmètica dels valors en la posició 5 i 6, és a dir

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

que en aquest cas també és igual a 46.

La **moda** és el valor de la variable que més vegades es repeteix, és a dir, el que presenta més freqüència absoluta.

En l'exemple que utilitzem hi ha dues modes, 43 i 46, que són les úniques que es repeteixen dues vegades. En aquest cas es diria que la distribució és bimodal.

Hi ha diverses **mesures de posició** com el valor **mínim** (39 anys en el nostre exemple, vegeu taula 2) i el **màxim** (63 anys en el nostre exemple).

A més, els **quantils** o **percentils** són les mesures de posició que generalitzen el concepte de mediana. Ordenades les observacions de menor a major, es defineix el percentil  $p$  ( $0 < p < 1$ ) com el valor que deixa  $p\%$  observacions a la seva esquerra i  $(1 - p)\%$  observacions a la dreta.

A la taula 2, per exemple, el percentil 20 (també denominat segon decil), correspon a 43, el tercer valor d'edat, ja que deixa a dalt (esquerra) un 20 % dels valors i a baix (a la dreta) un 80 %.

Alguns percentils tenen nom propi, com els decils (el primer decil acumula el 10 % dels valors, el segon decil el 20 %, etc.), els tercils (el primer tercil acumula el 33 % dels valors, el segon el 66 %, etc.) i, sobretot, els **quantils**. El primer quantil (o Q1) acumula el 25 % dels valors, el segon (o Q2 o mitjana) acumula el 50 % dels valors, el tercer quantil (o Q3) el 75 % dels valors i el quart quantil (o Q4 o màxim) el 100 % dels valors.

Tornant al nostre exemple a la taula 2, el primer quantil correspon al tercer valor, igual a 43, ja que acumula el 25 % dels valors; el segon quantil o la mediana ja vam veure que era igual a 46; i el tercer quantil correspon al novè valor, igual a 52, ja que acumula el 75 % dels valors. Els quantils s'interpreten com segueix: primer quantil: el 25 % dels 11 subjectes tenen com a màxim 43 anys; segon quantil o mediana: el 50 % dels 11 subjectes tenen com a màxim (i en aquest cas també com a mínim) 46 anys; i tercer quantil: el 75 % dels 11 subjectes tenen com a màxim 52 anys.

Pel que fa a les **mesures de dispersió**, explicarem primer el **rang**, **recorregut** o **amplitud**, calculat com la diferència entre el valor màxim i el mínim. Tornant a la taula 2, el rang, màxim-mínim,  $63 - 39$ , és de 24 anys.

La mesura de dispersió de la mitjana és la **variància** ( $s^2$ ), que no és més que la mitjana de les desviacions de cada valor de la variable respecte a la seva mitjana, és a dir:

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

El problema és que la variància no té les mateixes unitats que la mitjana, per la qual cosa s'utilitza la seva arrel quadrada, anomenada **desviació típica** o **desviació estàndard** ( $s$ ).

Tornant al nostre exemple:

Taula 4. Edat d'11 subjectes de la base de dades Framingham

| randid | age_basal                          | age_basal - $\bar{x}$ | (age_basal - $\bar{x}$ ) <sup>2</sup>                            |
|--------|------------------------------------|-----------------------|------------------------------------------------------------------|
| 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 | $\bar{x} = \frac{536}{11} = 48,73$ |                       | $s^2 = \frac{556,181}{11} = 50,562$<br>$s = \sqrt{50,562} = 7,1$ |

La mitjana és igual a 48,73 anys, la variància és igual a 50,562, i la desviació típica igual a 7,1 anys. Per veure si la distribució de la variable és molt o poc dispersa, és millor comparar la desviació típica amb la mitjana, cosa que dona com a resultat l'anomenat coeficient de variació ( $CV = (s / \bar{x}) * 100$ ). En el nostre cas

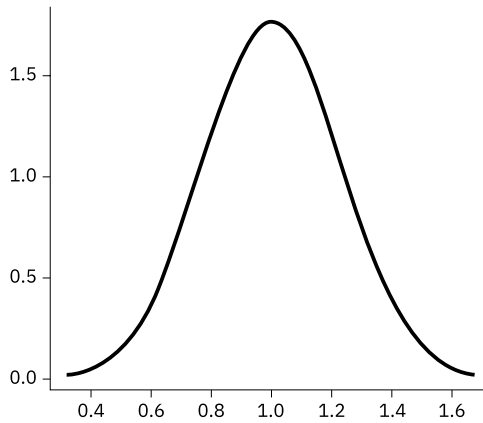
$$CV = (7,1 / 48,73) * 100 = 14,57\%$$

És a dir, la distribució de la variable edat entre aquests 11 subjectes no és gaire dispersa.

La mesura de dispersió de la mediana és el **rang (o recorregut) interquartilic** (IQR), que es calcula com el valor del tercer quartil menys el valor del primer quartil,  $Q3 - Q1$ .

En el nostre cas,  $IQR = Q3 - Q1 = 52 - 43 = 9$  (anys). Comparant l'IQR amb la mediana, que era de 46 anys, tampoc sembla una distribució gaire dispersa (9 és un 19,56 % de 46).

Les **mesures de forma** tracten d'identificar certes desviacions en la forma de la distribució de freqüències d'una variable respecte a un model de referència: la distribució normal, la funció de densitat de la qual adopta una forma característica coneguda com a campana de Gauss.



**Figura 1. Funció de densitat de la distribució normal, campana de Gauss**

Les variables quantitatives contínues segueixen una distribució de probabilitat normal, amb una funció de densitat simètrica, com veiem en la figura 1. Però això és en teoria. A la pràctica, les variables quantitatives contínues segueixen una distribució similar a la normal, depenent del nombre d'observacions (com més observacions tingui la variable més s'hi assemblarà la distribució) i de l'existència de valors extrems (com menys valors extrems més s'hi assemblarà).

Per explicar les mesures de forma és millor referir-se primer a una de les representacions gràfiques, l'**histograma**. Un histograma és una representació gràfica d'una variable en forma de barres, on la superfície de cada barra és proporcional a la freqüència dels valors representats.

La representació de l'histograma de la variable *bmi3*, índex de massa corporal corresponent a la tercera mesura, és bastant simètric, tot i que sembla més apuntat (veurem més endavant què significa) que el que correspondria a una variable que es distribuís com una distribució normal.

Però, a més, es poden utilitzar algunes mesures de forma.

**Coefficient d'asimetria.** Mesura si les observacions estan disposades de forma simètrica respecte a la mitjana. Si el coeficient d'asimetria és igual a 0, la distribució és simètrica (també denominada esbiaixada). Quan el coeficient sigui més gran que 0, la distribució serà asimètrica a la dreta (o presenta un biaix positiu) i quan sigui menor que 0 la distribució serà asimètrica a l'esquerra (o presenta un biaix negatiu). Quan la distribució sigui simètrica, la mitjana serà molt semblant a la mediana i, si és unimodal (només hi ha una moda), també a la moda.

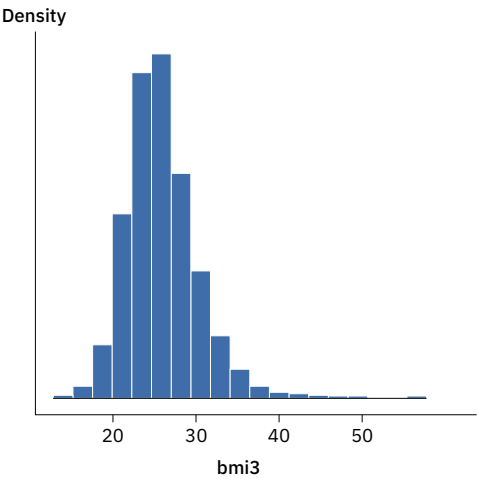


Figura 2. Histograma de la variable *bmi3* a la base de dades Framingham

En el cas de la variable *bmi3* (taula 5), veiem que la seva distribució és bastant simètrica, potser amb certa asimetria a la dreta, però no excessiva, ja que la mitjana i la mediana són pràcticament iguals.

Taula 5. Alguns estadístics descriptius per a la variable *bmi3*

| Descriptives        |             |
|---------------------|-------------|
|                     | <b>bmi3</b> |
| N                   | 3246        |
| Missing             | 1188        |
| Mean                | 25.9        |
| Median              | 25.5        |
| Skewness            | 0.916       |
| Std. error skewness | 0.0430      |

La distribució de la variable *x2* (simulada), però, és asimètrica a l'esquerra, amb un coeficient d'asimetria negatiu molt diferent de zero,  $-3,05$ . Quan la distribució és asimètrica a l'esquerra la mitjana és a l'esquerra de la mediana. En aquest cas, la mitjana de *x2* és igual a 1,524, mentre que la mediana és igual a 1,622.

Contràriament, la distribució de la variable *x3* (simulada) és asimètrica a la dreta, amb un coeficient d'asimetria positiu molt diferent de zero,  $1,47$ . Quan la distribució és asimètrica a la dreta la mitjana és a la dreta de la

mediana. En aquest cas, la mitjana de  $x_3$  és igual a 3,092, mentre que la mediana és igual a 2,497.

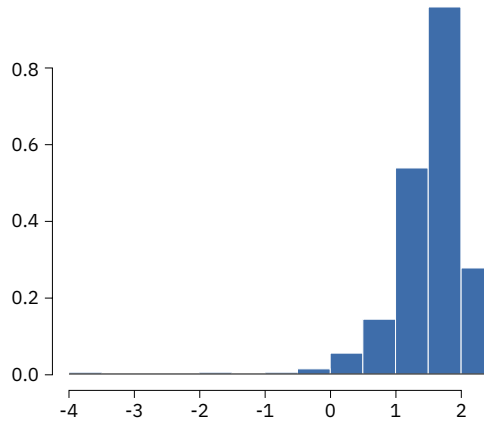


Figura 3. Histograma de la variable simulada  $x_2$

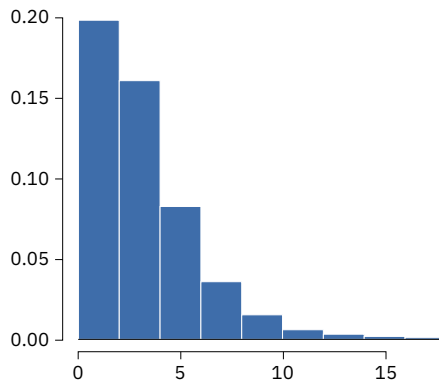
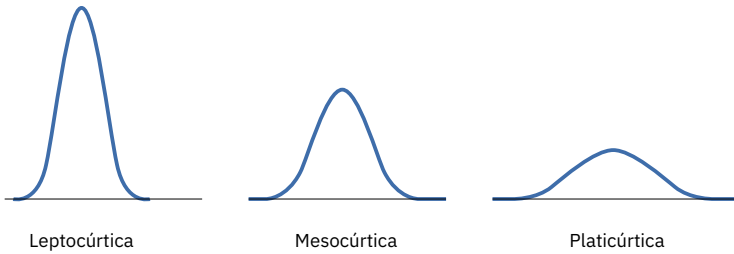


Figura 4. Histograma de la variable simulada  $x_3$

De fet, quan la distribució d'una variable quantitativa contínua no és simètrica, no s'han d'utilitzar la mitjana ni la desviació típica per descriure-la. En aquest cas són més representatives la mediana i el rang interquartílic.

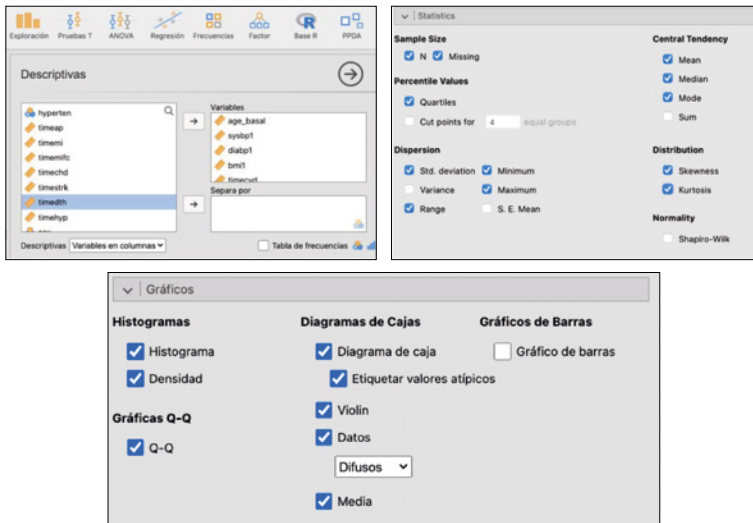
Pel que fa a l'**apuntament** (també denominat curtosi), quan el coeficient d'apuntament sigui igual a zero, la distribució serà igual d'apuntada que la normal (o mesocúrtica), quan sigui més gran que zero més apuntada (o

leptocúrtica) i quan sigui menor que zero menys apuntada o més aixafada (o platicúrtica).



**Figura 5. Densitat de variables amb diferent apuntament**

Utilitzant Jamovi descriurem, a continuació, un conjunt de variables quantitatives de la base de dades Framingham.



**Figura 6. Finestres que apareixen (Jamovi) a > Descriptiva**

Taula 6. Variables quantitatives

|                      | age_basal | sysbp1   | diabp1   | bmi1     | timecvd  |
|----------------------|-----------|----------|----------|----------|----------|
| N                    | 4434      | 4434     | 4434     | 4415     | 4434     |
| Perduts              | 0         | 0        | 0        | 19       | 0        |
| Mitjana              | 49.926    | 132.91   | 83.084   | 25.846   | 18.667   |
| Mitjana              | 49.000    | 129.00   | 82.000   | 25.450   | 24.000   |
| Moda                 | 40.000    | 120.00   | 80.000   | 23.480   | 24.000   |
| Desviació estàndard  | 8.6769    | 22.422   | 12.056   | 4.1018   | 7.6778   |
| RIC                  | 15.000    | 26.500   | 15.000   | 5.0000   | 9.9131   |
| Mínim                | 32.000    | 83.500   | 48.000   | 15.540   | 0.0000   |
| Màxim                | 70.000    | 295.00   | 142.50   | 56.800   | 24.000   |
| Asimetria            | 0.19244   | 1.1479   | 0.74286  | 0.98066  | -1.1981  |
| Error est. asimetria | 0.036773  | 0.036773 | 0.036773 | 0.036852 | 0.036773 |
| Curtosi              | -1.0262   | 2.0865   | 1.3312   | 2.6028   | 0.039049 |
| Error est. curtosi   | 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   |

Ara podem comentar els resultats d'aquesta taula de forma genèrica. S'observa que tenim 4434 casos i que només tenim dades perdudes en *bmi1*, que ens podrien mostrar valors que falten en algun cas. Pel que fa a la mitjana, són valors que varien segons el tipus de dades, des de 18 a 132, però sent variables de configuració molt diferent. Les medianes i les modes, molt similars entre elles.

Aquesta taula també ens mostra les mesures de variabilitat de les variables, amb una desviació estàndard alta i similar i amb alt rang interquartílic. Per finalitzar amb la taula, podem veure que les dades estan distribuïdes de forma molt semblant a la normal (gaussiana).

En resum, la taula descriptiva ens pot mostrar molta informació inicial de les dades per saber com tractar-les amb posterioritat en les diferents anàlisis.

### Tractament gràfic

Pel que fa al tractament gràfic, podem veure a la figura 6 que, per a variables quantitatives, els gràfics descriptius poden ser els histogrames, densitats, gràfics Q-Q i diagrames de caixes. Els diagrames de densitat, també coneguts com a gràfics de densitat de kernel o gràfics de densitat de traçat, són eines valuoses per visualitzar la distribució de dades

continúes. A diferència dels histogrames, que presenten les dades en barres dividides en intervals, els diagrames de densitat ofereixen una representació més fluida i precisa de la forma de la distribució. Els gràfics Q-Q ens representen les dades per veure si les variables són o no normals, i els diagrames de caixes, també coneguts com a diagrames de caixes i bigotis o *boxplot*, són una eina gràfica estandarditzada per representar la distribució d'un conjunt de dades numèriques.

### Components d'un diagrama de caixes

- **Caixa.** La caixa principal del diagrama representa el rang interquartil (IQR) de les dades. L'IQR abasta el 50% central de les dades, excloent-hi els valors atípics.
- **Mediana.** La línia que divideix la caixa en dues meitats representa la mediana del conjunt de dades.
- **Bigotis.** S'estenen des de la caixa fins als valors mínim i màxim, sense incloure els valors atípics.
- **Límits dels bigotis.** Solen correspondre al primer i tercer quartil (Q1 i Q3), respectivament.
- **Valors atípics.** Els punts individuals que es troben més enllà dels bigotis es representen com a cercles o estrelles i es consideren valors atípics.

En les següents figures (7 i 8), podem veure els gràfics de les variables. En el cas de la figura 7, apreciem dues distribucions força diferents, tot i que en els dos casos compleixen normalitat. En el cas de *sysbp1*, es mostren valors atípics, marcats amb requadres de números en els diagrames de caixes, que en aquest cas són, per exemple, els punts 501 o 3649. Això ens permet apreciar possibles errors en les mostres o valors atípics que podrien (o no) ser eliminats prèviament a les anàlisis. També cal comentar que la simetria de les distribucions és molt diferent i amb apuntament més elevat de les dades de *sysbp1*, és a dir, més concentrats i amb menor variabilitat.

En el cas de la figura 8, on tenim tres variables, s'ha afegit una temporal, *timecvd*, que té un comportament similar en totes les variables temps del fitxer, per la qual cosa el seu comportament és similar, apareixent en l'histograma la densitat més gran en els valors més alts. Podem veure, també, que els histogrames de les variables *bmi1* i *diabp1* són molt similars i característics de les variables amb distribució gaussiana, però, encara que tinguin un apuntament similar, *sysbp1* té valors cap a la seva dreta, per la qual cosa sembla que el màxim està desplaçat a l'esquerra.

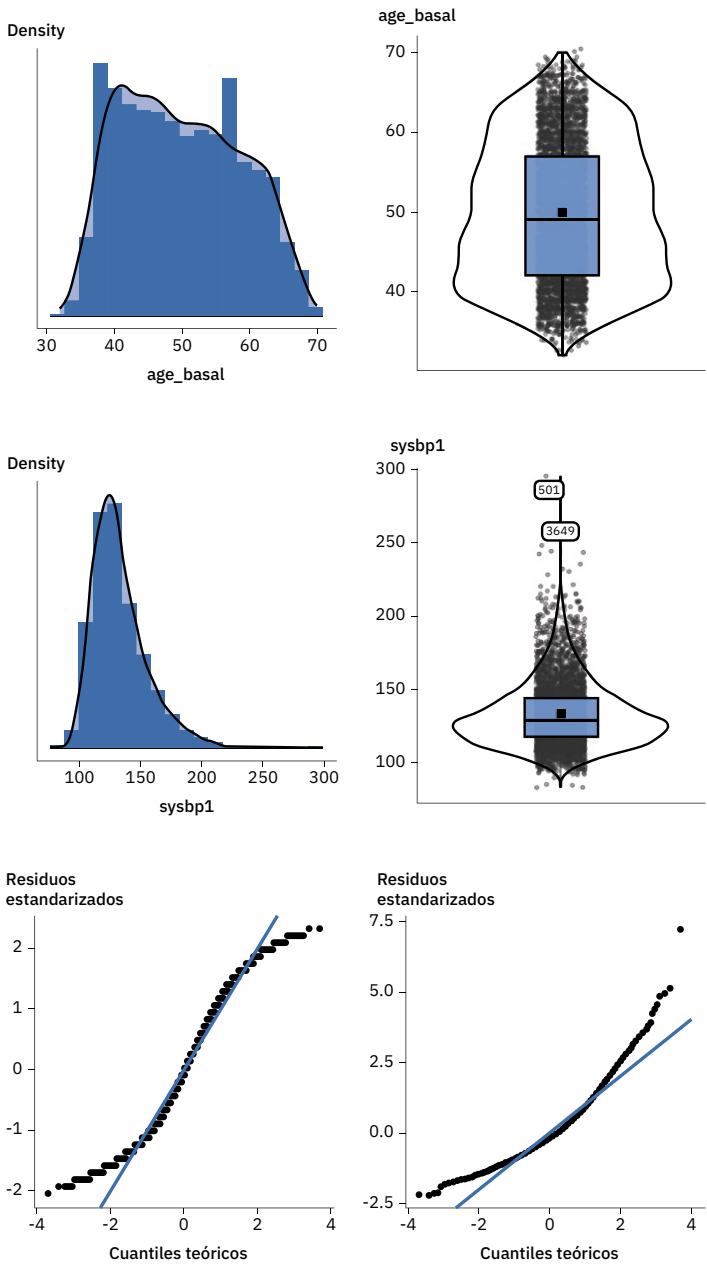


Figura 7. Freqüència de les variables dicotòmiques

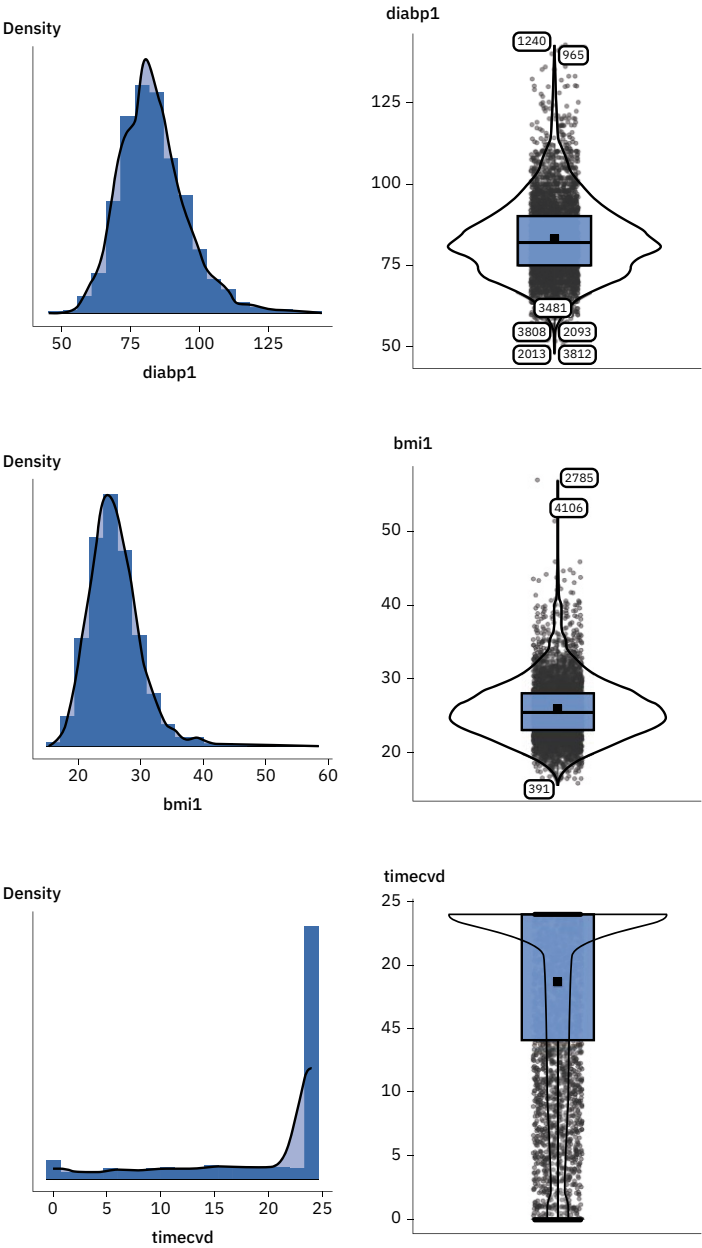


Figura 8. Diagrama Q-Q (esquerra) i diagrama de caixes amb dades i violi (dreta). Diabp1, Bmi1 y Timecvd

Si observem també els diagrames de caixa de la figura 8, podem veure que la variable *diabp1* té major nombre de dades atípiques, que podrien ser estudiades de forma separada.

En tots els diagrames de caixes, s'està introduint la màxima informació, com el punt negre central, que és la mitjana, la mida de caixa, el conjunt dels punts que són els valors, les dades atípiques i la densitat als dos costats en forma de violí, que ens mostra quins valors es repeteixen més.

Comentar finalment en aquestes gràfiques, més concretament en la figura 7, els gràfics que es mostren Q-Q de les variables *age\_basal* i *sysbp1*, per veure si aquestes dades es distribueixen de forma normal, és a dir, que les dades dels residus estan el més pròxim a la diagonal, quelcom que passa en els dos casos.

## Variables qualitatives

Per a aquesta segona part, les variables qualitatives que utilitzarem també seran les descrites en el capítol anterior com a variables dependents no contínues (*death*, *angina*, *hospmi*, *mi\_fchd*, *anychd*, *stroke*, *cvd* i *hypertens*). Podem veure que aquestes vuit variables només tenen dues opcions, si-no, per la qual cosa tindrem variables dicotòmiques, és a dir, nominals. D'altra banda, per prendre com a exemples variables qualitatives politòmiques, treballarem també amb dues variables ordinals: *age\_basal\_cat* i *Garrow*.

Pel que fa al procediment, és el mateix seguit per a les variables quantitatives.

## Variables dicotòmiques

Dins de les variables qualitatives, tenim les dicotòmiques, que podem veure les taules de les comentades amb anterioritat (taula 7).

**Taula 7. Freqüència de les variables dicotòmiques**

| Freqüències de death |             |             |            |
|----------------------|-------------|-------------|------------|
| death                | Freqüències | % del total | % acumulat |
| No                   | 2884        | 65.043 %    | 65.043 %   |
| Yes                  | 1550        | 34.957 %    | 100.000 %  |

**Freqüències de mi.fchd**

| mi.fchd | Freqüències | % del total | % acumulat |
|---------|-------------|-------------|------------|
| No      | 3703        | 83.514 %    | 83.514 %   |
| Yes     | 731         | 16.486 %    | 100.000 %  |

**Freqüències d'angina**

| angina | Freqüències | % del total | % acumulat |
|--------|-------------|-------------|------------|
| No     | 3709        | 83.649 %    | 83.649 %   |
| Yes    | 725         | 16.351 %    | 100.000 %  |

**Freqüències de hospmi**

| hospmi | Freqüències | % del total | % acumulat |
|--------|-------------|-------------|------------|
| No     | 3980        | 89.761 %    | 89.761 %   |
| Yes    | 454         | 10.239 %    | 100.000 %  |

**Freqüències d'anychd**

| anychd | Freqüències | % del total | % acumulat |
|--------|-------------|-------------|------------|
| No     | 3194        | 72.034 %    | 72.034 %   |
| Yes    | 1240        | 27.966 %    | 100.000 %  |

**Freqüències de stroke**

| stroke | Freqüències | % del total | % acumulat |
|--------|-------------|-------------|------------|
| No     | 4019        | 90.641 %    | 90.641 %   |
| Yes    | 415         | 9.359 %     | 100.000 %  |

**Freqüències de cvd**

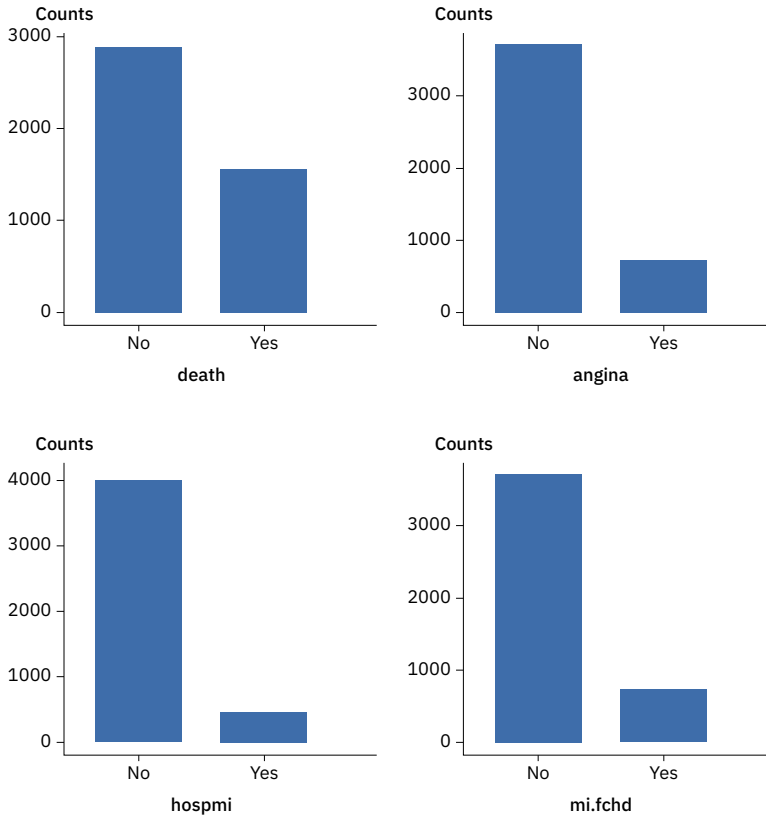
| cvd | Freqüències | % del total | % acumulat |
|-----|-------------|-------------|------------|
| No  | 327         | 73.906 %    | 73.906 %   |
| Yes | 1157        | 26.094 %    | 100.000 %  |

**Freqüències de hyperten**

| hyperten | Freqüències | % del total | % acumulat |
|----------|-------------|-------------|------------|
| No       | 1182        | 26.658 %    | 26.658 %   |
| Yes      | 3252        | 73.342 %    | 100.000 %  |

A les taules anteriors podem veure que totes aquestes variables són dicotòmiques, és a dir, dues possibilitats (No-Sí). A més, tenen diferents freqüències per als dos casos, no són homogènies.

Per a la seva millor comprensió, podem veure la forma dels diagrames de barres de cadascuna (figures 9 i 10). Com és normal en aquest tipus de dades, hi ha més casos de NO que de SÍ (que compleixin la condició).



**Figura 9. Diagrames de barres (I)**

Un diagrama de barres, també conegut com a gràfic de barres o gràfic de columnes, és una forma comuna de representar gràficament dades categòriques o discretes. Els seus components són:

- **Barres.** L'element principal del diagrama està format per barres rectangulars l'alçada o longitud de les quals és proporcional a la freqüència o magnitud de cada categoria.

- **Eix X.** L'eix horitzontal (X) representa les categories o variables qualitatives que s'estan analitzant.
- **Eix Y.** L'eix vertical (Y) representa la freqüència o magnitud de cada categoria. L'escala de l'eix Y ha de ser apropiada per als valors que s'hi representen.
- **Títols i etiquetes.** El diagrama ha d'incloure un títol clar que descrigui la variable que s'està representant, així com etiquetes en els eixos que identifiquin cada categoria.

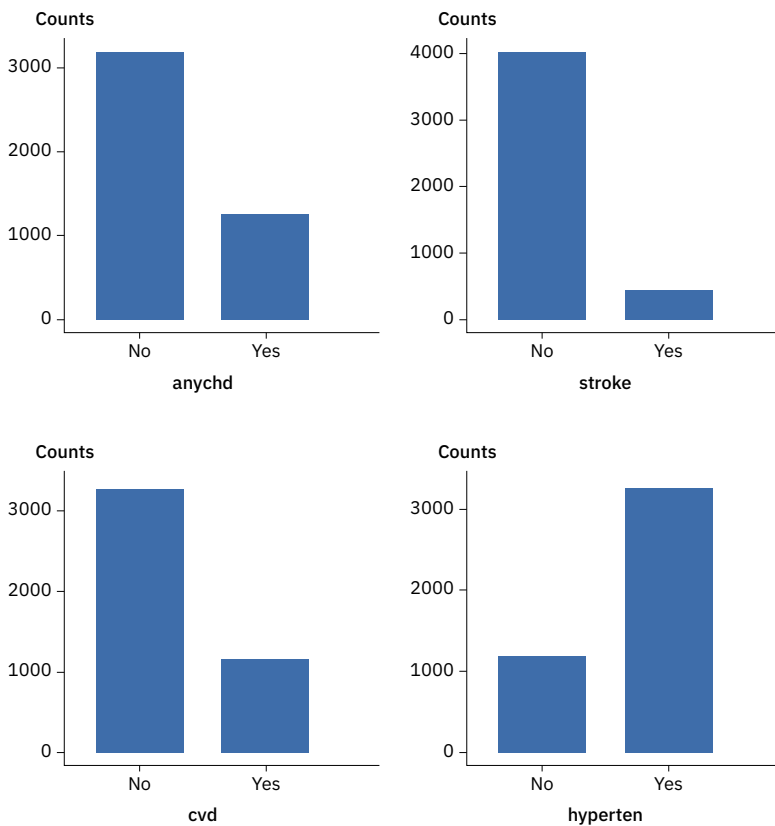


Figura 10. Diagrames de barres (II)

## Variables politòmiques

Per analitzar les variables qualitatives politòmiques, descriurem les variables *Garrow* i *age\_basal\_cat*, que tenen els seus valors ordenats. Veurem primer una descriptiva d'elles, en format taula, de la mateixa manera que s'ha efectuat amb les variables contínues. Tenim la descriptiva dels casos a la taula 3, on ens marca el nombre de perduts diferent de 0 per a cada cas de *Garrow*, quelcom normal pel tipus de dades. Les taules de freqüència (taula 8), ens mostren la distribució dels casos als *Garrow*, que són tres, i als *age\_basal\_cat*, que són quatre. En ambdós casos, ordenats per ser variables ordinals.

**Taula 8. Freqüències de les variables *Garrow* i *age\_basal\_cat***

### Freqüències de Garrow1

| Garrow1     | Freqüències | % del total | % acumulat |
|-------------|-------------|-------------|------------|
| Normoweight | 1993        | 45.142 %    | 45.142 %   |
| Overweight  | 1845        | 41.789 %    | 86.931 %   |
| Obese       | 577         | 13.069 %    | 100.000 %  |

### Freqüències de Garrow2

| Garrow2     | Freqüències | % del total | % acumulat |
|-------------|-------------|-------------|------------|
| Normoweight | 1744        | 44.298 %    | 44.298 %   |
| Overweight  | 1662        | 42.215 %    | 86.513 %   |
| Obese       | 531         | 13.487 %    | 100.000 %  |

### Freqüències de Garrow3

| Garrow3     | Freqüències | % del total | % acumulat |
|-------------|-------------|-------------|------------|
| Normoweight | 1462        | 45.040 %    | 45.040 %   |
| Overweight  | 1324        | 40.789 %    | 85.829 %   |
| Obese       | 460         | 14.171 %    | 100.000 %  |

### Freqüències d'*age\_basal\_cat*

| <i>age_basal_cat</i> | Freqüències | % del total | % acumulat |
|----------------------|-------------|-------------|------------|
| 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 ens mostra la distribució gràfica, mitjançant el diagrama de barres, de les diferents possibilitats d'aquestes dues variables ordinals, una més homogeniament distribuïda, l'*age\_basal\_cat*, i una altra amb diferència de casos, *Garrow*.

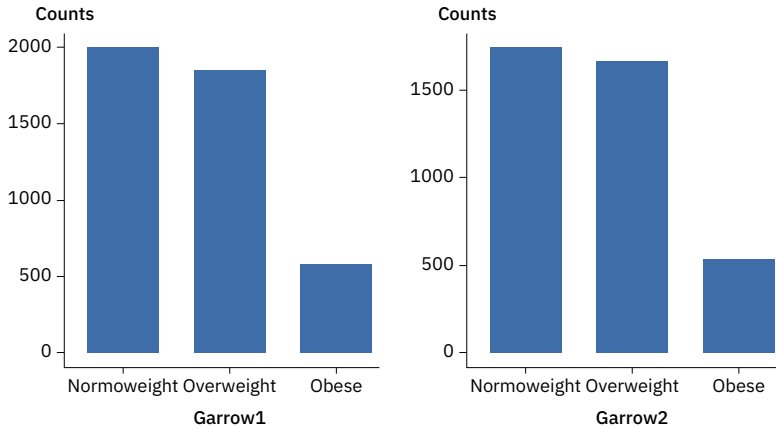


Figura 11. Diagrames de barres (III)

## Referències

- The Jamovi project (2022). *Jamovi*. (Versió 2.3) [Computer Software]. Recuperat de: <https://www.JAMOVI.org>
- R Core Team (2021). *R: A Language and environment for statistical computing*. (Versió 4.1) [Computer Software]. Recuperat de (R packages recuperat 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]. Recuperat de: <https://github.com/yanlinlin82/ggvenn>



# Capítol 5

## Descriptiva de dues o més variables

La tècnica d'estratificació resulta molt útil per treballar amb conjunts de dades. Estratificar consisteix a repetir l'anàlisi, en aquest cas descriptiu, tant de les taules com dels gràfics, per categories de variables qualitatives. Per exemple, el primer cas, tenim descrites les variables quantitatives *age\_basal*, *sysbp1*, *diabp1*, *bmi1*, separades segons si la variable *death* és No o Sí (taula 1). A continuació, els diagrames de caixes de les mateixes variables, que en aquest cas es denominen diagrames de caixes en paral·lel, igualment estratificat per la variable *death* (figura 1).

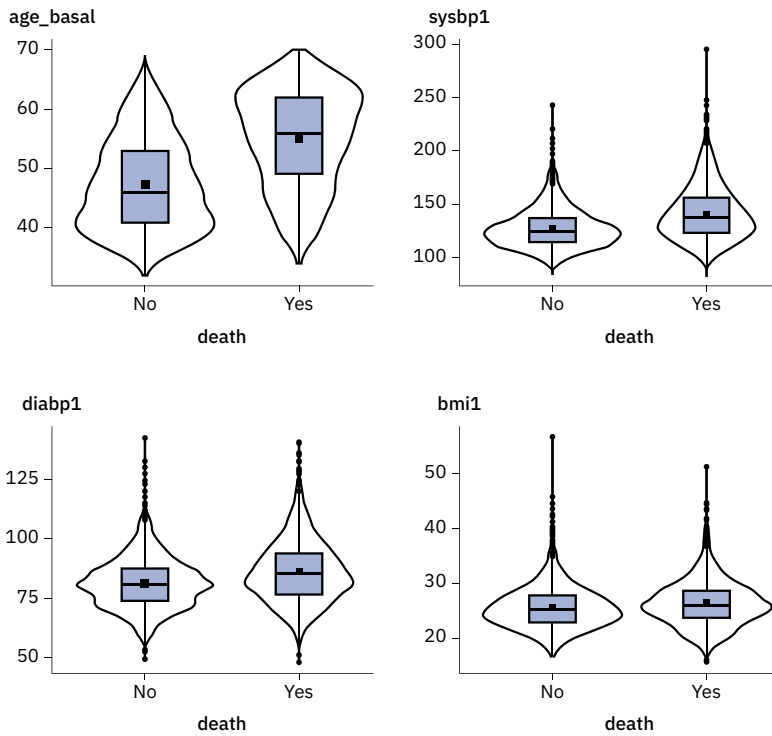
Taula 1. Variables

| Descriptives        |     | death | age_basal | sysbp1 | diabp1 | bmi1   |
|---------------------|-----|-------|-----------|--------|--------|--------|
| N                   | No  |       | 2884      | 2884   | 2884   | 2878   |
|                     | Yes |       | 1550      | 1550   | 1550   | 1537   |
| Mitjana             | 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 |
| Desviació estàndard | No  |       | 7.6705    | 18.538 | 10.795 | 3.9110 |
|                     | Yes |       | 8.2194    | 25.822 | 13.541 | 4.3829 |

Una altra possibilitat és no només separar o estratificar per una classe, sinó per dues. En aquest següent cas, les variables *age\_basal* i *timesysbp1hyp*, separades segons *death* i *stroke* alhora (figura 2).

En el cas de dues variables qualitatives, per exemple, el cas de *Garrow1* i *sex*, podem fer el diagrama de barres de la primera i veure com varia per la segona, a la part superior, o veure els casos en diagrames de barres de *Garrow1* segons la *age\_basal\_cat* (figura 3).

És interessant observar gràficament la variació en el comptatge dels casos de *Garrow1* en separar per gènere, ja que el cas d'*Overweight* té un significat diferent en comparació amb *Normoweight* i *Obese*, mentre que en el cas de la separació segons *age\_basal\_cat*, podem veure un comportament diferent de *Normoweight* respecte als altres dos casos.



**Figura 1. Diagrames de caixes en paral·lel, de les variables age\_basal, sysbp1, diabp1, bmi1, segons si la variable death**

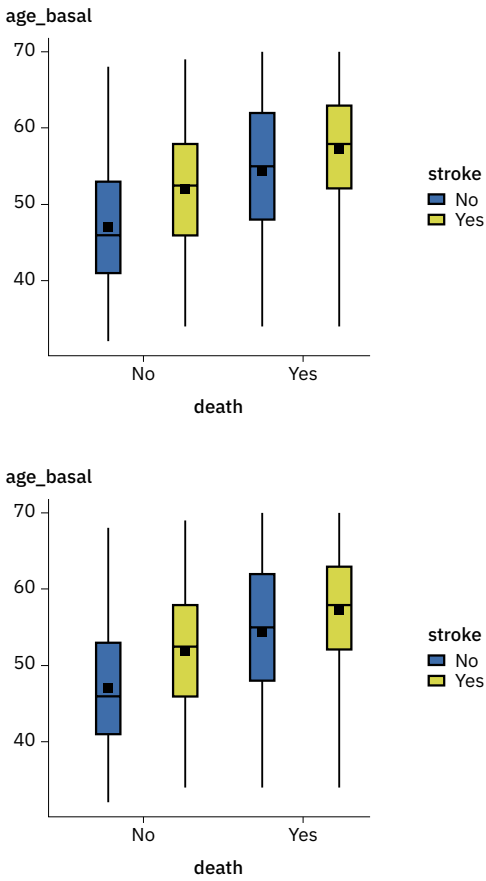


Figura 2. Diagrames de caixes en paral·lel (variables age\_basal i timesysbp1hyp, separades segons death i stroke)

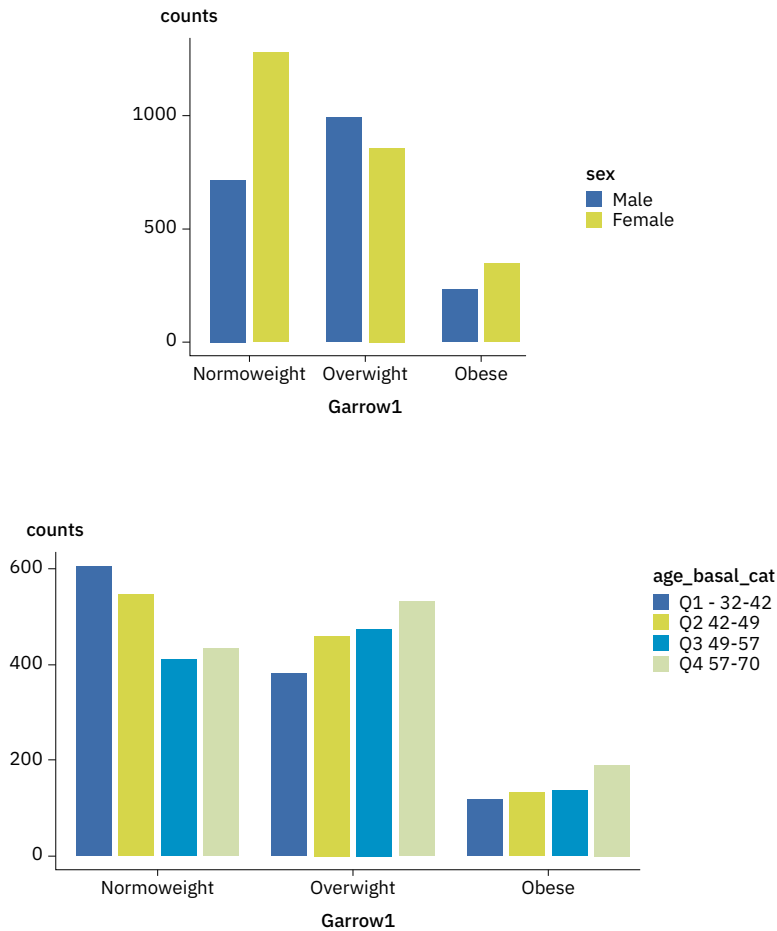


Figura 3. Diagrames de barres. Superior: Garrow 1, segons sexe. Inferior: Garrow1, segons age\_basal\_cat

## Altres tractaments gràfics

A més dels diagrames de barres, histogrames o gràfics de caixes ja mostrats, tenim altres possibilitats descriptives.

Una opció inicial és utilitzar diagrames de Venn per a les variables dicotòmiques i analitzar les relacions entre elles, sigui en quantitats o en percentatges.

Podem treballar amb dues, tres o quatre variables, com s'aprecia en els gràfics (figura 4).

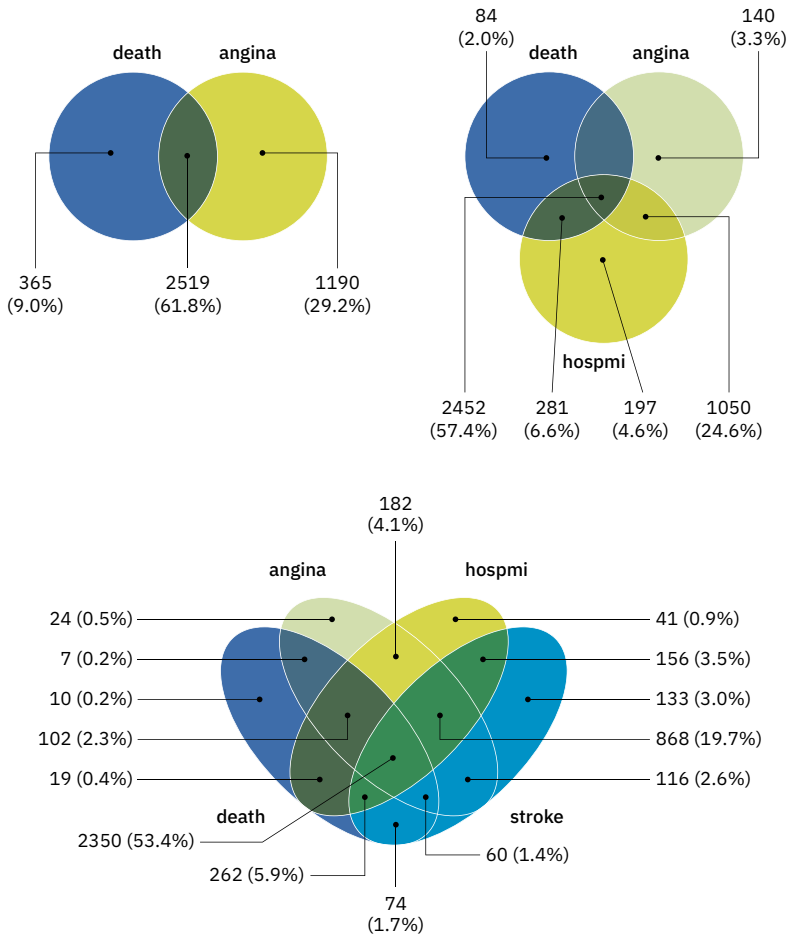


Figura 4. Diagrames de Venn

Aquests gràfics ens mostren tant les relacions com les quantitats i percentatges de casos, cosa que pot ajudar a veure la relació entre variables. Si ens centrem en els primers diagrames, veiem que el 61,8% dels casos del fitxer que tenien angines ha mort i que només un 9% dels que han mort no és per angines. En el següent diagrama de Venn, podem veure que en un 57,4% dels casos coincideixen les angines amb *hospmi*,

que han mort. Finalment, en l'últim cas es pot veure la relació en diagrames de Venn de quatre variables, on en un 53,4 % coincideixen totes.

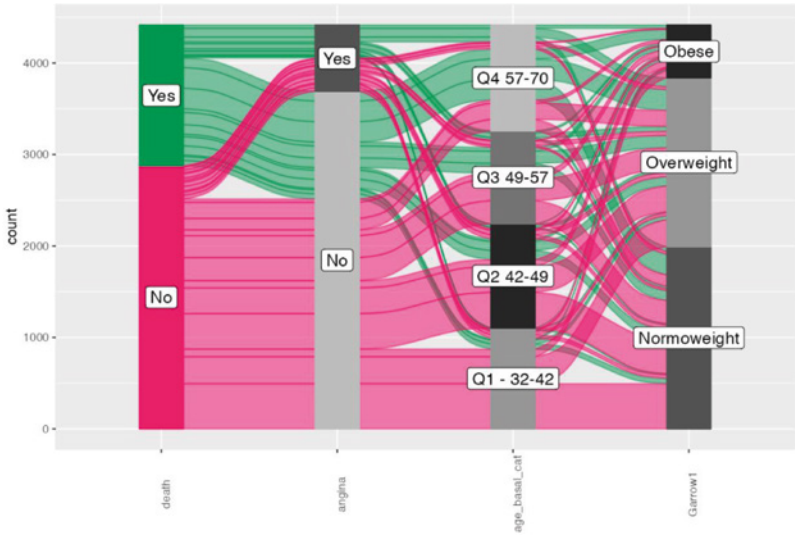


Figura 5. Diagrames Alluvial (I)

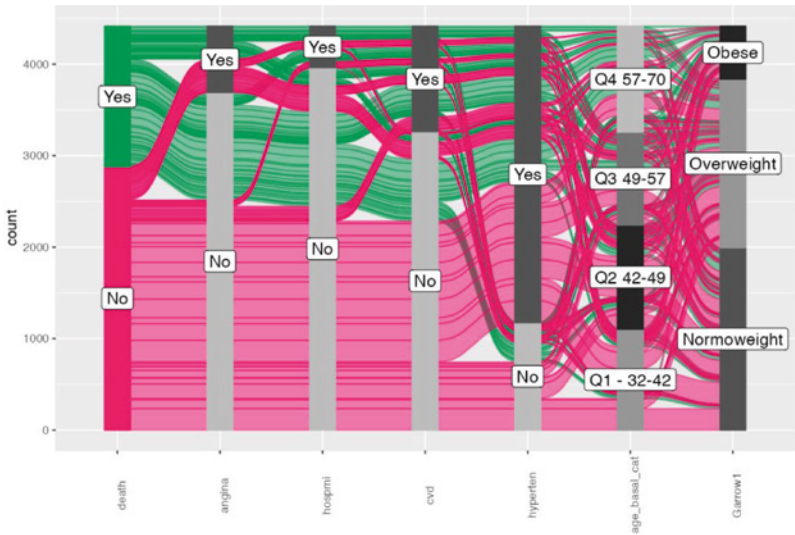


Figura 6. Diagrames Alluvial (II)

Una altra opció és el gràfic Alluvial. En el primer cas relacionem només quatre variables (figura 5) mentre que, en el següent, la relació és de més variables (figura 6).

Quina utilitat tindran, en la descriptiva de variables, els gràfics Alluvial? Ens permetran, entre altres coses, veure la quantitat de relacions diferents que hi ha entre variables i veure quina es repeteix més. Per exemple, el pas de *death* a *angina* mostra els casos (en verd) de *death* que són amb angines (Yes) i els que no tenen angines (No). A més, podem veure que tenim casos de *death* i *no death* tant per a gent amb angines com per a gent sense angines. En el supòsit que tinguéssim una relació entre dues variables en què un cas no es donés —per exemple, *no death* amb *no angina*— podria establir-se una colinealitat entre variables. En resum, es tracta d'un gràfic descriptiu entre variables qualitatives i en el qual es mostren relacions directes.

| 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   |

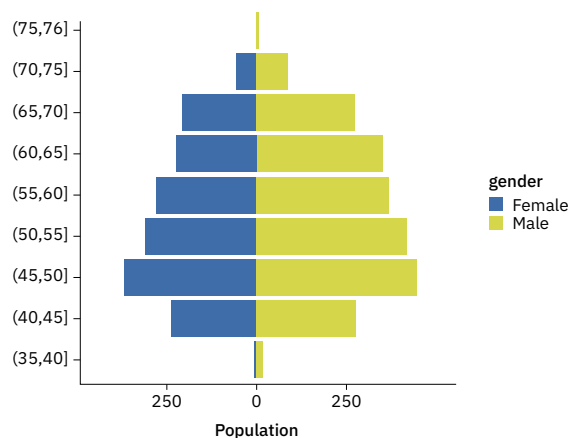


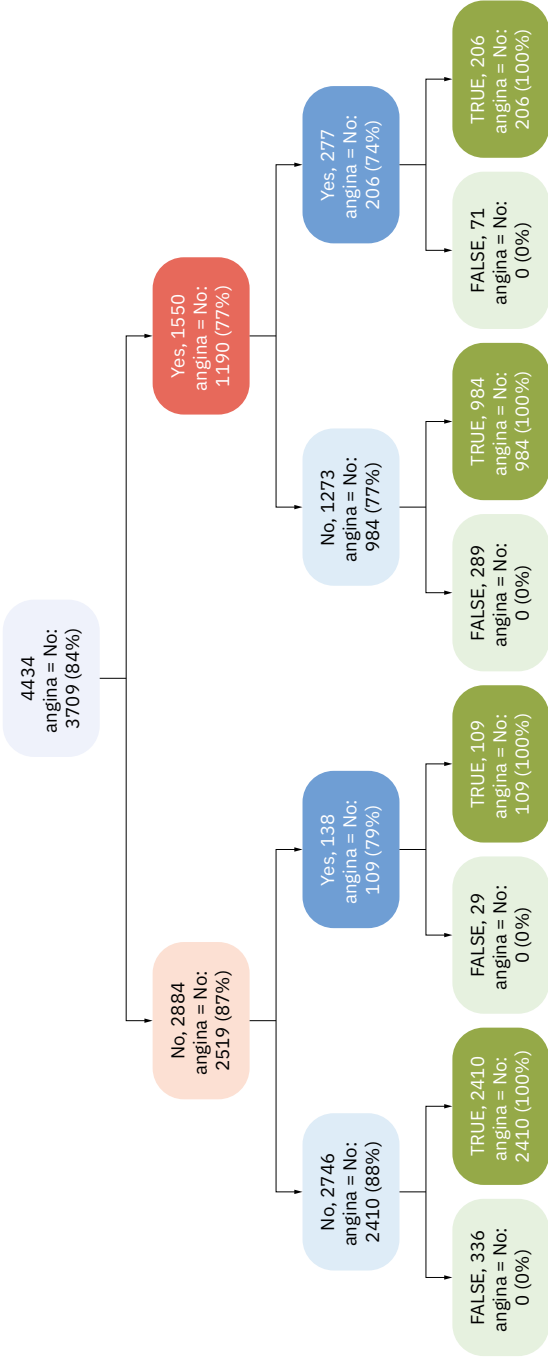
Figura 7. Diagrames piràmide poblacional

Finalment, es presenta un exemple de piràmide de població, amb la utilització de les dades de les variables *edat* i *sexe* de Framingham (figura 7), a més de l'arbre de relacions (figura 8).

Com a exemple final, s'observa a la figura 7, la població femenina supera la masculina en tots els rangs d'edats en aquest estudi, de forma general tenim 2490 dones i 1944 homes. En el cas de la figura 8, aquesta ens mostra com es distribueixen els casos, a partir d'*angina* (Yes-No) per a cada fase, gràfic que podem elaborar amb qualsevol de les variables tant dicotòmiques com polinòmiques. Amb això, hem vist algunes pinzellades de l'obtenció de gràfics descriptius de variables quantitatives i qualitatives que ens serviran com a pas previ a l'anàlisi més detallada de les dades que farem més endavant i que, a més, ens serveixen per conèixer millor les variables.

## Referències

- The Jamovi project (2022). *Jamovi*. (Versió 2.3) [Computer Software]. Recuperat de <https://www.JAMOVl.org>
- R Core Team (2021). *R: A Language and environment for statistical computing*. (Versió 4.1) [Computer programari]. Recuperat de (R packages recuperat de MRAN snapshot 2022-01-01): <https://cran.r-project.org>
- Serdar Balci (2022). *ClinicoPath Jamovi Module* doi:10.5281 zenodo.3997188. [R package]. Recuperat 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]. Recuperat de: <https://CRAN.R-project.org/package=easyalluvial>
- Nick Barrowman (2020). *vtree: Display Information About Nested Subsets of a Data Frame*. [R package]. Recuperat de: <https://CRAN.R-project.org/package=vtree>





## Capítol 6

# Introducció a la inferència amb una variable

En aquest capítol desenvoluparem la inferència estadística amb una variable. L'objectiu del capítol és guiar en la realització d'operacions amb Jamovi relatives a la inferència matemàtica o estimació de paràmetres poblacionals a partir de dades mostrals.

Primer es tractarà la construcció d'interval de confiança per a una variable i, a continuació, els contrastos d'hipòtesis d'una variable en relació amb un paràmetre teòric. En ambdós casos es distingirà entre tipus de variables diferents: estimació de proporcions per a variables qualitatives nominals/ordinals, estimació de mitjanes per a variables contínues).

Per a aquest propòsit, treballarem amb les variables contínues següents:

- **sysbp.** Pressió arterial sistòlica (mitjana dels dos últims de tres mesuraments) (mmHg).
- **diabp.** Pressió arterial diastòlica (mitjana dels dos últims de tres mesuraments) (mmHg).
- **imc.** Índex de massa corporal, pes en quilograms/altura en metres quadrats.
- A més de les ja comentades en capítols anteriors, com són *hyperten* i *angina*, que són variables qualitatives dicotòmiques, és a dir, que només tenen dos nivells, yes/no.

## Conceptes bàsics: població, mostra, estimació i contrastos

La **inferència estadística** és la part de l'estadística que tracta de les condicions i procediments mitjançant els quals podem extreure conclusions o inferències de la població a partir de la informació mostral. El seu objectiu és obtenir conclusions (estimacions, decisions, prediccions o altres generalitzacions) útils per fer raonaments deductius sobre una totalitat, basant-se en la informació numèrica donada per la mostra.

Donarem algunes definicions per als conceptes que acabem de nomenar.

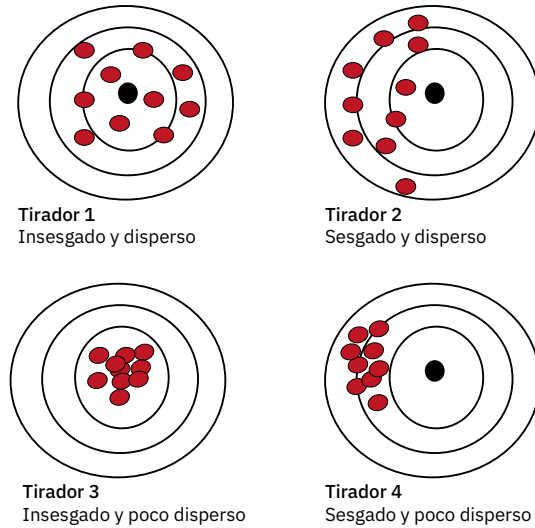
- **Univers.** És el conjunt d'individus o objectes que considerem en un estudi estadístic. Dins d'un mateix univers podem distingir diverses poblacions.
- **Població.** Són tots els valors que adopta la característica (o característiques) que volem estudiar en l'univers. Per exemple: si estem analitzant només una característica, l'edat dels espanyols, la població seria el conjunt de les edats de tots els espanyols i l'univers seria el conjunt de tots els espanyols.
- **Mostra.** És qualsevol subconjunt de la població. Perquè la mostra ens serveixi per extreure conclusions sobre la població haurà de ser tan representativa com sigui possible.
- **Mostreig (o tècniques de mostreig).** Són els mètodes per triar les mostres.

La inferència (**estadística**) pot prendre la forma d'estimacions d'unes característiques numèriques (**estimació**), respostes a preguntes sí/no o veritable/fals (**contrast o prova d'hipòtesis**), pronòstics de futures observacions, descripcions d'associació (o correlació) o modelització de relacions entre variables, així com anàlisi de la variància, sèries de temps i mineria de dades.

L'estimació es divideix en **estimació puntual** i **estimació per interval**.

L'**estimació puntual** consisteix a donar un valor numèric a un paràmetre poblacional desconegut (per exemple, la proporció poblacional, la mitjana poblacional, etc.). Tot i que hi ha infinits estimadors puntuals per a cada paràmetre poblacional només un d'ells és òptim. Un estimador òptim, anomenat ELIO, és un estimador **lineal** (obtingut com una combinació lineal), **no esbiaixat** (s'espera que el seu valor coincideixi amb el veritable valor del paràmetre poblacional) i de **variància mínima** (d'entre tots els estimadors no esbiaixats el que tingui mínima variància).

Il·lustrarem aquests conceptes amb la figura 1. Hi tenim les dianes de 4 tiradors sobre les quals han realitzat 10 trets cadascun d'ells. Si el paràmetre poblacional desconegut fos el centre de la diana i l'estimador cadascun dels trets, els tiradors 1 i 3, amb tots els trets al voltant del centre de la diana serien no esbiaixats, mentre que els tiradors 2 i 4, amb trets lluny de la diana serien esbiaixats. Entre els estimadors no esbiaixats el tirador 3 té menys dispersió, és a dir menys variància, i el tirador 1 major dispersió (variància). L'estimador ELIO seria, per tant, el tirador 3.



**Figura 1. Estimadors esbiaixats i no esbiaixats, amb diferent variància**

L'estimador òptim de la mitjana poblacional (representat per la lletra grega  $\mu$ ) és la mitjana mostral (representat per  $\bar{x}$ ) i el de la proporció poblacional (representat per la lletra grega  $\pi$ ) és la proporció mostral (representada per  $p$ ). És a dir, si disposéssim d'una mostra (representativa) d'una població amb informació del sexe i de l'edat d'un conjunt de subjectes, la proporció de dones a la mostra seria un estimador (òptim) de la proporció de dones en la població i l'edat mitjana dels subjectes un estimador de l'edat mitjana de la població. El que passa és que si disposéssim d'una altra mostra de la població (per exemple, obtinguda en un altre moment del temps), ni la proporció de dones a la mostra ni l'edat mitjana dels subjectes serien les mateixes (tot i que podrien semblar-se). Ara bé, si aquesta segona mostra fos representativa, els estimadors obtinguts a partir d'ella també serien òptims. Aquest fenomen es coneix com a **variabilitat mostral** que, teòricament, es defineix com la variació dels estimadors (puntuals) d'un paràmetre entre mostres representatives d'una població. De fet, si un estimador no té variabilitat mostral tenim un error. La variabilitat mostral es mesura mitjançant l'error estàndard.

Posarem un exemple per aclarir tots aquests conceptes. Suposem que tenim una població composta per 5 subjectes, 1, 2, 3, 4 i 5. La mitjana poblacional (que desconeixem) és, per tant, igual a 3

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

Per estimar aquest paràmetre poblacional farem totes les mostres de mida 2 (sense repetició) que puguem de la població (taula 1).

**Taula 1. Mostres de mida 2 de la població fictícia de l'exemple**

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

Com hem dit, l'estimador òptim de la mitjana poblacional és la mitjana mostral. Així que calculem les mitges mostrals de cadascuna de les mostres (taula 2). Així, per exemple, a la mostra (1,2), la mitjana mostral seria

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

**Taula 2. Mitjanes mostrals de les mostres de la població fictícia de l'exemple**

|       | $\bar{x}$ |       | $\bar{x}$ |       | $\bar{x}$ |       | $\bar{x}$ |
|-------|-----------|-------|-----------|-------|-----------|-------|-----------|
| (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 taula podem observar, en primer lloc, que les mitges mostrals varien molt, és a dir, que hi ha variabilitat mostral. En segon lloc, que excepte a les mostres (1,5) i (2,4), les mitjanes mostrals no són iguals que el valor de la mitjana poblacional (que, recordem-ho, era igual a 3). Llavors, què significa que la mitjana mostral sigui un estimador no esbiaixat de la mitjana poblacional? Si fem la mitjana de totes les mitges mostrals

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

És a dir, significa que si tinguéssim totes les mostres possibles d'una població el valor (mitjà) de l'estimador seria igual que el valor del paràmetre. Però mai es disposa de totes les mostres d'una població, sinó usualment d'una de sola (de vegades, amb sort de 2 o 3, però mai de totes).

Continuant amb el nostre exemple, si tinguéssim la mostra (1,5) o la (2,4), el valor de l'estimador coincidiria amb el veritable valor del paràmetre. Però, i si treballem amb la mostra (2,5), per exemple? Podem estar segurs que el valor de l'estimador (puntual) (igual a 3,5) sigui igual al valor del

paràmetre poblacional? No. L'únic que sabem és que 3,5 és un estimador no esbiaixat del veritable paràmetre poblacional.

Per això ens cal acompanyar l'estimació puntual de l'**estimació per interval**. Es tracta d'un rang de valors en el qual amb una determinada probabilitat (denominada grau de confiança) està contingut el paràmetre poblacional.

En el nostre exemple, si utilitzem la mostra (2,5), l'estimador puntual serà igual a 3,5 i l'estimador per interval al 95 % és igual a  $(-0,88, 7,88)$  (la fórmula de l'estimador per interval de la mitjana mostral l'explicarem més endavant). És a dir, amb un 95 % de probabilitat el veritable valor del paràmetre estarà situat entre  $-0,88$  i  $7,88$ . Observeu com l'interval és molt ampli. Això és així perquè la mida mostral és molt petita (igual a 2). Com més gran sigui la mida mostral, menor serà l'interval i com menor sigui l'interval major precisió tindrà l'estimador.

En aquest cas utilitzem dues mostres. Una, la mostra (2,5), que ja hem utilitzat, i la segona, una nova mostra formada per la unió de les mostres (1,2), (1,3) i (1,4), és a dir, mostra (1,2,3,4). L'estimador puntual de la mitjana poblacional utilitzant aquesta nova mostra és 2,5

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

i l'interval de confiança al 95 % és  $(-0,604, 5,604)$ . Podem plantejar-nos diverses qüestions:

- Per què l'amplitud de l'interval a la mostra (2,5)  $(7,88 - (-0,88) = 8,76)$  és més gran que a la mostra (1,2,3,4)  $(5,604 - (-0,604) = 6,208)$ ?
- Per què la mida de la mostra és menor (2 vs. 4).
- Quin estimador és més vàlid? el de la mostra (2,5), igual a 3,5, o el de la mostra (1,2,3,4), igual a 2,5?

Tots dos són igual de vàlids, el que passa és que el de la mostra (1,2,3,4) és més precís.

## Estadística descriptiva

Primerament, vegem com són les variables contínues de forma descriptiva, dades estadístiques i histogrames. Fem notar que aquestes variables també s’han obtingut en els tres períodes de l’estudi, per la qual cosa aplicarem inferència en els tres moments per a les tres variables.

Podem veure a la taula 3 que, mentre que BMI (IMC) es manté de forma mitjana constant en els tres temps, tant sysbp com *diabp* han variat, la primera augmentant en el temps i la segona amb disminució en l’últim període (baixada bastant significativa).

**Taula 3. Descriptiva de les variables en els tres moments temporals de l’estudi**

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

Podem veure en els histogrames de les tres variables (figura 2), que hi ha més simetria a *diabp* durant el temps i que les altres variables tenen asimetria cap a la dreta.

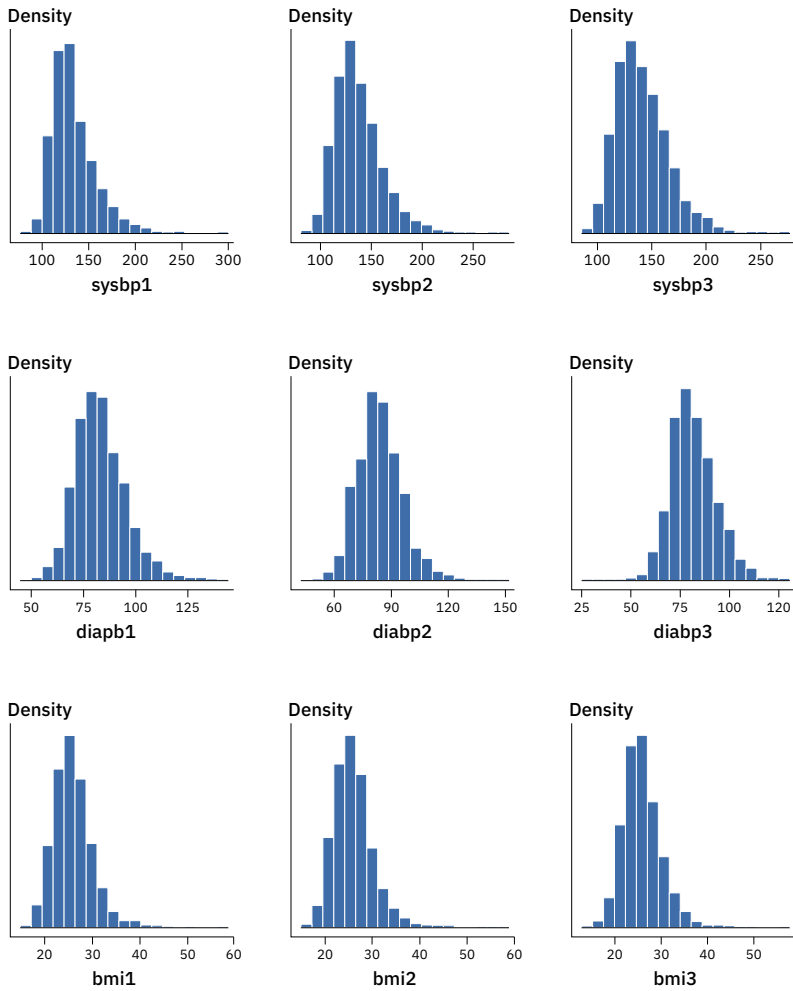


Figura 2. Histograma de les tres variables per a cada període de temps

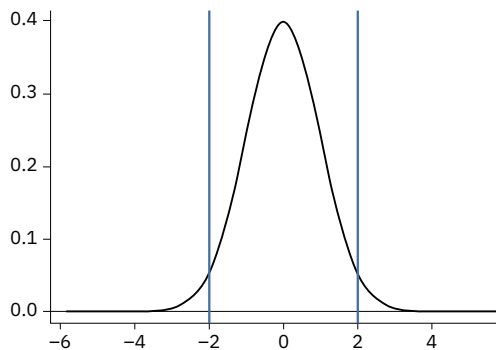
## Construcció d'un interval de confiança per a una proporció

Com a primer pas d'inferència, introduïm la construcció de l'interval de confiança per a una proporció. Per a això, utilitzarem les variables dicotòmiques *hyperten* i *angina com a* exemple.

La fórmula per calcular l'interval de confiança per a una proporció és la següent (seguint l'aproximació a la normal):

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

On  $p$  és la proporció mostral (l'estimador puntual de la proporció poblacional),  $n$  és la mida de la mostra i  $z_{\alpha/2}$ ,  $z_{(1-\alpha/2)}$ , són els percentils  $\alpha/2$  i  $(1 - \alpha/2)$  de la distribució normal tipificada ( $N(0, 1)$ ) (vegeu figura 3), sent  $\alpha = 1 - \text{grau de confiança}$  (en tant per un).



**Figura 3. Distribució normal tipificada, en la qual s'assenyalen els percentils 0,025 (-1,959) i 0,975 (1,959), que corresponen a un grau de confiança del 95 %**

Així, si treballem al 95 % de confiança (l'habitual),  $\alpha = 1 - 0,95 = 0,05$  i  $\alpha = 0,05 / 2 = 0,025$  que es correspon amb el valor -1,959 per a  $z_{\alpha/2}$  i a l'1,959 per al percentil  $z_{(1-\alpha/2)}$ .

Els passos per seguir a Jamovi a fi d'obtenir els resultats s'inicien primer amb triar «Anàlisi» (figura 4), on seleccionarem «Prova per a la proporció en una mostra» i «2 Resultats. Prova binomial», seleccionar les variables (figura 4) i, finalment, decidir quins resultats volem obtenir (figura 5).

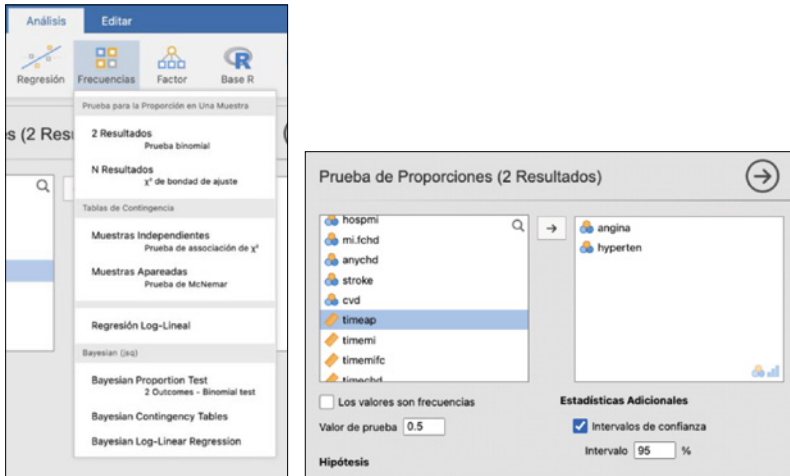


Figura 4. Obtenció de proporcions (I)

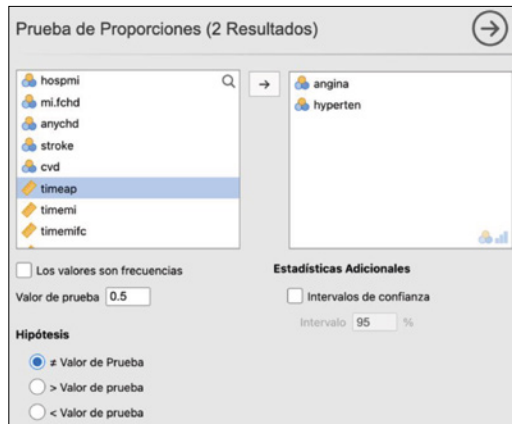


Figura 5. Obtenció de proporcions (II)

Els resultats obtinguts en utilitzar les dues variables binomials apareixen a la taula següent.

Taula 4. Resultats de la prova de proporcions d'una variable, per al cas d'angina i hyperten

Prova binomial

|          |        |            |       |           |          | Interval de confiança al 95 % |          |
|----------|--------|------------|-------|-----------|----------|-------------------------------|----------|
|          | Nivell | Freqüència | Total | Proporció | 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.  $H_\alpha$  és proporció  $\neq 0.5$

Obtenim el nombre de casos per a cada possibilitat, total de la mostra, proporció de cadascun dels dos casos i el p-valor per al contrast d'aquesta proporció, concepte que veurem més endavant. A més, podem veure l'estimació puntual i per interval. Per exemple, el resultat ens diu que la proporció d'hipertensos és de 73,34 %, estimador puntual de la proporció d'hipertensos a la població. D'altra banda, amb un 95,00 % de confiança, la proporció d'hipertensos en la població està entre el 72,01 % i el 74,64 %. La proporció de subjectes que han patit angina de pit és del 16,35 % i, amb confiança al 95,00 %, la proporció està entre el 15,27 % i el 17,47 %. Observeu com l'amplitud dels intervals és força petita, per la qual cosa els estimadors són bastant precisos. Podem canviar el nivell de confiança, per exemple, al 99,00 %, com s'observa en la figura 5.

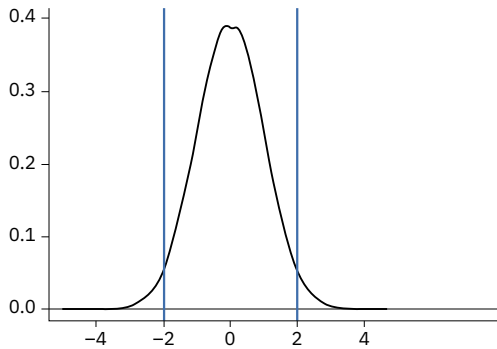
## Construcció d'un interval de confiança per a la mitjana

A continuació, veurem com construir un interval de confiança per a una mitjana. Es tracta d'un cas que implica inferència d'una variable del tipus contínua (o dels tipus escala o numèric).

La fórmula per construir un interval de confiança per a una mitjana és la següent:

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

On  $\bar{x}$  és la mitjana mostral (l'estimador puntual de la mitjana poblacional),  $s$  és la desviació típica,  $n$  és la mida de la mostra i  $t_{n-1, \alpha/2}$ ,  $t_{n-1, (1-\alpha/2)}$  són els percentils  $\alpha/2$  i  $(1 - \alpha/2)$  de la distribució t de Student amb  $n-1$  graus de llibertat, respectivament (vegeu figura 6), sent  $\alpha = 1 - \text{grau de confiança}$  (en tant per un).



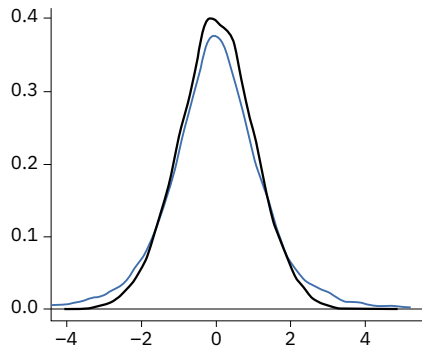
**Figura 6. Distribució t de Student amb 4433 graus de llibertat ( $4433 - 1$ ), en els quals s'assenyalen els percentils 0,025 ( $-1,960$ ) i 0,975 ( $1,960$ ), que corresponen amb un grau de confiança del 95 %**

En aquest cas, si utilitzem la mostra de la base de dades de Framingham, amb  $n = 4434$  subjectes, treballem al 95 % de confiança (l'habitual),  $\alpha = 1 - 0,95 = 0,05$  i  $\alpha / 2 = 0,025$ , i, per simetria  $(1 - \alpha / 2) = 0,975$ , es correspon amb els valors  $-1,960$  per a  $t_{4434-1,0,025}$  i,  $1,96$  a  $t_{4434-1,0,975}$ .

La distribució t de Student és més apuntada que la distribució normal i com més graus de llibertat, més apuntada. Així, en la figura 7, la corba negra correspon a una distribució amb 100 graus de llibertat i en blau a una distribució amb 4 graus de llibertat.

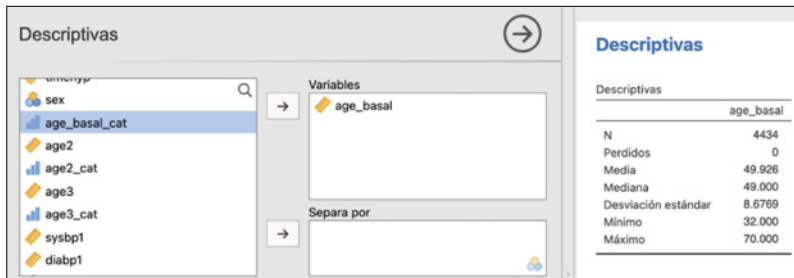
Abans de construir l'interval de confiança és convenient comprovar que la variable no tingui valors impossibles. Per a això, en l'anàlisi descriptiva de la variable demanarem el màxim i el mínim i comprovarem que tots els valors siguin lògics i, en cas contrari, definirem com a perduts els valors impossibles.

Així, per exemple, seleccionem a Jamovi «Anàlisi», «Exploració» i «Descriptius» i comprovem els valors impossibles de l'edat en el moment d'inclusió en la cohort (*age\_basal*) (figura 8).



**Figura 7. Distribucions t de Student amb 100 graus de llibertat (en negre) i amb 4 graus de llibertat (en blau)**

En la figura 8 veiem que l'edat màxima és de 165 anys, valor impossible. El que fem és eliminar el valor 165 (és a dir, fer-ho *missing*) i repetir el procediment per veure si hi ha altres valors impossibles (figura 9). Com veiem, no sembla haver-hi valors impossibles per *age\_basal*.



**Figura 8. Alguns descriptius de *age\_basal***

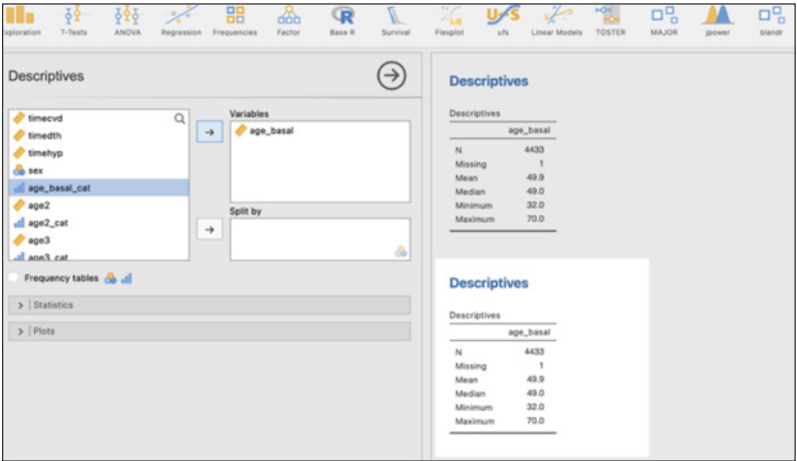


Figura 9. Alguns descriptius de age\_basal, després de fer faltant (missing) el valor 165 anys

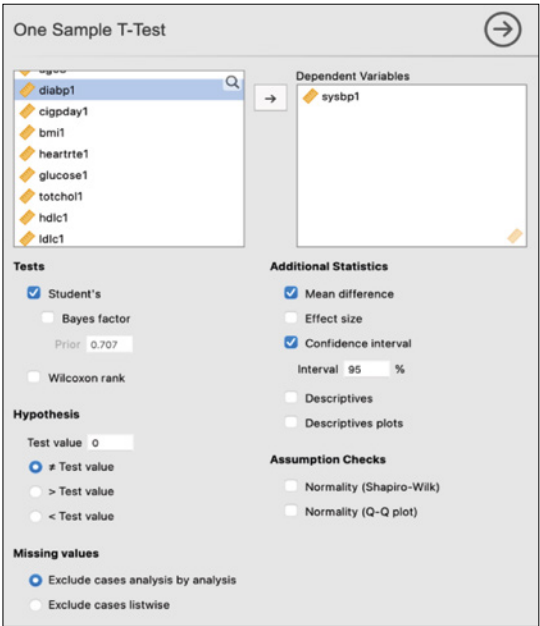


Figura 10. Alguns descriptius de age\_basal, després de fer faltant (missing) el valor 165 anys

Després de comprovar que no existeixen valors impossibles per a les variables *sysbp*, *diabp* i *imc* en les seves tres observacions, construirem, utilitzant Jamovi, intervals de confiança per a les seves mitjanes.

Seleccionarem «Anàlisi», «Prova T en una mostra» (figura 10).

**Taula 5. Resultats de la prova d’inferència per a la mitjana de tres variables en tres moments diferents**

**Prova T en una mostra**

|        |              |            |        |          |                        | Interval de confiança al 95% |          |
|--------|--------------|------------|--------|----------|------------------------|------------------------------|----------|
|        |              | Estadístic | gl     | p        | Diferència de mitjanes | 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 els resultats de la taula 5 podem veure que, amb un 95% de confiança, la mitjana de *sysbp1* està entre 132,24 i 133,56. El mateix podem dir per a cadascuna de les variables que apareixen a la taula. Tenim els valors de l’estadístic, els graus de llibertat, la mitjana i el seu interval de confiança corresponent.

## Contrastos d’hipòtesis

Els intervals de confiança es poden emprar per contrastar hipòtesis. Per exemple, utilitzant els resultats de la taula 4 podríem contrastar si la prevalença poblacional d’angina de pit és igual al 20%. Utilitzant la base de dades Framingham com a mostra, estímem que la proporció (mostral) de l’angina de pit, és a dir, la prevalença de l’angina a la mostra és igual a 16,351%, amb un interval de confiança al 95% igual a (15,274%, 17,473%). El 20% no està contingut en l’interval. És a dir, amb un 95% de probabilitat, rebutjarem que la prevalença d’angina de pit

per a aquesta població sigui igual al 20 %. No obstant això, no és habitual utilitzar els intervals de confiança per realitzar contrastos d'hipòtesis. El més habitual és utilitzar contrastos d'hipòtesis per si sols.

Una prova de contrast d'hipòtesis (també denominat contrast d'hipòtesis o prova de significació) és un procediment per jutjar si una propietat que se suposa en una població estadística és compatible amb l'observat en una mostra d'aquesta població. Una altra definició, en la qual utilitzarem diversos conceptes bàsics de contrastació d'hipòtesis, és que un contrast calcula la probabilitat que els resultats obtinguts en una investigació puguin ser deguts a l'atzar en el supòsit que la hipòtesi nul·la sigui certa.

Teòricament, el procés de contrast es divideix en diverses etapes:

### 1. Plantejament de la hipòtesi nul·la ( $H_0$ ) i de la hipòtesi alternativa ( $H_A$ o $H_1$ ).

Una hipòtesi és una determinada conjectura que pot prendre un paràmetre poblacional.

Continuant amb el nostre exemple,  $\pi = 20\%$ . Observeu que utilitzem la nomenclatura poblacional, ja que volem contrastar si la prevalença poblacional d'angina de pit és igual al 20 %. Concretament,

$$H_0: \pi = 0$$

$$H_A: \pi \neq 0$$

Si en la hipòtesi nul·la els símbols són igual (=), com en el nostre cas, el contrast es denomina **bilateral** o a **dues cues**. Si tingués un símbol diferent igual, es denominaria **unilateral** o a **una cua**. Que seria aquest cas:

$$H_0: \pi > 0$$

$$H_A: \pi \leq 0$$

Habitualment es treballa amb contrastos bilaterals o a dues cues, com farem nosaltres.

Des del punt de vista científic, ens interessa rebutjar la hipòtesi nul·la.

### 2. Decisió d'un estadístic que resumeixi adequadament la informació mostral del paràmetre d'interès.

Aquest estadístic, denominat **estadístic de contrast**, és l'estimador ELIO del paràmetre. En el nostre cas, l'estimador ELIO de la proporció poblacional ( $\pi$ ) és la proporció mostral ( $\hat{p}$ ).

### 3. Divisió de l'espai mostral en dues regions no encavalcades, la regió crítica i la regió d'acceptació.

Aquest apartat es divideix en tres subetapes:

#### 3a. Representació gràfica de la distribució mostral de l'estadístic.

En el nostre cas, la proporció mostral es distribueix com una distribució khi quadrat que, com veiem en la figura 11, és asimètrica.

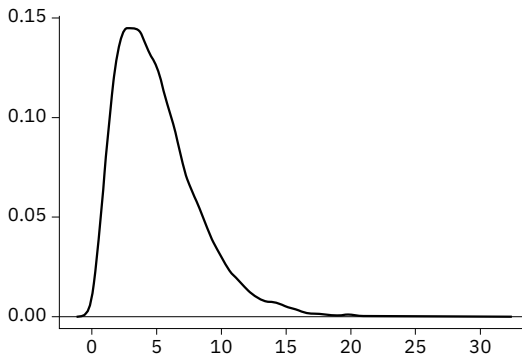


Figura 11. Distribució khi quadrat

#### 3b. Elecció del risc del contrast.

El risc del contrast, també denominat **nivell de significació**, és la probabilitat de rebutjar la hipòtesi nul·la quan aquesta és certa. Aquests conceptes es poden explicar millor a la següent taula (taula 6).

Taula 6. Quadre de decisió

|                           |       | Decisió        |               |
|---------------------------|-------|----------------|---------------|
|                           |       | No rebutjar    | Rebutjar      |
| Hipòtesi nul·la ( $H_0$ ) | Certa | Correcte       | Error tipus I |
|                           | Falsa | Error tipus II | Correcte      |

L'**error tipus I** (també denominat **falsos positius**) es comet en rebutjar la hipòtesi nul·la sent certa. La probabilitat de cometre error tipus I es denomina **risc**, **nivell de significació**, **nivell alfa** o  $\alpha$ . A més, el grau de confiança no és més que  $1 - \alpha$ .

L'**error tipus II** (també denominat **falsos negatius**) es comet perquè no es rebutja la hipòtesi nul·la sent falsa. La probabilitat de cometre error tipus II es denomina **beta**,  $\beta$ . La **potència del contrast** és igual a  $1 - \beta$ . És a dir, la potència del contrast és la probabilitat de rebutjar la hipòtesi nul·la sent falsa.

Se sol treballar amb un risc del 5 % (grau de confiança del 95 %), que es correspon amb un  $\beta = 20\%$  i una potència del 80 %.

### 3c. Dibuixar la regió crítica i la regió d'acceptació.

Sobre la representació de la distribució mostral de l'estadístic (en el nostre cas khi quadrat) es dibuixen els percentils  $\alpha = 0,05 / 2 = 0,025$  i  $(1 - \alpha / 2) = 0,975$ . El risc es divideix entre 2, perquè el contrast és a dues cues. En el nostre cas, vegeu la figura 12.

Notem que, en ser una distribució asimètrica, les regions de rebuig no són d'igual grandària. Això no passaria quan la distribució fos simètrica (com una t de Student).

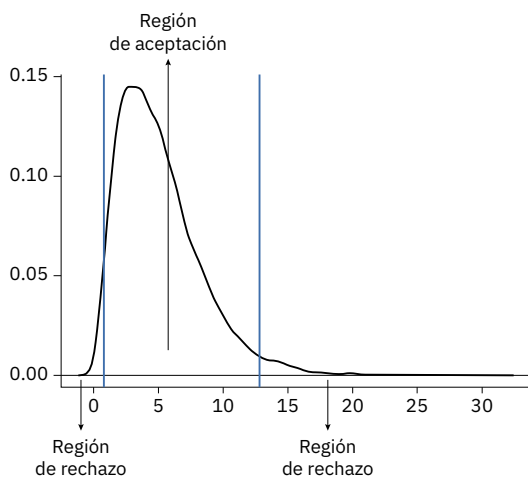
### 4. Obtenció d'una mostra per mesurar el paràmetre d'interès.

En el nostre exemple, la base de dades de Framingham.

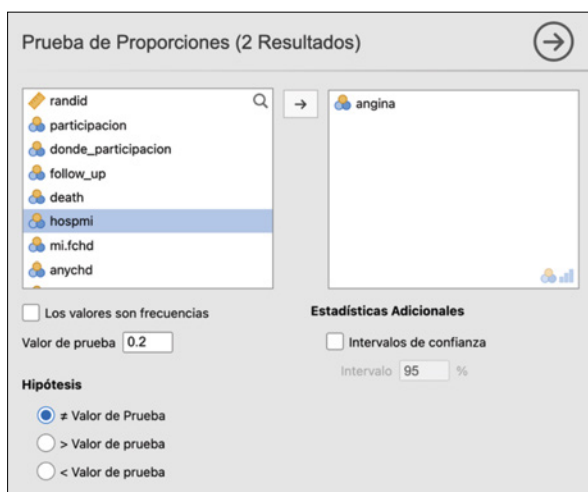
### 5. Càlcul de l'estadístic de contrast en la mostra escollida.

Quan es contrasten proporcions (una o més d'una), el contrast de l'estadístic és el contrast khi quadrat.

A Jamovi, seleccionem «Anàlisi», «Freqüències», «Prova per a la proporció en una mostra» i «2 Resultats. Prova binomial». A «Valor del contrast» posar 0.2 (el 20 % en tant per un) i assenyalar «Interval de confiança al 95 %» (figura 13).



**Figura 12.** Distribució khi quadrat en la qual hem dibuixat les regions d'acceptació i de rebuig



**Figura 13.** Contrast d'hipòtesis d'una proporció a Jamovi

El resultat del contrast està en la figura 14.

### Prova de proporcions (2 resultats)

#### Prova binomial

|        | Nivell | Freqüència | Total | Proporció | p       |
|--------|--------|------------|-------|-----------|---------|
| angina | No     | 3709       | 4434  | 0.83649   | <.00001 |
|        | Yes    | 725        | 4434  | 0.16351   | <.00001 |

Nota.  $H_a$  és proporció  $\neq 0.2$

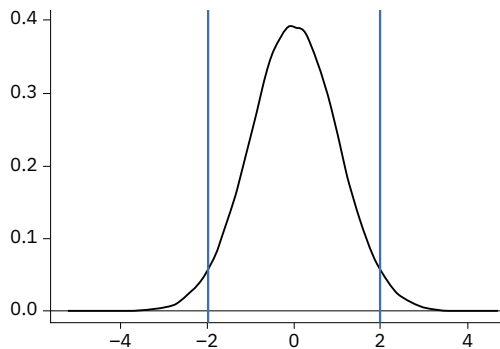


Figura 14. Resultat del contrast d'hipòtesis d'una proporció a Jamovi

#### 6. Resolució del contrast.

La resolució es podria fer gràficament (figura 15). Si el valor mostral de l'estadístic de contrast se situa en algunes de les regions de rebuig (com en el nostre cas), es rebutjarà la hipòtesi nul·la.

Per contra, si el valor mostral de l'estadístic de contrast se situa a la regió d'acceptació, no es rebutja la hipòtesi nul·la.

Numèricament, en el passat s'utilitzaven les taules estadístiques, com la que mostrem a la taula 7.

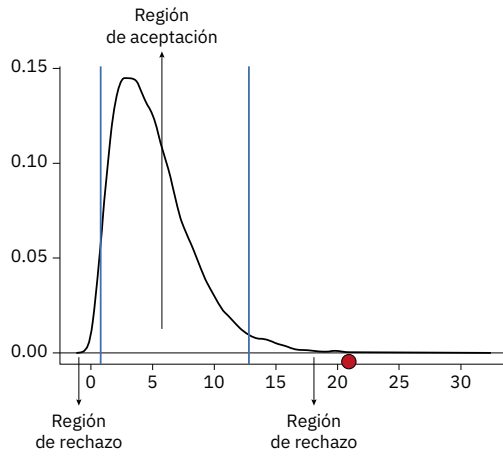


Figura 15. Resolució del contrast. Rebuig de la hipòtesi nul·la

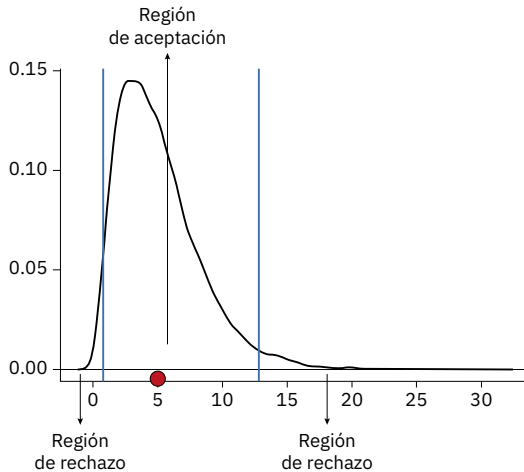


Figura 16. Resolució del contrast. No rebuig de la hipòtesi nul·la

Però és molt més còmode utilitzar el **p-valor** (o valor p). El p-valor podria ser definit com la probabilitat d'observar els resultats de l'estudi, o altres més allunyats de la hipòtesi nul·la, si la hipòtesi fos certa.

**Taula 7. Part dels valors crítics de la distribució khi quadrat segons els graus de llibertat (GL) i el risc**

| GL | $\alpha = 0,10$ | $\alpha = 0,05$ | $\alpha = 0,01$ |
|----|-----------------|-----------------|-----------------|
| 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 contrast en particular té associat el seu p-valor. En el nostre cas, observant la figura 14, el p-valor del contrast és  $p < 0,001$ , és a dir, menor de l'1 per 1000. En termes pràctics, si el p-valor és menor que el risc, es rebutja la hipòtesi nul·la i, en cas contrari, no es rebutja la hipòtesi nul·la. Així doncs, com  $0,001 < 0,05$  (assumim un risc del 5%), rebutgem la hipòtesi nul·la. És a dir, al 95 % de confiança, no podem dir que la prevalença d'angina de pit a la població sigui el 20 %.

Cal assenyalar que el p-valor només serveix per rebutjar o no la hipòtesi nul·la. No proporciona cap informació sobre la significació estadística. Per exemple, tenim dos contrastos d'una proporció, un amb un p-valor igual a 0,03 i un altre amb un p-valor igual a 0,01. En els dos casos rebutgem la hipòtesi nul·la, però en cap cas podem dir que el segon contrast és més significatiu que el primer.

## Contrast d'una mitjana

En aquest cas, també podem fer una anàlisi sobre un valor concret de la mitjana, ja que aquí hem comparat en tots ells amb el valor 0 com a valor de la prova per a la mitjana. Per exemple, podem veure si el cas de l'índex de massa corporal observat en el moment 3 (variable *bmi3*) es

diferència amb un valor concret, que podria ser 24,9 a partir del qual es pressuposa sobrepès.

Tornem a plantejar les etapes de contrastació. A la pràctica, no solen seguir-se totes.

**1. Plantejament de la hipòtesi nul·la ( $H_0$ ) i de la hipòtesi alternativa ( $H_A$  o  $H_1$ ).**

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

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

Es tracta, com abans, d'un contrast bilateral o a dues cues.

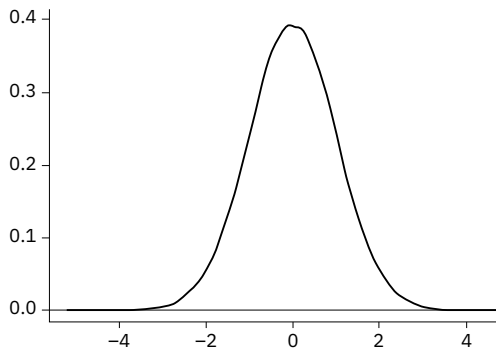
**2. Decisió d'un estadístic que resumeixi adequadament la informació mostral del paràmetre d'interès.**

L'estimador ELIO de la mitjana poblacional ( $\mu$ ) és la mitjana mostral ( $\bar{x}$ ).

**3. Divisió de l'espai mostral en dues regions no encavalcades, la regió crítica i la regió d'acceptació.**

**3a. Representació gràfica de la distribució mostral de l'estadístic.**

En el nostre cas, la mitjana mostral es distribueix com una distribució t de Student.

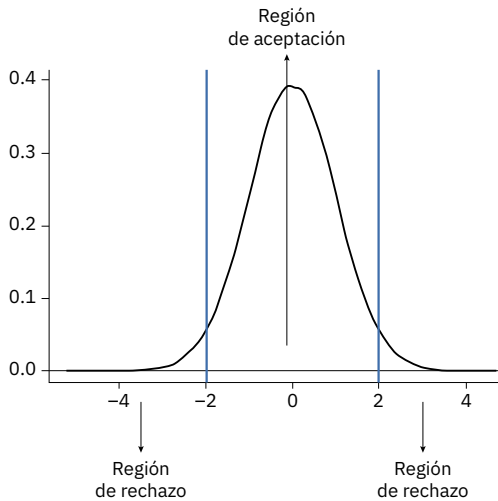


**Figura 17. Distribució t de Student**

### 3b. Elecció del risc del contrast.

Treballarem amb un risc del 5 %.

### 3c. Dibuixar la regió crítica i la regió d'acceptació.



**Figura 18. Distribució t de Student en la qual hem dibuixat les regions d'acceptació i de rebuig**

Vegeu que, en ser una distribució simètrica, les regions de rebuig són d'igual grandària.

### 4. Obtenció d'una mostra per mesurar el paràmetre d'interès.

En el nostre exemple, la base de dades de Framingham.

### 5. Càlcul de l'estadístic de contrast en la mostra escollida.

Quan es contrasten proporcions (una o més), el contrast de l'estadístic és el contrast t de Student.

En Jamovi, seleccionem «Anàlisi», «Contrastos t» i «Prova t en una mostra». A «Valor del contrast» posar 24.9 i assenyalar «Interval de confiança al 95 %» (figura 19).

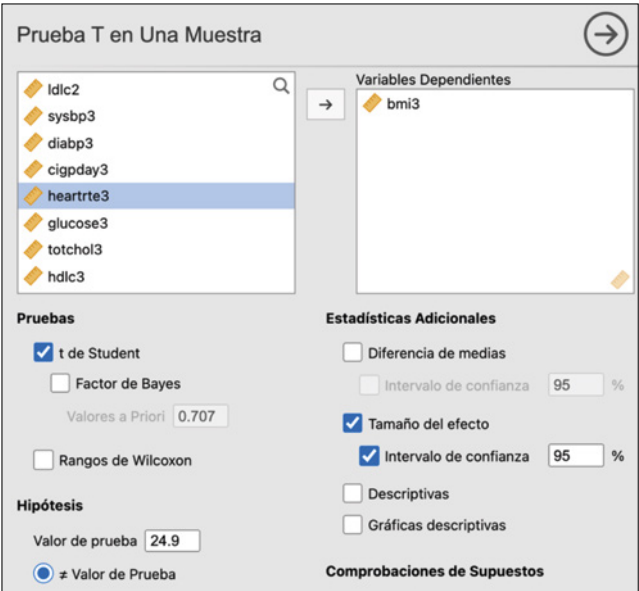


Figura 19. Contrast d’hipòtesis d’una mitjana a Jamovi

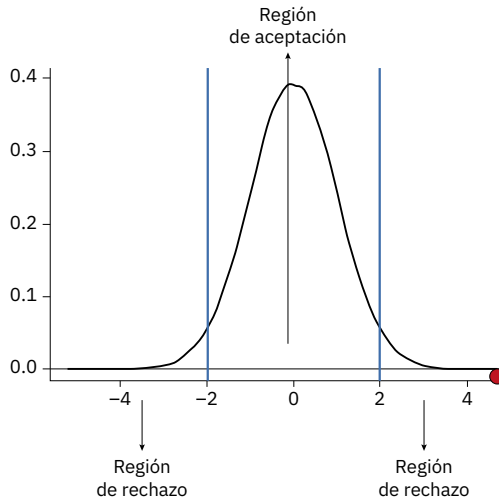
El resultat del contrast és a la taula 8.

Taula 8. Resultat del contrast d’hipòtesis d’una mitjana a Jamovi

| Prova T en una mostra            |              |            |        |          |                  |                               |                 |
|----------------------------------|--------------|------------|--------|----------|------------------|-------------------------------|-----------------|
|                                  |              |            |        |          |                  | Interval de confiança al 95 % |                 |
|                                  |              | Estadístic | gl     | p        | Mida de l'efecte | Inferior                      | Superior        |
| bmi3                             | T de Student | 13.889     | 3245.0 | < .00001 | d de Cohen       | 0.24378                       | 0.20885 0.27867 |
| Nota. $H_{\alpha} \mu \neq 24.9$ |              |            |        |          |                  |                               |                 |

6. Resolució del contrast.

La resolució es podria fer gràficament (figura 20). En el nostre cas, com el valor mostral de l'estadístic de contrast (igual a 13,9, vegeu *statistic* en la figura 20) se situa en algunes de les regions de rebuig (com en el nostre cas), es rebutjarà la hipòtesi nul·la.



**Figura 20. Resolució del contrast. Rebuig de la hipòtesi nul·la**

Però és molt més còmode utilitzar el **p-valor** (o valor p). En el nostre cas, observant la figura 20, el p-valor del contrast és  $p < 0,001$ , menor que el risc (0,05). És a dir, al 95 % de confiança no podem dir que l'IMC en el moment 3 sigui 24,9.

## Referències

- The Jamovi project (2022). *Jamovi*. (Versió 2.3) [Computer Software]. Recuperat de:  
<https://www.JAMOVI.org>
- R Core Team (2021). *R: A Language and environment for statistical computing*. (Versió 4.1) [Computer Software]. Recuperat 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]. Recuperat de:  
<https://github.com/yanlinlin82/ggvenn>
- Serdar Balci (2022). *ClinicoPath Jamovi Module* doi:10.5281/zenodo.3997188. [R package]. Recuperat 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]. Recuperat de:  
<https://CRAN.R-project.org/package=easyalluvial>
- Nick Barrowman (2020). *vtree: Display Information About Nested Subsets of a Data Frame.* [R package]. Recuperat de:  
<https://CRAN.R-project.org/package=vtree>
- JASP Team (2018). *JASP*. [Computer Software]. Recuperat de:  
<https://jasp-stats.org>
- Morey, R. D. i Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs.* [R package]. Recuperat de:  
<https://cran.r-project.org/package=BayesFactor>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., i 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>

# Capítol 7

## Introducció a la inferència de relació entre dues variables

Quan es consideren hipòtesis explicatives, cal treballar amb més d'una variable. Una d'elles, a la qual anomenarem dependent, presenta una distribució concreta de valors que no és aleatòria, sinó que és determinada per la distribució d'una altra (o altres) variables. Segons un model d'anàlisi determinada, fonamentat en un marc de referència teòric, aquestes altres variables «explicatives», a les quals anomenarem independents, tenen una relació amb la variable dependent que pot ser de simple covariació o d'associació i influència causal.

Aquesta relació explicativa també pot ser contrastada empíricament amb dades quantitatives, fins al punt d'afirmar o desmentir hipòtesis. Aquesta estratègia correspon a hipòtesis explicatives respecte a la població.

A la taula següent, podem veure les diferents anàlisis que podem utilitzar segons la relació entre les variables.

**Taula 1. Combinació de variables i anàlisi corresponent**

| Combinació de variables           | Possible anàlisi                                   |
|-----------------------------------|----------------------------------------------------|
| V. Qualitativa → V. Qualitativa   | Taules de contingència. Contrast de la khi quadrat |
| V. Qualitativa → V. Quantitativa  | Comparació de mitjanes. Anàlisi de la variància    |
| V. Quantitativa → V. Quantitativa | Anàlisi de regressió. Regressió lineal             |
| V. Quantitativa → V. Qualitativa  | Regressió logística                                |

De la mateixa manera que hem treballat en el capítol anterior, emprarem les variables quantitatives contínues *sysbp*, *diabp* i *imc*, i les variables qualitatives *age\_basal\_cat*, *angina* i *death*.

Previ a les seves anàlisis, descriurem les variables de forma conjunta, amb l'ús del paquet JJStat de Jamovi.

Primer ho farem, com a exemple, per a dues contínues, tant l'*scatter plot* com la matriu de correlacions. Podem veure el resultat en els gràfics següents.

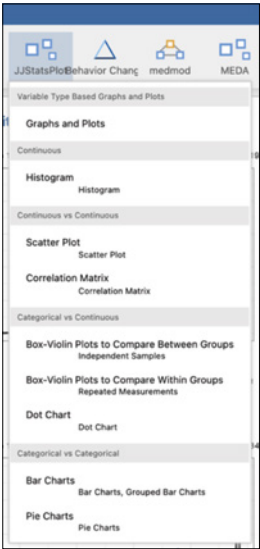


Figura 1. Possibilitats de JJStat

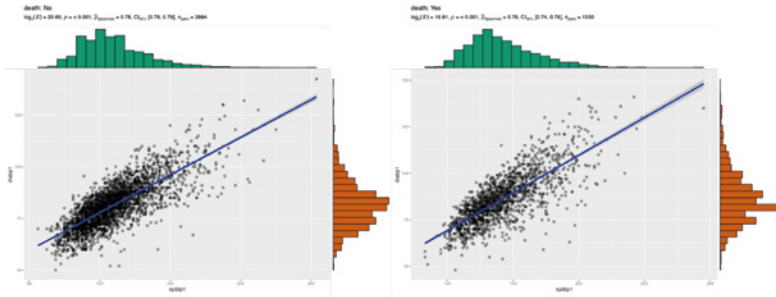


Figura 2. Scatter plot entre sysbp1 i diabp1, separades per death

Aquestes figures 2 i 3, ens mostren, prèviament a dur a terme anàlisis d'inferència, com estan relacionades les variables.

Les següents figures, 4 i 5, mostren la relació entre una variable contínua i una altra variable discreta (*sysbp1* i *age\_basal\_cat*). Podem obtenir molta informació a través d'elles, però ens centrarem en la possibilitat de fer, a partir d'elles, inferència per a dues o més variables.

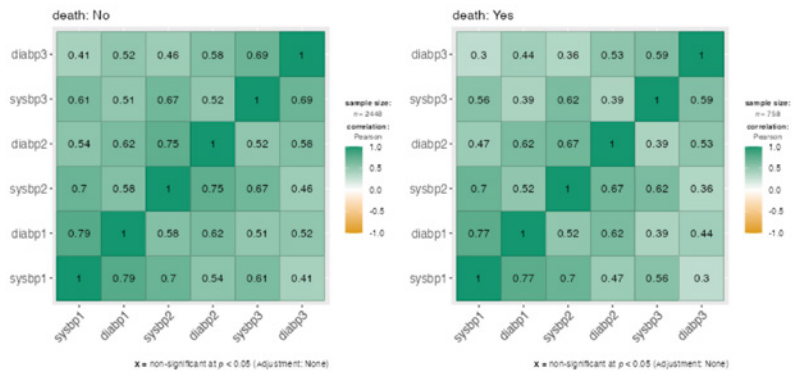


Figura 3. Matriu de correlacions entre sysbp i diabp, separades per death, per als tres estats

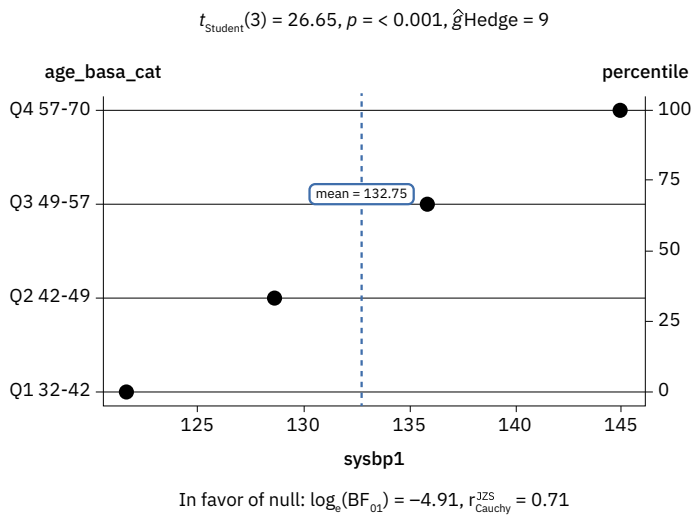
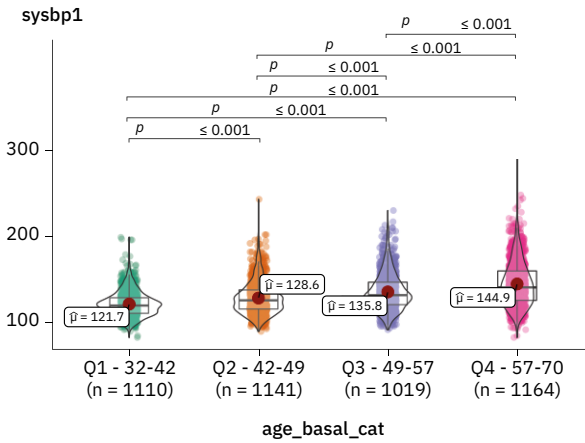


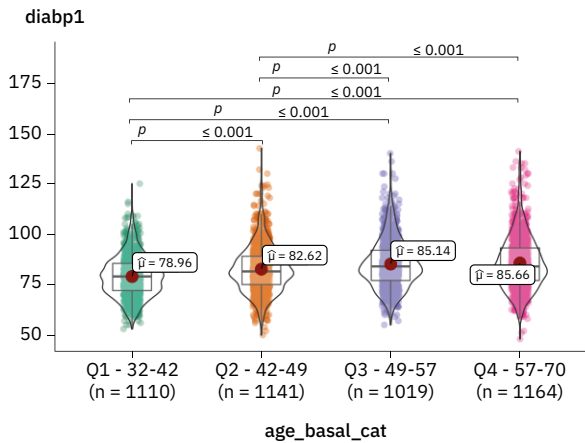
Figura 4. sysbp1 i age\_basl\_cat

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



In favor of null:  $\log(\text{BF}_{01}) = -353.91$ ,  $r_{\text{Cauchy}}^{225} = 0.71$   
 Pairwise comparisons: **Games-Howell test**; Adjustment (p-value): **Holm**

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



In favor of null:  $\log(\text{BF}_{01}) = -353.91$ ,  $r_{\text{Cauchy}}^{225} = 0.71$   
 Pairwise comparisons: **Games-Howell test**; Adjustment (p-value): **Holm**

**Figura 5. sysbp1 i age\_basl\_cat. (gràfic de violí)**

Finalment, com a exemple, si tenim dues variables contínues, podem veure la seva relació prèvia a la figura 6.

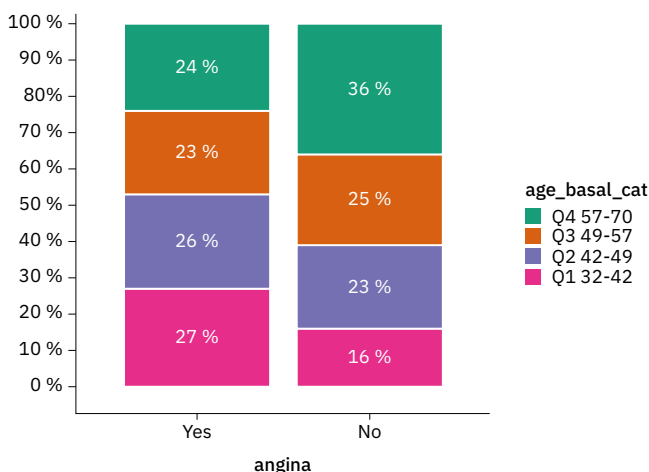


Figura 6. *angina* i *age\_basl\_cat*

## Comparació de proporcions. Taules de contingència

Després de conèixer com es poden relacionar les variables, començarem amb les taules de contingència, on relacionarem dues variables en format factor. Per a això, a Jamovi, podem observar la figura 7 de la finestra de «Freqüències», centrant-nos en «Mostres independents» per obtenir la taula de contingència.

Seleccionarem dues variables, en aquest cas *angina* i *death*, que són dos a dos casos, la qual cosa ens permetrà obtenir les odds ratio i risc relatiu, desenvolupat en el següent capítol (figura 8).

Primerament (figura 8), podem veure la descriptiva en format de taula de les dades de les dues variables, en files i columnes, però incloent, en aquest cas, els percentatges per files. També els valors de les mesures comparatives com són les odds ratio i els riscos relatius amb els seus respectius valors i intervals.



Figura 7. Procediment «Taules de contingència»

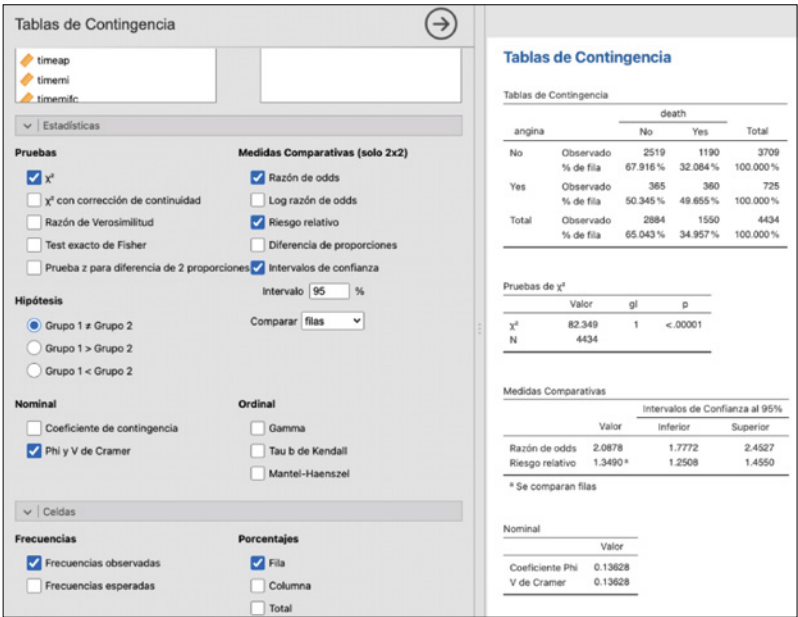


Figura 8. Resultats de la taula de contingència

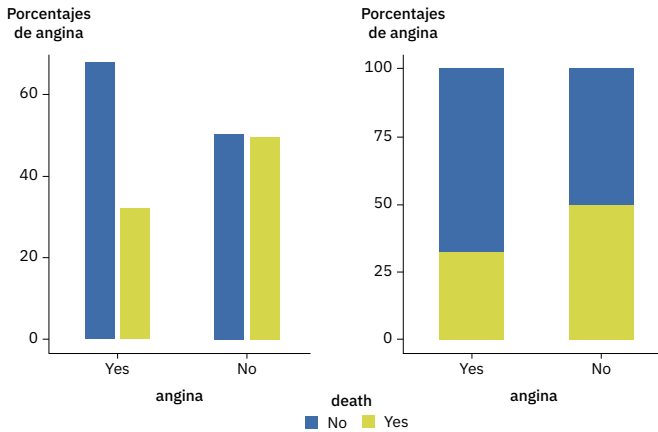


Figura 9. Gràfics exemple de la taula de contingència

En la figura 9, podem veure la representació gràfica que ens permet aquesta finestra, separant per percentatges per files. A la part esquerra, un al costat de l'altre i, a la dreta, alineats, que són les opcions que ens permet Jamovi.

L'objectiu de l'estadística inferencial, expressada aquí a partir del contrast de la khi quadrat ( $\chi^2$ ), és avaluar fins a quin punt aquestes diferències són extrapolables al conjunt de la població.

La conclusió pot ser:

- Les diferències **són significatives**. Arribarem a aquesta conclusió quan les diferències difícilment siguin degudes a l'atzar. Si unes diferències observades en la mostra no són degudes a l'atzar, és que es donen en la població general i, per tant, és molt probable que hi hagi relació entre les dues variables considerades.
- Les diferències **no són significatives**. Les diferències es poden deure a l'atzar, atès que les dades han estat obtingudes d'una mostra, de manera que les conclusions derivades de l'anàlisi no es poden extrapol·lar al conjunt de la població.

El valor que ens informa d'aquesta influència de l'atzar és la significació del contrast de la khi quadrat, també anomenat p-valor. Aquesta significació

ens informa sobre la probabilitat que les diferències observades siguin degudes a l'atzar:

- Quan la significació de  $\chi^2$  és gran no podem descartar que les diferències siguin degudes a l'atzar i, per tant, no extrapolables a la població.
- Mentre que si la significació de  $\chi^2$  és petita és que les diferències difícilment seran degudes a l'atzar i, en conseqüència, són extrapolables a la població.

Si el p-valor és igual o menor ( $p\text{-valor} \leq 0,05$ ) direm que podem acceptar que existeixen diferències significatives entre les categories (i, per tant, que hi ha relació entre les variables) amb un 95% de confiança. De fet, com menor sigui la significació més improbable és que l'atzar ens hagi jugat una mala passada i més probable és que les diferències es donin veritablement en el conjunt de la població. Això també ens servirà per a tots els apartats següents.

En aquest cas, podem observar un p-valor molt petit, menor que 0,001, per la qual cosa les diferències observades entre les dues variables es poden assegurar amb un risc d'error molt petit. És a dir, que la variable *angina* estigui relacionada amb *death* per a la població en general. El resultat d'aquesta etapa és que hi ha significació i, per tant, es pot continuar amb l'anàlisi. En cas que no hi hagi significació estadística es poden treure conclusions de la taula, però només seran aplicables a les persones entrevistades a la mostra (n). En aquest cas, no es podrà inferir cap conclusió a la població (N) a partir de les dades disponibles.

Un cop establert que hi ha relació significativa entre ambdues variables se sol calcular una mesura per avaluar la intensitat d'aquesta relació. És habitual fer-ho a partir de l'indicador d'associació bilateral V de Cramer, que varia entre 0 i 1: un valor de 0 equivaldria a dir que no hi ha cap relació (o, amb valors propers a 0, que és molt baixa) mentre que un valor proper a 1 hauria de ser llegit en termes d'associació forta entre les variables. En ciències de la salut no és habitual trobar valors molt elevats d'aquest indicador, per la qual cosa superar el valor de 0,3 ja es considera una intensitat elevada. En els resultats del cas d'*angina* i *death*, podem veure que hi ha un valor proper a 0 (0,13), intensitat baixa. Normalment, mentre aquest valor no sigui inferior a 0,1, es pot continuar amb l'anàlisi.

En casos extrems, quan treballem amb mostres molt petites, podem trobar-nos que fins i tot agrupant al màxim continuïn sense complir-se les condicions d'aplicació. Així, en taules 2 files x 2 columnes, quan alguna cel·la esperada és inferior a 5, però totes són superiors a 3, s'aplica una correcció en el contrast de la khi quadrat de Pearson anomenada correcció de Yates, i que Jamovi identifica com l'opció de continuïtat de  $\chi^2$ . Si l'incloem en aquest cas gairebé no canvia, a causa del nombre de dades que hi ha.

## Comparació de mitjanes

En aquesta secció, analitzarem la relació entre una variable independent qualitativa i una variable dependent quantitativa (de la qual es pot calcular la mitjana) mitjançant els procediments «T-tests» i «One-way ANOVA», així com per realitzar la representació gràfica pertinent i interpretar la informació que s'obté de Jamovi per donar resposta a les hipòtesis de relació entre aquestes dues variables d'acord amb un model d'anàlisi bàsica (aspecte descriptiu i de contingut de l'anàlisi), i determinar la significació d'aquesta informació per extrapolar-la al conjunt de la població (aspecte inferencial de l'anàlisi).

L'objectiu principal és mostrar com comparar les mitjanes d'una variable quantitativa entre els diferents grups o categories d'una variable qualitativa. Ens trobem davant la possible relació entre dues variables, una mètrica qualitativa, que farà el paper de variable independent (a Jamovi, de tipus NOMINAL o ORDINAL), i una altra quantitativa, la variable dependent a Jamovi, de tipus CONTÍNUA, amb dades SENCERA o DECIMAL).

### Comparació de dues mitjanes. T de Student

Començarem amb el cas de mostra independent, per exemple, dues variables preses en diferents moments (*dyabp* o *imc*) i per a diferents condicions, per exemple, *angina* o *death*.

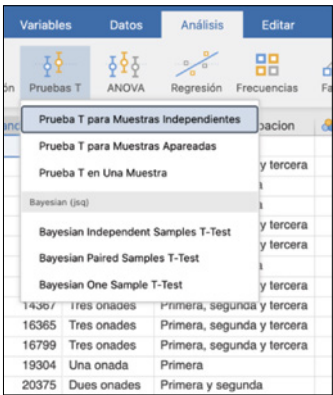


Figura 10. Tipus de proves T, centrada ara en mostres independents

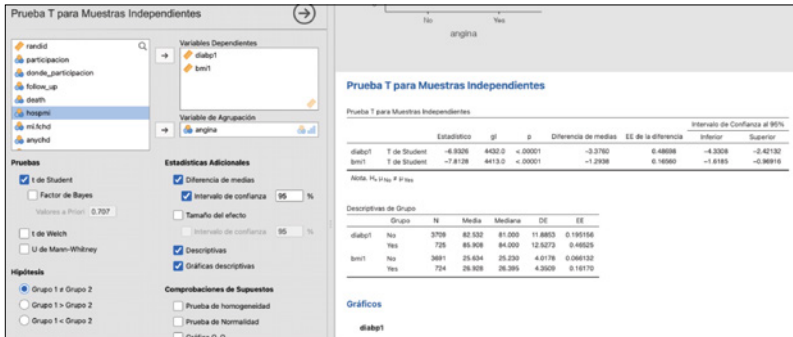


Figura 11. Resultats de la prova T amb dades independents

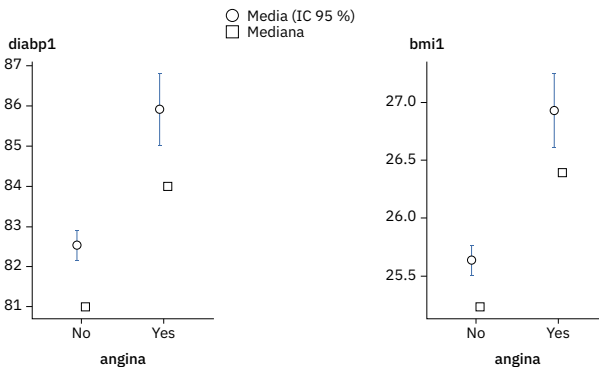


Figura 12. Resultats gràfics de la prova T amb dades independents

La interpretació d'aquesta visualització gràfica és:

1. En la mostra obtinguda, hi ha evidència de diferències entre la mitjana dels dos casos, ja que les línies verticals no coincideixen.
2. A més, aquestes diferències són prou elevades per assegurar que hi ha diferències entre *diabp* i *bmi* per a la separació entre *angina* No i Sí en la població en general perquè els nivells de confiança no coincideixen.

A més, podem completar amb el contrast de Shapiro-Wilk per a la normalitat de la variable. Tal com s'ha esmentat anteriorment, per poder efectuar un contrast de tipus paramètric (que és el més recomanat) cal que les mostres siguin «prou grans», o bé que la distribució «no difereixi gaire» d'una distribució normal. Quan no es compleix la condició, és a dir quan en algun grup poc nombrós no hi ha normalitat, és recomanable utilitzar

un contrast no paramètric. La majoria dels autors considera que si un dels grups comparats supera els 30 casos és suficient. Així, només avaluem la normalitat en els grups que no arribin a aquest nombre de casos.

Per avaluar la normalitat drem a terme la prova de normalitat de Shapiro-Wilk que ofereix Jamovi. Aquest contrast avalua les següents hipòtesis:

- $H_0$ : la distribució és normal.
- $H_1$ : la distribució no és normal.

S'interpreta de manera que si el p-valor és menor (o igual) a 0,05 considerem evidència de suport a  $H_1$  (no hi ha normalitat), mentre que si és superior a 0,05 podem assumir la normalitat (amb un 95 % de nivell de confiança). A continuació, podem veure els resultats obtinguts per a les nostres dades en el tema de normalitat (taula 2).

**Taula 2. Contrast de normalitat de diabp1 i 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 |

### Comparació de més de dues mitjanes. ANOVA d'un factor

La pregunta que ens formularem en aquest cas és similar a l'anterior. L'únic que canvia és que la variable explicativa qualitativa té més de dues categories o valors i, per tant, genera més de dos grups que comparar. En aquest cas ens preguntem si la variable contínua, per exemple, *bmi* o *diabp*, s'agrupa segons més d'un cas, *ag-basal\_cat*. Elegirem segons apareix a la imatge.

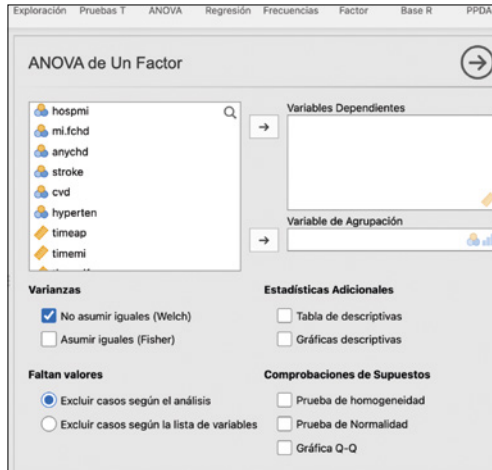


Figura 13. Procediment ANOVA d'un factor

Els resultats els podem veure a la figura 14.

El resultat ens el torna a donar en forma de p-valor. En aquest cas pren un valor menor a 0,001, de manera que hi ha prou evidència per afirmar que les variàncies no són iguals: cal realitzar el contrast ANOVA en cas de variàncies diferents.

En el panell de definició del procediment anterior es pot comprovar que existeixen dos contrastos possibles:

- Contrast paramètric en cas de variàncies diferents: *Don't assume equal (Welch's)*.
- Contrast paramètric en cas de variàncies iguals: *Assume equal (Fisher's)*.

També podem veure els resultats descriptius de l'anàlisi.

Els resultats gràfics d'ANOVA, els podem veure a la figura 15, on es mostra de forma clara la diferència per a la majoria dels casos.

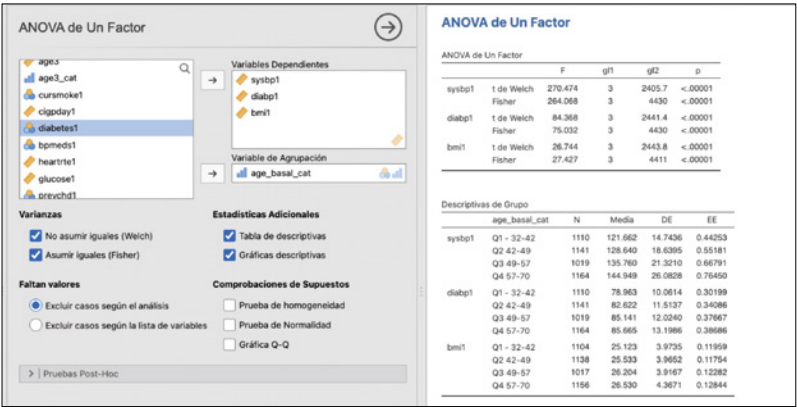


Figura 14. Resultats ANOVA d'un factor

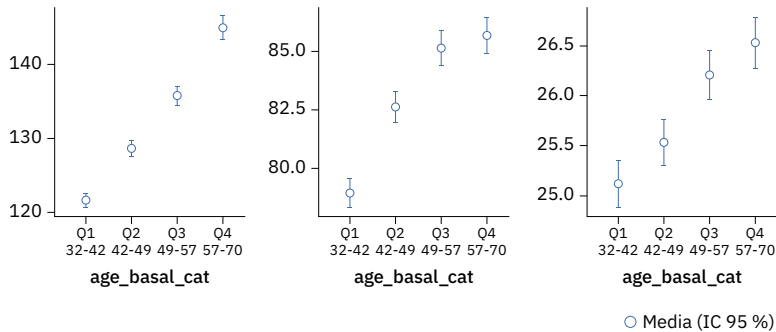


Figura 15. Resultats gràfics ANOVA d'un factor

Com més grups tinguem més comparacions 2 a 2 haurem de fer. Per avaluar-les cal:

1. En el mateix panell de definició del procediment «Anàlisi de variàncies simple», fes clic a «Post-Hoc Test» per activar les comparacions múltiples a posteriori, o comparacions 2 a 2.
2. Per defecte no es calcula cap comparació. Cal activar el tipus adequat, segons les conclusions del contrast de Levene previ: «Games-Howell» en aquest cas perquè existeix.

**Taula 3. Games-Howell de variàncies diferents****Prova Post-Hoc de Games-Howell – sysbp1**

|            |                        | <b>Q1 – 32-42</b> | <b>Q2 42-49</b> | <b>Q3 49-57</b> | <b>Q4 57-70</b> |
|------------|------------------------|-------------------|-----------------|-----------------|-----------------|
| Q1 – 32-42 | Diferència de mitjanes | —                 | –6.9776         | –14.0974        | –23.2872        |
|            | valor p                | —                 | < .00001        | < .00001        | < .00001        |
| Q2 42-49   | Diferència de mitjanes |                   | —               | –7.1198         | –16.3095        |
|            | valor p                |                   | —               | < .00001        | < .00001        |
| Q3 49-57   | Diferència de mitjanes |                   |                 | —               | –9.1897         |
|            | valor p                |                   |                 | —               | < .00001        |
| Q4 57-70   | Diferència de mitjanes |                   |                 |                 | —               |
|            | valor p                |                   |                 |                 | —               |

**Prova Post-Hoc de Games-Howell – diabp1**

|            |                        | <b>Q1 – 32-42</b> | <b>Q2 42-49</b> | <b>Q3 49-57</b> | <b>Q4 57-70</b> |
|------------|------------------------|-------------------|-----------------|-----------------|-----------------|
| Q1 – 32-42 | Diferència de mitjanes | —                 | –3.6588         | –6.1778         | –6.70146        |
|            | valor p                | —                 | < .00001        | < .00001        | < .00001        |
| Q2 42-49   | Diferència de mitjanes |                   | —               | –2.5190         | –3.04270        |
|            | valor p                |                   | —               | < .00001        | < .00001        |
| Q3 49-57   | Diferència de mitjanes |                   |                 | —               | –0.52369        |
|            | valor p                |                   |                 | —               | 0.76662         |
| Q4 57-70   | Diferència de mitjanes |                   |                 |                 | —               |
|            | valor p                |                   |                 |                 | —               |

**Prova Post-Hoc de Games-Howell – bmi1**

|            |                        | <b>Q1 – 32-42</b> | <b>Q2 42-49</b> | <b>Q3 49-57</b> | <b>Q4 57-70</b> |
|------------|------------------------|-------------------|-----------------|-----------------|-----------------|
| Q1 – 32-42 | Diferència de mitjanes | —                 | –0.41044        | –1.08157        | –1.40706        |
|            | valor p                | —                 | 0.06870         | < .00001        | < .00001        |
| Q2 42-49   | Diferència de mitjanes |                   | —               | –0.67112        | –0.99662        |
|            | valor p                |                   | —               | 0.00047         | < .00001        |
| Q3 49-57   | Diferència de mitjanes |                   |                 | —               | –0.32549        |
|            | valor p                |                   |                 | —               | 0.25871         |
| Q4 57-70   | Diferència de mitjanes |                   |                 |                 | —               |
|            | valor p                |                   |                 |                 | —               |

Es mostren els resultats en format de taula, on se situen les diferents categories de la variable en files i columnes, de manera que el punt d'encreuament ens dona la significació de la diferència entre els grups corresponents. Així, el valor  $p < 0,001$  és la significació en la majoria dels casos ( $p\text{-valor} < 0,05$ ).

## Altres mòduls per a relacions entre variables

Com ja hem vist amb anterioritat en el capítol d'inferència per a una variable, podem utilitzar altres mòduls no genèrics per estudiar la relació entre variables. En la figura 16, podem veure l'ús del mòdul Jeva i Toster. Tots dos ens permeten realitzar alguns casos com els anteriorment desenvolupats.

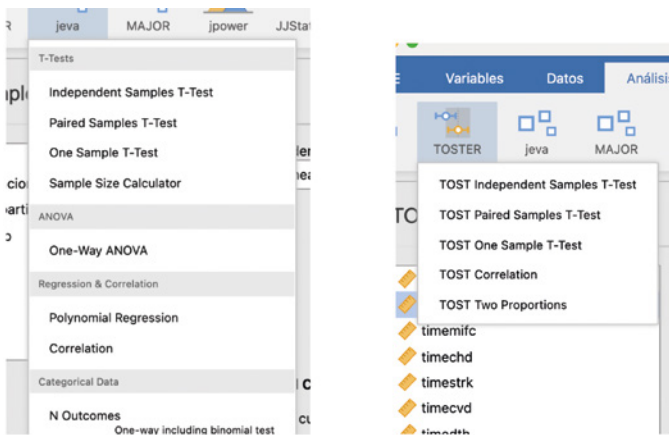


Figura 16. Jamovi de Jeva i Toster

En la següent imatge, podem veure la comparació de mitjanes de dues mostres independents amb Jeva. Els resultats són molt més complexos i complets, tot i que els resultats són els mateixos.

Altres dues possibilitats són esci o anàlisi robusta amb Walrus (figura 18).

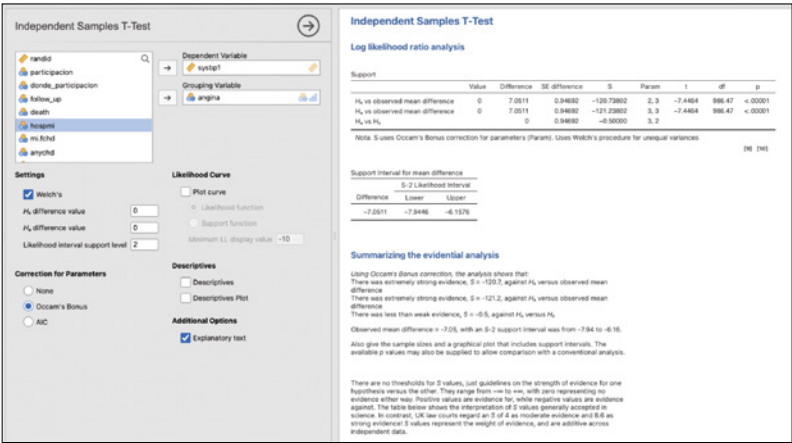


Figura 17. Resultats amb JEVA

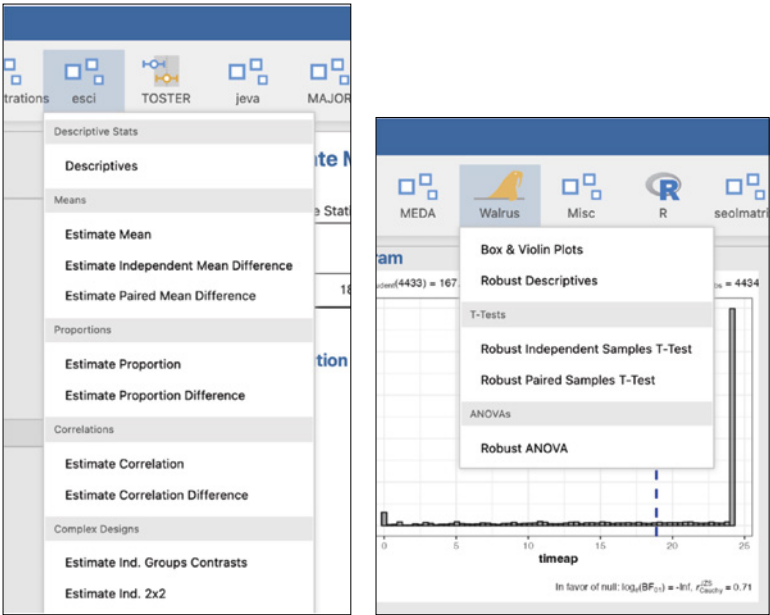
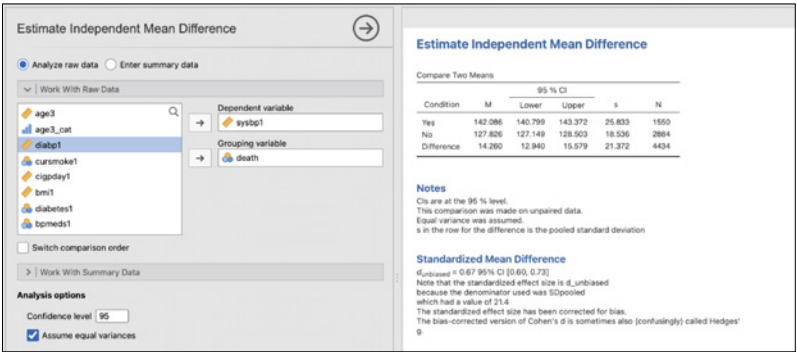
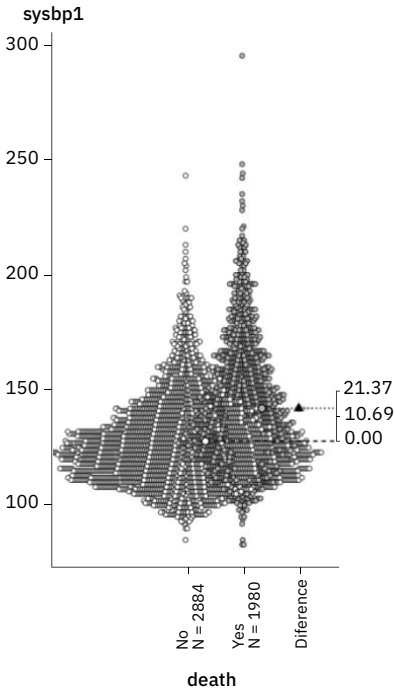


Figura 18. Jamovi d'esci i Walrus



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. Resultats d'esci

## Referències

- The Jamovi project (2022). *Jamovi*. (Versió 2.3) [Computer Software]. Recuperat de: <https://www.JAMOVl.org>
- R Core Team (2021). *R: A Language and environment for statistical computing*. (Versió 4.1) [Computer Sowntware]. Recuperat de (R packages recuperat 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]. Recuperat de: <https://github.com/yanlinlin82/ggvenn>.
- Serdar Balci (2022). *ClinicoPath Jamovi Module* doi:10.5281/zenodo.3997188. [R package]. Recuperat 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]. Recuperat de: <https://CRAN.R-project.org/package=easyalluvial>
- Nick Barrowman (2020). *vtree: Display Information About Nested Subsets of a Data Frame*. [R package]. Recuperat de: <https://CRAN.R-project.org/package=vtree>
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., i RStudio (2018). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. [R package]. Recuperat de: <https://CRAN.R-project.org/package=ggplot2>
- Patil, I. (2018). *ggstatsplot: 'ggplot2' Based Plots with Statistical Details*. [R package]. Recuperat 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>

## Capítol 8

# Mesures utilitzades en epidemiologia

L'**epidemiologia** és una disciplina científica en l'àrea de la salut pública (no només la medicina), que estudia la distribució, freqüència, magnitud i factors determinants de les malalties existents en poblacions humanes definides. Rich la va descriure el 1979 com la ciència que estudia la dinàmica de salut a les poblacions; per tant, involucra l'anàlisi i interpretació de les persones que també estan sanes.

## Quocients

En epidemiologia s'utilitzen diversos tipus de mesures, d'entre les quals assenyalarem els **quocients**, entre els quals es troben les proporcions, les raons, les taxes i els odds.

Les **proporcions** són quocients en els quals el numerador està inclòs en el denominador, no tenen unitats i el seu rang varia entre 0 i 1 (0 % i 100 % si s'expressa en tant per cent).

Per exemple, en una població amb 500 homes i 700 dones, la proporció d'homes és  $500 / (500 + 700) = 0,4167$  (41,67 %) i la de dones és  $700 / (500 + 700) = 0,5833$  (58,33 %). Es podria interpretar que en aquesta població un 41,67 % són homes i un 58,33 % dones.

Les **raons** (*ratio* en anglès) són quocients en què el numerador no està inclòs en el denominador, no tenen unitats, però el seu rang no està acotat, variant entre 0 i infinit.

En l'exemple anterior, la raó de masculinitat és  $500 / 700 = 0,7143$  i la raó de feminitat  $700 / 500 = 1,4$ . Es podria interpretar que per cada home hi ha 1,4 dones.

Les **taxes** (*rates* en anglès) són quocients que indiquen el nombre d'individus d'una població (o mostra) amb un atribut. El denominador indica la població total, però inclou el temps. S'interpreta com la velocitat d'ocurrència de l'atribut en la població (o mostra). Les taxes més utilitzades en epidemiologia són la taxa de natalitat (naixements/població en un període determinat), la taxa de fecunditat (naixements/dones en edat fèrtil en un període determinat), la taxa bruta de mortalitat (defuncions/població en un període determinat). Per convenció, en el

denominador es fa servir el nombre d'individus estimat en la meitat del període (usualment un any). Normalment es donen en tant per mil.

Per exemple, suposem una població de 8500 habitants en els quals en un any es produeixen 80 infarts de miocardi: la taxa d'infart és de  $80 / 8500 = 0,0094$ , és a dir, 9,4 infarts per cada 1000 habitants per any.

Els **odds** (de vegades es tradueixen com a 'avantatges') són un cas particular de raó i són el quocient entre individus que tenen un atribut i d'altres que no en tenen. Tampoc tenen unitat.

Per exemple, de 90 pacients donats d'alta a l'hospital A, tenim 60 reingressos. L'odds (de reingressar) es calcularia com a  $\text{odds} = 60 / (90 - 60) = 2$ . És a dir, els reingressos són 2 vegades més freqüents que els no reingressos.

## Mesures de freqüència

Les **mesures de freqüència** quantifiquen quantificant o comú resulta determinat fenomen d'interès (malaltia, lesió, mort).

Entre les mesures de freqüència assenyalarem la prevalença i la incidència.

Les mesures de **prevalença** informen de la proporció de persones d'una població amb una característica determinada (per exemple, una malaltia o un factor de risc, etc.) en un moment o període de temps determinat. Es pot interpretar com la probabilitat individual d'una persona de la població de tenir la característica. És una mesura pròpia dels estudis transversals. Hi ha dues mesures de prevalença, la **prevalença puntual** (la més utilitzada) i la **prevalença de període**.

Per exemple, en una població de 10000 persones, s'informa que 500 persones pateixen una determinada malaltia. La prevalença (o prevalença puntual) és de  $500 / 10000 = 0,05$ , és a dir, del 5%. També es pot interpretar que la probabilitat que una persona d'aquesta població tingui la malaltia és del 5%. Si en aquesta mateixa població, 2430 persones desenvolupen la malaltia en un període de 5 anys, la prevalença de període és  $2430 / 10000 = 0,243$ , 24,3%. És a dir, que la probabilitat que una persona d'aquesta població tingui la malaltia en un període de 5 anys és del 23,4%.

Les mesures d'**incidència** es fixen en el nombre de casos nous d'ocurrència d'un esdeveniment (per exemple, malaltia) que apareixen entre les persones susceptibles (per exemple, d'emmalaltir) d'una població al llarg d'un període de temps. Cal un seguiment. Es calculen sobre cohorts.

Igualment, hi ha dues mesures d'incidència, la **incidència acumulada** i la **taxa o densitat d'incidència**.

En la **incidència acumulada**, les persones controlades són fixes i han d'estar lliures de l'esdeveniment a l'inici del seguiment. Es pot interpretar com la probabilitat individual que una persona de la població desenvolupi la característica.

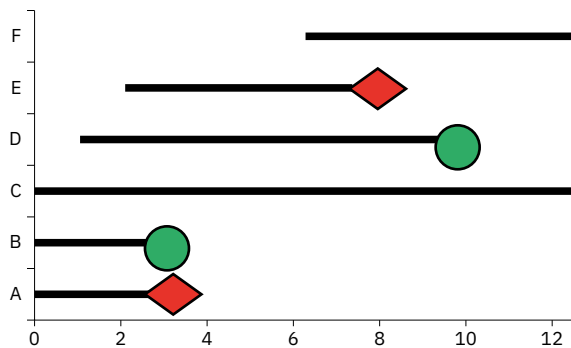
Per exemple, en una població de 1230 persones sense COVID-19, en una setmana se'n van infectar 28. La **incidència acumulada** es calcularà com:

$$(\text{nombre de casos nous durant el període}) / (\text{nombre de casos susceptibles al principi del període}) = 28 / 1230 = 0,0227$$

També se sol expressar per mil. En aquest cas la incidència acumulada en 7 dies és de 22,7 per cada 1000 habitants.

La **densitat d'incidència** és el quocient entre els nous casos durant el període i la suma dels temps de seguiment de cada individu. Més adequat per a poblacions dinàmiques. Informa sobre la velocitat de l'esdeveniment. A diferència de la incidència acumulada no té interpretació individual.

Per exemple, se segueix durant dotze mesos un grup de 6 individus que han patit un accident cerebrovascular (ACV), per avaluar la incidència de recidiva. La situació es pot esquematitzar en la imatge:



**Figura 1.** Seguiment de 6 individus amb ACV durant dotze mesos

És una cohort dinàmica en la qual es permeten incorporacions. Per tant, el seguiment no va començar alhora per a tots els individus. Els rombes vermells assenyalen recidives (pacients A i E), mentre que els cercles verds assenyalen pèrdues. Així, el pacient B va desaparèixer de l'estudi al cap de 2,5 mesos, sense que fins aquell moment hagués patit una recidiva; el pacient D va morir per una altra causa no relacionada amb l'ACV; els pacients C i F van acabar el període d'estudi sense recidiva.

La densitat d'incidència es calcularia com:

(nombre de casos nous durant el període) / / (suma del temps de seguiment dels casos susceptibles, és a dir sense la malaltia)

El nombre de casos nous durant el període de seguiment és igual a 2.

El temps de seguiment dels casos susceptibles:

- Individu A: 3 mesos (fins que va tenir la recidiva)
- Individu B: 2,5 mesos
- Individu C: 12 mesos
- Individu D: 8 mesos
- Individu E: 6 mesos (fins que va tenir la recidiva).
- Individu F: 6 mesos
- Total temps de seguiment:  $3 + 2,5 + 12 + 8 + 6 + 6 = 37,5$  mesos.

Per tant,

densitat d'incidència =  $2 / 37,5 = 0,053$  per cada mes

## Mesures d'associació

Les **mesures d'associació** mesuren l'associació entre dues variables dicotòmiques, el factor de risc i la malaltia o mort (en tots dos casos, presència o absència).

En general, les mesures d'associació quantifiquen el risc. El terme *risc* implica que la presència d'una característica o factor augmenta la probabilitat de conseqüències adverses. El risc es pot mesurar mitjançant:

- **Risc absolut.** Mesura la incidència del dany en la població total.
- **Risc relatiu.** Compara la freqüència amb què passa el dany entre els que tenen el factor de risc i els que no el tenen.

Per exemple, mostrem la taula d'un estudi de seguiment on 853 dones van estar passivament exposades al fum del tabac durant la gestació i 1620 no ho van estar, i la seva associació amb el baix pes de la criatura en néixer.

**Taula 1. Distribució de gestants segons exposició passiva al fum de tabac i nounats segons baix pes o pes normal**

| Exposició passiva al tabac | Nounat baix pes |      | Total |
|----------------------------|-----------------|------|-------|
|                            | Sí              | No   |       |
| Sí                         | 20              | 833  | 853   |
| No                         | 14              | 1606 | 1620  |
| Total                      | 34              | 2439 | 2473  |

Risc (absolut) de baix pes en les gestants exposades al tabac:

$$20 / 853 = 0,0234$$

Risc (absolut) de baix pes en les gestants no exposades al tabac:

$$14 / 1620 = 0,0086$$

És a dir, el risc (o probabilitat, ja que es tracta d'una proporció) que una gestant exposada al tabac tingui un nounat amb baix pes és del 2,34 %, mentre que en una gestant no exposada al tabac és del 0,86 %.

El risc relatiu mesura la força de l'associació entre l'exposició i la malaltia. Indica la probabilitat que es desenvolupi la malaltia en els exposats a un factor de risc en relació amb el grup dels no exposats.

Continuant amb el nostre exemple, calcularem el **risc relatiu**.

Risc relatiu (RR) = (Risc de baix pes en les gestants exposades al tabac) /  
(Risc de baix pes en les gestants no exposades al tabac)

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

El risc que una gestant exposada al tabac tingui un nounat amb baix pes és 2,72 més gran que en una gestant no exposada al tabac. Com és un quocient de dues porcions també es pot expressar com a probabilitat. D'altra banda, es pot interpretar de la manera següent:

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

Una gestant exposada al tabac té un 172,09 % més risc (o probabilitat) que una gestant no exposada, que el nounat tingui baix pes.

D'altra banda, si els esdeveniments són infreqüents (el més habitual en ciències de la salut) el risc també es pot mesurar mitjançant l'**odds ràtio** (OR) (també denominat *raó d'avantatges*, entre molts altres termes). En

els estudis de **casos i controls**, atès que la incidència és desconeguda, el mètode d'estimació del risc relatiu és diferent i s'estima calculant l'OR.

En l'exemple que hem vist més amunt, de 90 pacients donats d'alta a l'hospital A, en tenim 60, mentre que a l'hospital B, de 50 pacients donats d'alta, en tenim 30.000.

- Odds (de reingressar) a l'hospital A =  $60 / (90 - 60) = 2$ .
- Odds (de reingressar) a l'hospital B =  $30 / (50 - 30) = 1,5$ .

Així, l'odds ràtio (raó d'odds) és igual a  $2 / 1,5 = 1,33$

El risc de reingressar a l'hospital A és 1,33 més gran que a l'hospital B. També

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

L'hospital A té un 33 % més gran de risc de reingrés que l'hospital B.

Com que els odds no són proporcions, l'OR no es pot interpretar com a probabilitat.

Tant en el RR com en l'OR, quan són majors que 1 el factor és un **factor de risc** de la malaltia, quan és menor que 1 és un **factor protector** i quan és igual a 1, no existeix relació entre el factor i la malaltia.

## Mesures d'impacte

Les **mesures d'impacte** descriuen el possible impacte potencial en la població que pogués tenir l'eliminació del factor de risc en qüestió sobre el desenvolupament de la malaltia; són les mesures d'impacte potencials, de les quals només és possible el seu càlcul en dissenys de prevenció de malaltia, com els estudis de cohorts i casos i controls. Les mesures d'impacte potencials es divideixen en absolutes, quan mesuren l'excés de risc en els exposats en la població que es pren de referència, i en relatives, quan mesuren el percentatge de risc que es deu a exposició.

### Mesures d'impacte absolut

El risc atribuïble en els exposats (RAE) o, també, **reducció absoluta de risc (RAR)**, es calcula com la diferència entre la incidència d'exposats i no exposats ( $RAE = RAR = I_{\text{exposats}} - I_{\text{no exposats}}$ ). La diferència entre els dos valors dona el valor del risc de malaltia en la cohort exposada, que es deu exclusivament a l'exposició.

En l'exemple de la taula 1:

- Incidència de baix pes en les gestants exposades al tabac:  $20 / 853 = 0,0234$
- Incidència de baix pes de les gestants no exposades al tabac:  $14 / 1620 = 0,0086$

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

Es podria interpretar que suprimint l'exposició al tabac en les gestants evitariem de mitjana 1,48 (arrodonint, 1,5) nounats de baix pes per cada 100 naixements en un període determinat.

Utilitzant el RAE es pot calcular el **nombre necessari de pacients a tractar** per reduir un efecte o **NNT** com a

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

Si el RAE és negatiu es fa servir el **nombre necessari per danyar, NNH**, que s'obté simplement canviant l'ordre del denominador,

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

El NNT i el NNH són unes mesures d'associació utilitzades en assajos clínics.

Com menor sigui l'NNT (major la diferència entre la incidència en els exposats) més efectivitat tindrà el tractament de l'assaig clínic. I com més gran sigui el NNH més segur.

Per exemple, a la taula 2, veiem que els règims intensius d'insulina en pacients diabètics per prevenir les retinopaties són molt efectius. El NNT és igual a 4. És a dir, necessitaríem tractar (amb un règim intensiu d'insulina) 4 pacients diabètics per evitar 1 cas de retinopatia. Tanmateix, per prevenir la neuropatia diabètica és poc efectiu. El NNT és igual a 15. Ens caldria tractar (amb un règim intensiu d'insulina) 15 pacients diabètics per evitar 1 cas de neuropatia diabètica. Molt menys efectiva és l'exploració de mames, a més de la mamografia, en dones sanes de 50 a 69 anys, per prevenir la mort per càncer de mama. Caldrien 1075 dones per evitar 1 cas.

## Mesures d'impacte relatiu

La **fracció atribuïble** estima la proporció de la malaltia entre els exposats que pot ser atribuïble al fet d'estar exposats. La fracció atribuïble es pot calcular en els exposats i en la població.

La **fracció atribuïble en el grup exposat** (fracció etiològica o percentatge de risc atribuïble en els exposats) (FAE), estableix el grau d'influència que té l'exposició en la presència de malaltia entre els exposats. El seu càlcul es realitza com:

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

Amb les dades de la taula 1,

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

que s'interpretaria com que el 63,25% del baix pes en els nounats de gestants exposades al tabac es deu a l'exposició.

La **fracció atribuïble en la població** (FAP), mostra la proporció en què el dany podria ser reduït si els factors de risc causals desapareguessin de la població total. Es calcularia com:

$$FAP = (I_{\text{població}} - I_{\text{no exposats}}) / I_{\text{població}}$$

Si disposem de la incidència de l'exposició en la població, el càlcul també es pot efectuar de la manera següent amb aquesta fórmula alternativa,

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

on Pt denota la rellevància de l'exposició (o factor de risc) en la població. Així doncs, la fracció atribuïble en la població total és una mesura d'associació influenciada per la prevalença del factor de risc en la població total.

Per calcular-la en el nostre exemple necessitarem la prevalença del baix pes en néixer que, a Espanya, s'estima al voltant del 7,8 %. Així doncs,

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

El risc de baix pes en la població atribuïble a l'exposició de les gestants al tabac és de l'11,83 %.

Taula 2. NNT per a diferents tractaments

| Malaltia                                                | Intervenció                             | Episodis que es prevenen                                    | Taxa en el grup control | Taxa en el grup experimental | Durada del seguiment | NNT per evitar un episodi addicional |
|---------------------------------------------------------|-----------------------------------------|-------------------------------------------------------------|-------------------------|------------------------------|----------------------|--------------------------------------|
| Diabetis (DMID) <sup>(1)</sup>                          | Règims intensius d'insulina             | Neuropatia diabètica                                        | 0,096                   | 0,028                        | 6,5 anys             | 15                                   |
| Diabetis (DMNID) <sup>(2)</sup>                         | Règims intensius d'insulina             | Retinopaties                                                | 0,38                    | 0,13                         | 6 anys               | 4                                    |
|                                                         |                                         | Nefropatia                                                  | 0,30                    | 0,10                         | 6 anys               | 5                                    |
| Infart de miocardi <sup>(3)</sup>                       | Estreptoquinasa i aspirina              | Mort a les 5 setmanes                                       | 0,134                   | 0,081                        | 5 setmanes           | 19                                   |
|                                                         |                                         | Mort als 2 anys                                             | 0,216                   | 0,174                        | 2 anys               | 24                                   |
| Pressió arterial diastòlica 115–129 mmHg <sup>(4)</sup> | Fàrmacs antihipertensius                | Mort, apoplexia o infart de miocardi                        | 0,0545                  | 0,0467                       | 5,5 anys             | 128                                  |
| Persones grans independents <sup>(5)</sup>              | Estudi geriàtric exhaustiu              | Permanència en residències durant un llarg període de temps | 0,10                    | 0,04                         | 3 anys               | 17                                   |
| Dones embarassades amb eclàmpsia <sup>(6)</sup>         | MgSO4 iv (vs. diazepam)                 | Convulsions recurrents                                      | 0,279                   | 0,132                        | Hores                | 7                                    |
| Dones sanes de 50–69 anys <sup>(7)</sup>                | Exploració de mames a més de mamografia | Mort per càncer de mama                                     | 0,00345                 | 0,00252                      | 9 anys               | 1075                                 |
| Estenosi greu simptomàtica de l'artèria <sup>(8)</sup>  | Endarterectomia                         | Apoplexia total o mort                                      | 0,181                   | 0,08                         | 2 anys               | 10                                   |
| Nens prematurs <sup>(9)</sup>                           | Corticoides prenatals                   | Síndrome de distrès respiratori                             | 0,23                    | 0,13                         | Dies                 | 11                                   |

Font: Pita Fernández i López de Ullibarri Galparsoro (2001).

## Aplicació amb la base de dades de Framingham

Per exemplificar les mesures d'associació, ens centrarem en les variables de la base de dades Framingham:

- **death.** Mort per qualsevol causa.
- **sexe.** Sexe del participant.
- **obese.** 1 Obesitat ( $\text{IMC} \geq 30 \text{ kg/m}^2$ ), 0 un altre cas.

Utilitzarem Jamovi per facilitar l'anàlisi que efectuarem. En primer lloc, carreguem les dades del fitxer Framingham (figura 2).

Mesurarem l'associació entre l'obesitat en el moment d'entrada a la cohort (*obese1*) i la mort (*death*). En la pestanya «Anàlisi» triem el menú «Freqüències» i en l'apartat «Taules de contingència» ubiquem la variable *death* per columnes i la variable *obese* per files. A «Cel·les» marquem percentatges per files. A més, a la barra «Estadístiques» desmarquem totes les opcions.

Com veiem, la incidència de mort entre els subjectes que van entrar obesos en la cohort és del 44,9 %, mentre que entre els subjectes que no eren obesos en entrar va ser del 33,3 %.

Estimarem el risc relatiu, assenyalant en «Estadístics», risc relatiu i intervals de confiança.

| Data        |        | Analyses     |                            |            |             |        |         |          |          |     |               |        |
|-------------|--------|--------------|----------------------------|------------|-------------|--------|---------|----------|----------|-----|---------------|--------|
| Exploration |        | T-Tests      | ANOVA                      | Regression | Frequencies | Factor | Base R  | Survival | Flexplot | ufs | Linear Models | TOSTER |
|             | randid | participa... | donde_participacion        | follow_up  | death       | angina | hospmli | 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      | 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      | No       |          |     |               |        |
| 17          | 33077  | Tres onades  | Primera, segunda y tercera | 12         | No          | No     | No      | No       |          |     |               |        |
| 18          | 33555  | Una onada    | Primera                    | Yes        | No          | 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     | 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     | 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      | No       |          |     |               |        |
| 35          | 70948  | Tres onades  | Primera, segunda y tercera | 12         | Yes         | Yes    | No      | No       |          |     |               |        |
| 36          | 76273  | Una onada    | Primera                    | Yes        | 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       |          |     |               |        |

Ready

Filters 0

Row count 4434

Filtered 0

Deleted 0

Added 0

Cells edited 0

Figura 2. Finestra de les dades Framingham

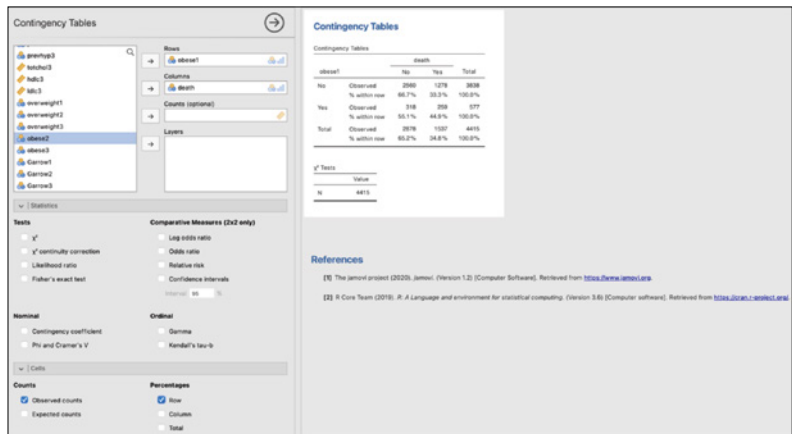


Figura 3. Finestres que apareixen a «Taules de contingència»

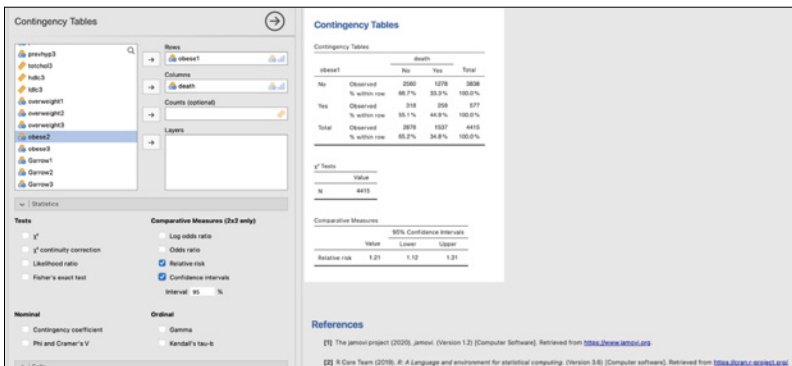


Figura 4. Finestres que apareixen a «Taules de contingència»

El risc relatiu és igual a 1,21. Vegeu que és estadísticament significatiu (al 95 %), ja que el seu interval de confiança no conté l'1. Tingueu en compte que la mateixa informació l'obtenim utilitzant el contrast khi quadrat d'independència (o, també, diferència de proporcions).

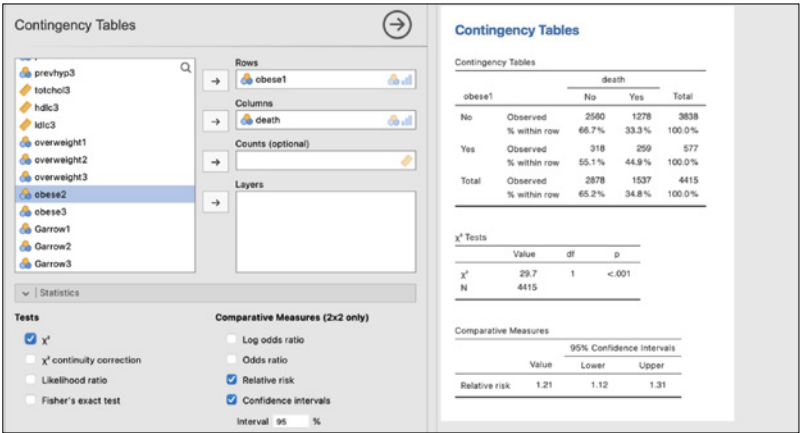


Figura 5. Resultats del comandament de «Taules de contingència»

Rebutgem la hipòtesi nul·la (d'independència o d'igualtat de proporcions), ja que el p-valor del contrast és menor que el risc ( $< 0,001$  és menor que 0,05).

Així doncs, ser obès és un factor de risc de morir. En concret, si el subjecte va entrar obès a la cohort va tenir un 21 % més de risc de morir que si no era obès quan va entrar.

Tot i que sent una cohort no seria apropiat mesurar el risc utilitzant l'odds ràtio, l'estimarem també.

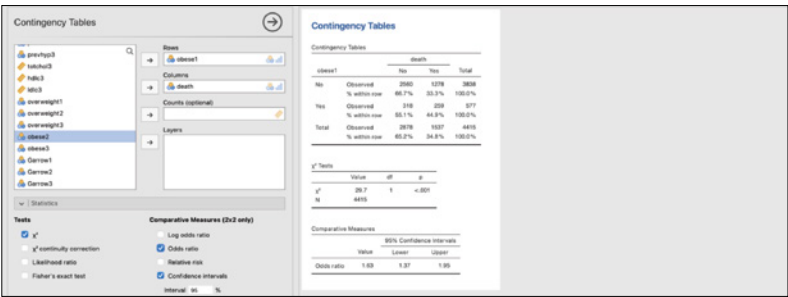


Figura 6. Resultats del comandament de «Taules de contingència»

L'odds ràtio, també estadísticament significatiu, és igual a 1,63.

Calcularem ara les mesures d'impacte.

El risc atribuïble en els exposats (RAE) es calcularia com a

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

Si cap dels subjectes hagués estat obès hauríem evitat, de mitjana, 11,6 (arrodonint, 12) morts per cada 100 subjectes vius.

La fracció atribuïble en el grup exposat (FAE) es calcularia com a

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

El 25,84 % de les morts es deu a l'obesitat.

## Referències

The Jamovi project (2022). *Jamovi*. (Versió 2.3) [Computer Software].

Recuperat de:

<https://www.JAMOVI.org>

Pita Fernández, S., Vila Alonso, M. T. i Carpena Montero, J. (1997).

Determinació dels factors de risc. *Cad Aten Primària*; 4:75-78.

Disponible a (últim accés el 20 de maig de 2024):

<https://www.fisterra.com/formacion/metodologia-investigacion/determinacion-factores-riesgo/>

Pita Fernández, S. i López de Uribarri Galparsoro, I. (1998, revisat el 2001).

Nombre necessari a tractar per reduir un esdeveniment. *Cad Aten Primària*; 5:96-98. Disponible a (últim accés el 20 de maig 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. (2014) (2014). *A dictionary of epidemiology* [6a edició]. Nova York:

Oxford University Press.

R Core Team (2021). *R: A Language and environment for statistical computing*. (Versió 4.1) [Computer programari]. Recuperat de (R

packages recuperat de MRAN snapshot 2022-01-01):

<https://cran.r-project.org>

## Capítol 9

# Introducció a la regressió

### Regressió lineal

En aquest capítol, veurem de forma introductòria la regressió lineal. La regressió estadística és una eina fonamental per comprendre les relacions entre variables i prendre decisions informades en diversos àmbits. La seva versatilitat i capacitat per modelar relacions complexes la converteixen en un mètode indispensable per a l'anàlisi de dades.

La **regressió lineal** és una tècnica estadística àmpliament utilitzada per modelar i analitzar la relació entre una **variable dependent** i una o més **variables independents**. La regressió lineal busca trobar la millor línia recta que s'ajusti a les dades observades. En la seva forma més simple, assumeix una relació **lineal** entre les variables. Va més enllà de la simple relació lineal; s'aplica en diversos camps, com economia, ciències socials, medicina i enginyeria.

#### Tipus de regressió:

- Regressió lineal simple: analitza la relació entre una variable dependent i una variable independent.
- Regressió lineal múltiple: estudia la relació entre una variable dependent i dues o més variables independents, considerant múltiples factors.

### Regressió lineal simple

Començarem amb la regressió lineal simple, on utilitzarem variables contínues per veure la seva relació. En general, farem servir les mateixes variables que en els capítols 6 i 7.

Inicialment, podem veure el diagrama de dispersió, que, en la imatge següent, ens mostra Jamovi (figura 1). També podem veure en la imatge següent, el diagrama de dispersió, però separant per una altra variable (figura 2).

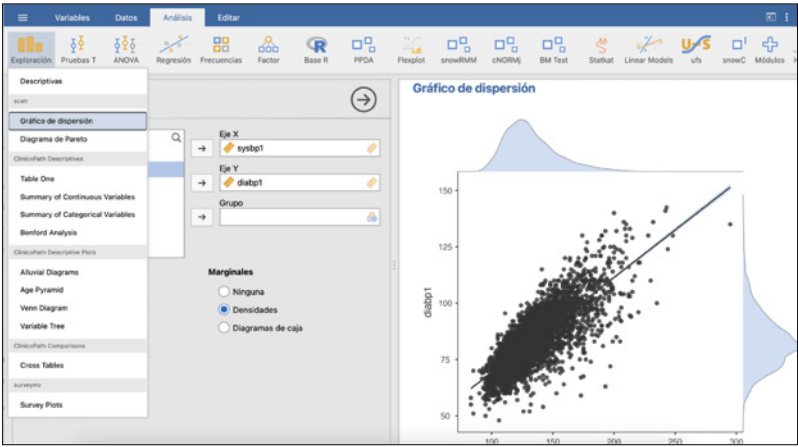


Figura 1. Imatge de Jamovi de diagrama de dispersió

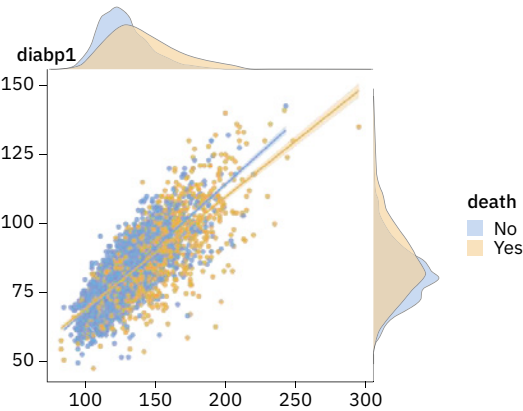


Figura 2. Imatge de Jamovi de diagrama de dispersió, separant per classes

Després de veure gràficament si hi ha relació entre les variables, podem dur a terme l’anàlisi pròpiament dita (figura 3).

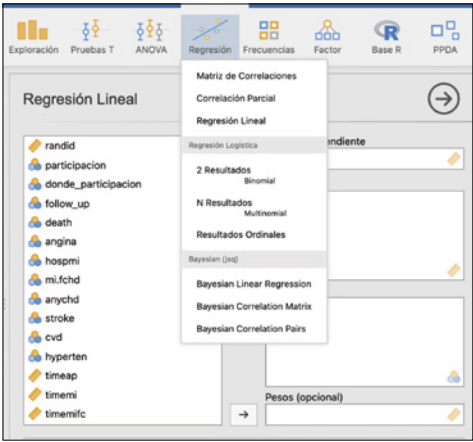


Figura 3. Imatge de Jamovi per efectuar la regressió lineal

Amb tot això, obtenim els següents resultats com a exemple d'aplicació de regressió lineal:

Taula 1. Resultats de la regressió lineal entre pressió arterial sistòlica (sysbp1) i pressió arterial diastòlica (diabp1)

| Mesures d'ajustament del model |         |  |                |  |
|--------------------------------|---------|--|----------------|--|
| Model                          | R       |  | R <sup>2</sup> |  |
| 1                              | 0.78418 |  | 0.61493        |  |

| Coeficients del model - sysbp1 |           |          |         |          |
|--------------------------------|-----------|----------|---------|----------|
| Predictor                      | Estimador | EE       | t       | p        |
| Constant                       | 11.6226   | 1.455594 | 7.9848  | < .00001 |
| diabp1                         | 1.4586    | 0.017338 | 84.1286 | < .00001 |

La correlació entre dues variables quantitatives —el coeficient *r* de Pearson— es pot avaluar sempre, no comporta que hi hagi una variable dependent i una independent, simplement ens mostra si ambdues variables canvien (correlacionen) alhora. En aquest cas ens diu que la correlació entre ambdues variables és bastant elevada (78,418%) i que quan augmenta la pressió arterial diastòlica també ho fa la sistòlica.

En canvi, el model de regressió lineal —com passa també amb el diagrama de dispersió— comporta ja pensar en una variable dependent (*Y* = efecte o variable resposta) i una independent (*X* = variable explicativa o també variable tractament). En aquests casos, la intensitat de la relació ens la dona

el «coeficient de determinació», un estadístic que ens diu quina part de la variabilitat total és explicada per la variable independent. Es tracta d'un coeficient que varia entre 0 i 1, i que en percentatge ens indica precisament això, la proporció de la variabilitat total de Y explicada per X. En aquest cas el coeficient de determinació és igual a 61,493 %. A més, sol ser habitual que ens interressi quina és la funció de la recta de regressió (la recta que JAMOMI ha dibuixat en el gràfic de dispersió). Aquesta funció ens permetrà predir, per a cada valor de la variable independent, el valor que pren, de mitjana, la variable dependent.

A partir de la segona part de la taula, podem obtenir la recta de regressió entre les dues variables ( $Y = a + bX$ ), sent en aquest cas  $a=11,62$  i  $b =1,46$ . El pendent és l'augment de la variable Y en incrementar en una unitat la variable X. En aquest cas direm que, per cada mil·límetre de mercuri en la pressió arterial diastòlica (diabp1), la pressió arterial sistòlica (sysbp1) augmenta, en mitjana, en 1,45 mil·límetres de mercuri. El signe del pendent serà sempre el mateix que el del coeficient de correlació de Pearson, per la qual cosa també indica si es tracta d'una relació directa o inversa. Òbviament, el signe ha de ser el mateix que el pendent de la recta dibuixada en el diagrama de dispersió. A més, tenir la recta ens permet també fer predicció de dades.

A JAMOMI també podem obtenir molta més informació sobre la regressió efectuada, com podem veure en les següents imatges (figura 4).

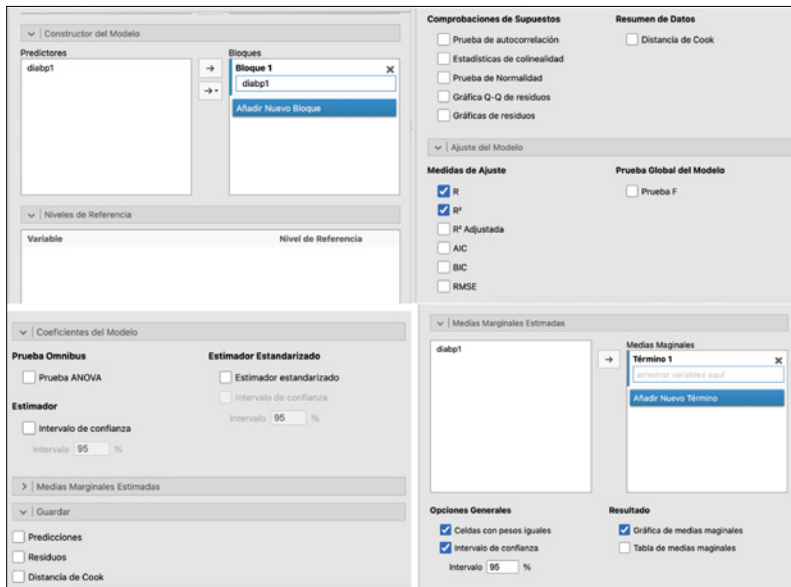
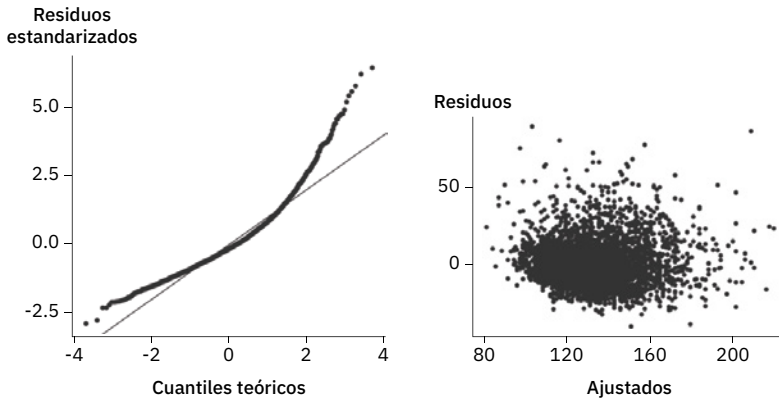


Figura 4. Imatge de Jamovi de les diferents opcions de la regressió lineal

També analitzarem les dades per veure normalitat i homoscedasticitat dels residus.



**Figura 5. Imatge de Jamovi de les diferents opcions per estudiar normalitat i homoscedasticitat**

Per acabar amb aquest apartat, introduïrem la matriu de correlacions, juntament amb la inferència associada a aquesta pròpia anàlisi (figura 6). En aquest cas, podem veure les correlacions entre els diferents valors de la pressió arterial sistòlica i diastòlica en diferents moments temporals.



**Figura 6. Imatge de Jamovi de les diferents opcions per estudiar la matriu de correlacions**

A la taula següent podem veure les correlacions entre cada parell de valors i, a més, el seu p-valor associat, que en tots els casos coincideix amb  $<0,0001$ , per la qual cosa hi ha relació entre les variables. A continuació, també la imatge del gràfic dels diagrames de dispersió de tots els casos amb els seus valors de regressió, juntament amb les densitats de cada cas.

**Taula 2. Matriu de correlacions entre sysbp i diabp en diferents moments**

| Matriu de correlacions |              | 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      | $< .00001$ | $< .00001$ | $< .00001$ | $< .00001$ | $< .00001$ | —      |

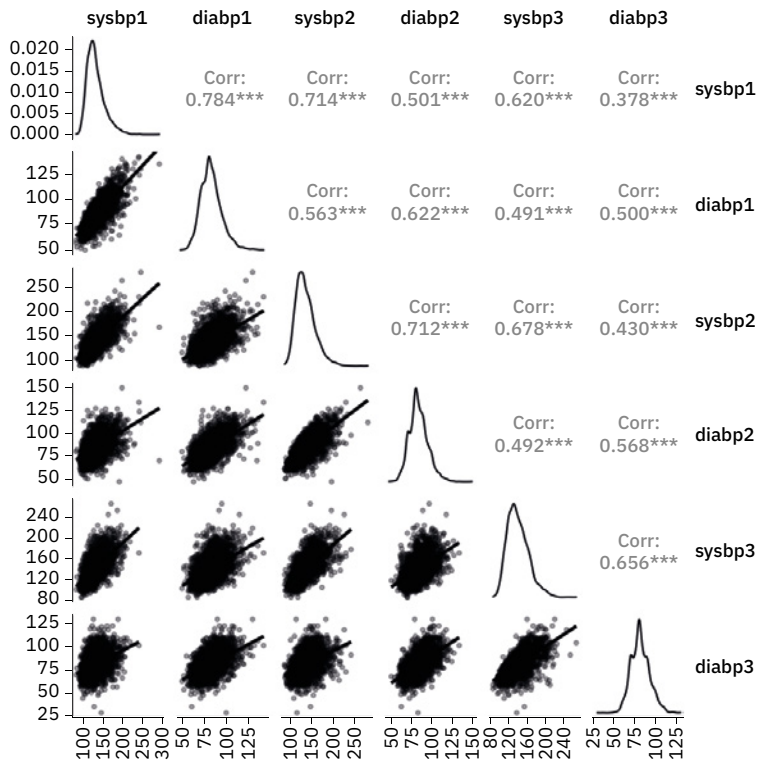


Figura 7. Imatge de JAMOVI de les diferents correlacions

Altres opcions, de les quals només comentarem en aquest punt, però que poden ser desenvolupades, són les diferents relacions que podem obtenir amb les variables d'estudi. En la imatge següent, podem veure que amb el paquet seolmatrix es pot desenvolupar tota mena de correlacions, des de Sperman, fins i tot multinivell (figura 8). A més, ens permet també obtenir la matriu de correlacions i la seva relació de forma gràfica.

| 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 |

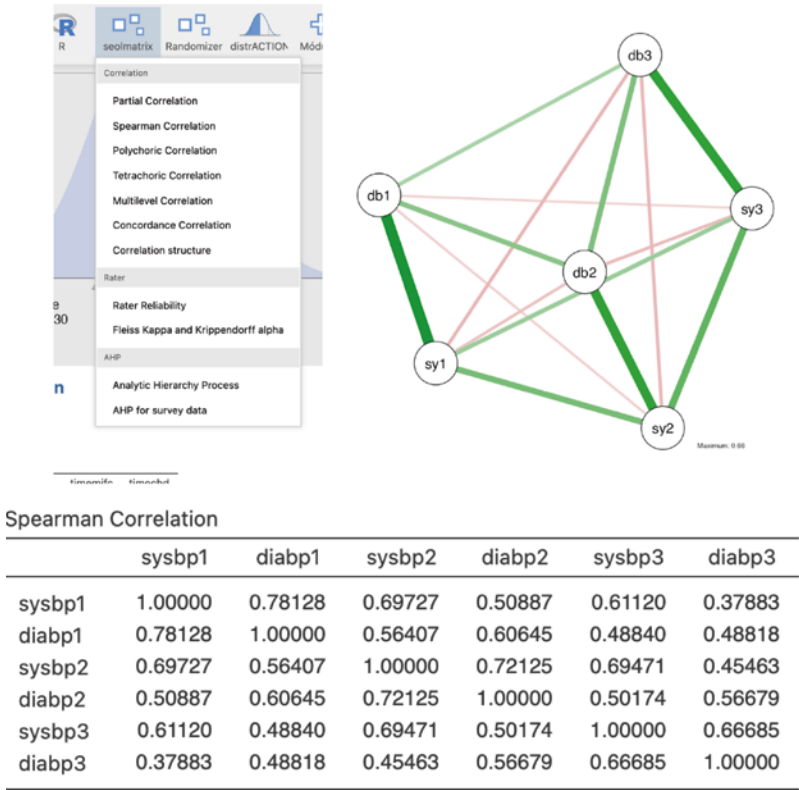


Figura 8. Imatge de JAMOVI de les diferents opcions de seolmatrix (esquerra) i relacions de forma gràfica entre variables (dreta). Part inferior matriu de correlacions obtinguda amb seolmatrix

### Regressió lineal simple

Si ara volem relacionar una variable amb diverses altra, també ho podem fer. A la taula següent, obtenim els resultats d'introduir més d'una variable independent, que fer que millori l'ajustament, augment del valor de R.

Taula 3. Resultats de regressió lineal múltiple

| Mesures d'ajustament del model |         |                |
|--------------------------------|---------|----------------|
| Model                          | R       | R <sup>2</sup> |
| 1                              | 0.87312 | 0.76234        |

Coefficients del model - sysbp1

| Predictor | Estimador | EE       | t        | p        |
|-----------|-----------|----------|----------|----------|
| Constant  | 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 |

Per acabar aquest apartat, elaborarem la regressió lineal múltiple amb el paquet Lineal Models, que ens donarà els mateixos resultats per als coeficients, però ens donarà més opcions (figura 9).

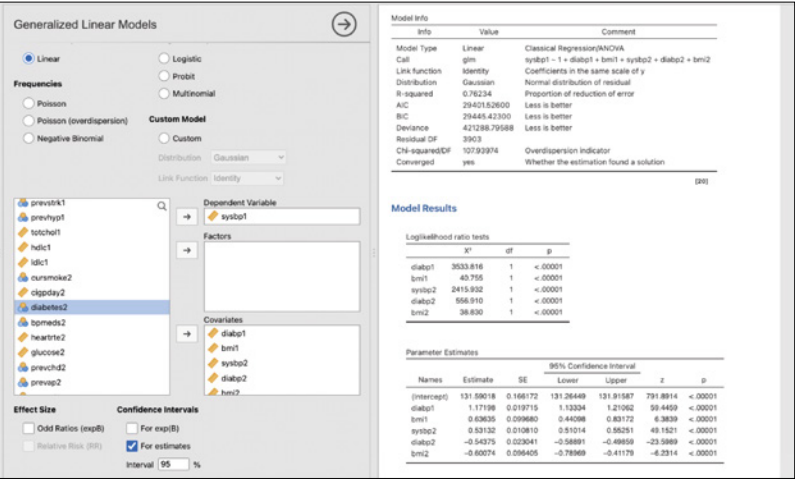


Figura 9. Imatge de Jamovi regressió lineal múltiple amb l'ús del paquet Lineal Model

Veiem que la informació que ens dona (figura 9) és molt més completa, amb la significació de les variables, coeficients i fins i tot tenim l'opció de comparar diversos models alhora.

## Referències

- The jamovi project (2022). *jamovi*. (Versió 2.3) [Computer Software]. Recuperat de <https://www.jamovi.org>.
- R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer Software]. Recuperat de <https://cran.r-project.org>. (R packages recuperat de MRAN snapshot 2022-01-01).
- Linlin Yan (2020). *Venn Diagram by ggplot2, with really easy-to-use API*. [R package]. Recuperat de <https://github.com/yanlinlin82/ggvenn>.
- Serdar Balci (2022). *ClinicoPath jamovi Module doi:10.5281/zenodo.3997188*. [R package]. Recuperat de <https://github.com/sbalci/ClinicoPathJamoviModule>. link.
- Bjoern Koneswarakantha (2019). *easyalluvial: Generate Alluvial Plots with a Single Line of Code*. [R package]. Recuperat de <https://CRAN.R-project.org/package=easyalluvial>.
- Nick Barrowman (2020). *vtree: Display Information About Nested Subsets of a Data Frame*. [R package]. Recuperat de <https://CRAN.R-project.org/package=vtree>.
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., i RStudio (2018). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. [R package]. Recuperat de <https://CRAN.R-project.org/package=ggplot2>.
- Patil, I. (2018). *ggstatsplot: 'ggplot2' Based Plots with Statistical Details*. [R package]. Recuperat de <https://CRAN.R-project.org/package=ggstatsplot>.
- Seol, H. (2023). *seolmatrix: Correlations suite for jamovi*. (Versió 3.7.1) [jamovi module]. URL <https://github.com/hyunsooseol/seolmatrix>.
- Ripley, B., Venables, W., Bates, D. M., Hornik, K., Gebhardt, A., i Firth, D. (2018). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. [R package]. Recuperat de <https://cran.r-project.org/package=MASS>.

## Capítol 10

# Introducció a les regressions amb variables dependents qualitatives

## Regressió logística

La regressió logística és una tècnica d'anàlisi estadística utilitzada per modelar la relació entre una variable dependent binària i una o més variables independents. És particularment útil quan la variable dependent és categòrica i dicotòmica, és a dir, que pot prendre un de dos valors possibles (com «sí» o «no», «veritable» o «fals»).

Les característiques principals de la regressió logística són:

- Variable dependent binària: la variable dependent i és binària, per la qual cosa pren valors de 0 o 1.

Exemples:

Diagnòstic mèdic (malalt = 1, sa = 0)

Resultat d'una campanya publicitària (compra = 1, no compra = 0)

- Risc com a resultat:

La regressió logística estima el risc que ocorre l'esdeveniment d'interès (per exemple, la probabilitat d'aprovar un examen). Els coeficients s'interpreten com el canvi en el risc que s'esdevingui l'esdeveniment per una unitat de canvi en la variable independent, mantenint constants les altres variables.

Com a exemple de regressió logística, pot ser que suposem que estem interessats a predir si un pacient té una malaltia cardíaca (1 = sí, 0 = no) en funció de l'edat i el nivell de colesterol. I com a aplicacions de la regressió logística, **Medicina** (predicció de malalties, risc de complicacions, etc.), **Màrqueting** (probabilitat que un client compri un producte), **Finances** (predicció de la probabilitat d'incompliment de crèdit) o **Sociologia** (estudi de comportaments i característiques demogràfiques). La regressió logística és una eina poderosa en estadística i aprenentatge automàtic per a la classificació binària, oferint una forma clara i interpretable d'entendre les relacions entre les variables i els resultats binaris.

En la següent imatge, tenim les possibilitats del Jamovi per a la regressió logística, de forma genèrica amb l'opció de regressió o amb el paquet Linear Models. Tot i que cal dir que tenim altres paquets, que no desenvoluparem aquí, que també ens donen aquesta opció.

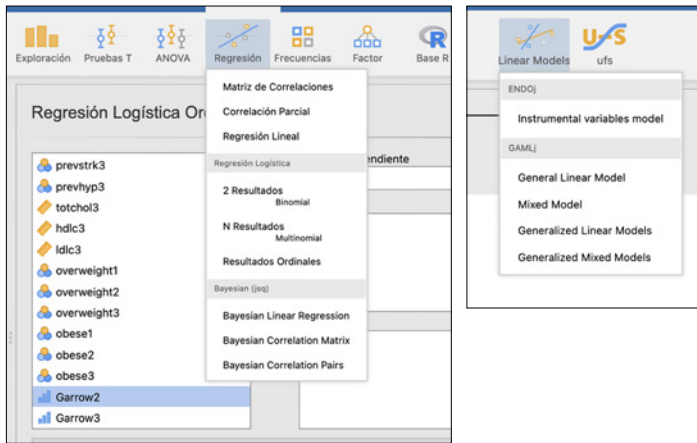


Figura 1. Imatge de Jamovi de les diferents opcions per executar regressió logística

Començarem amb el cas binomial. En la imatge següent, podem veure com analitzar la variable binomial «mort» (*death*) enfront de tres variables qualitatives: sexe (*sex*), edat a l'entrada a la cohort categoritzada en quartils (*age\_basal\_cat*) i un indicador si el subjecte era obès quan es va incorporar a la cohort), les quals s'introdueixen a «Factors»; i una variable quantitativa el nivell de glucosa del subjecte en el moment en què es va incorporar a la cohort (*glucose1*), que s'introdueix a «Covariates» (figura 2).

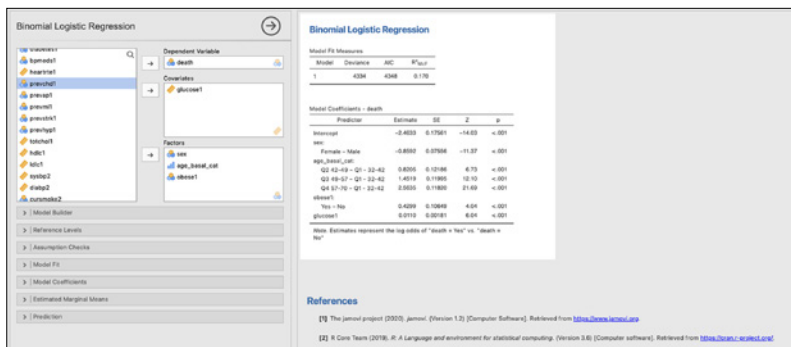
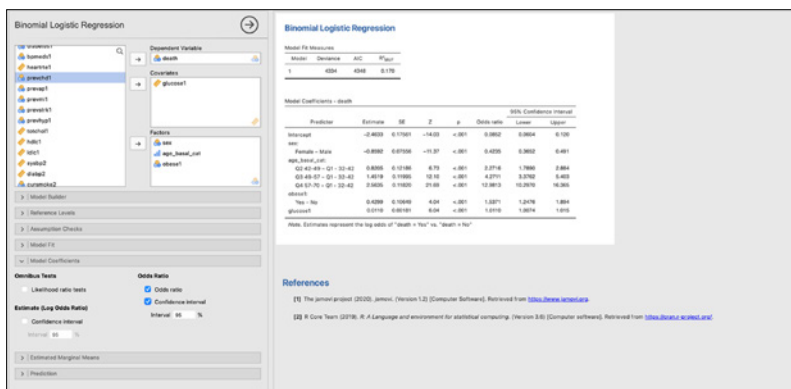


Figura 2. Imatge de Jamovi dels resultats de regressió logística

Podem observar que, de la mateixa manera que teníem a la regressió lineal, es presenten els coeficients, i si les variables són o no significatives. Però, com sabem, més que els coeficients ens interessin els odds ràtios, que en Jamovi s'expressen com a  $\exp(B)$ . Per a això polsarem «Model coefficient» i marcarem els odds ràtio i els seus intervals de confiança (figura 3).



**Figura 3.** Imatge de Jamovi dels resultats de regressió logística demanant-li els odds i els seus intervals de confiança

En aquest senzill exemple, s'interpretaria com segueix. En primer lloc, veiem que tots els odds ràtio són estadísticament significatius, ja que els valors p són menors que 0,05 (si es treballa amb un nivell de significació del 5 %) i, per tant, els intervals dels odds ràtio no contenen el valor 1. Interpretarem a continuació els odds ràtio de les variables qualitatives.

Les dones van tenir un 57,65 % menys de risc de morir que els homes, categoria de referència

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

Observeu que les categories de referència Jamovi les col·loca a la dreta de cada categoria en les variables qualitatives. Si el subjecte era obès en la primera mesura va tenir un 53,71 % més risc de morir que els que no ho eren

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

Observeu que a més edat més risc de morir, sempre respecte als subjectes de 32 a 42 anys (primer quartil). Així, els subjectes de 42 a 49 van tenir un 127,16 % més risc de morir que els subjectes de 32 a 42 anys; els subjectes de 49 a 57 anys, un 327,11 % més risc; i els de 57 a 70 anys un 1198,13 % més risc.

Els odds ràtio de les variables quantitatives s'interpreten de forma diferent. Així, per cada unitat de la variable *glucosa* el risc de morir va augmentar un 1,1 %. Recomanem, això no obstant, categoritzar totes les variables quantitatives, ja que són més fàcils d'interpretar i són més realistes. El risc per unitat de glucosa és el mateix amb un nivell baix que amb un nivell alt?

## Regressió multinomial

Continuem amb el cas en què la variable dependent té més possibilitats, no només dues. Tenim la regressió multinomial, també coneguda com a regressió logística multinomial. És una extensió de la regressió logística binària que s'utilitza per modelar la probabilitat de **més de dues categories** en una variable dependent. Igual que la regressió logística binària, la regressió multinomial prediu probabilitats en lloc de valors absoluts.

Les seves característiques principals són:

- **Variable dependent:** categòrica amb **més de dues categories** (per exemple, tipus de transport: automòbil, autobús, bicicleta).
- **Variables independents:** poden ser quantitatives o categòriques.
- **Objectiu:** estimar la probabilitat que s'esdevingui cadascuna de les categories dependents en funció de les variables independents.
- **Funció de sortida:** probabilitats entre 0 i 1 per a cada categoria.
- **Interpretació de coeficients:** canvi en la probabilitat d'una categoria específica per una unitat de canvi en la variable independent, mantenint constants les altres variables.

En què s'assembla a la binomial:

- Utilitza la funció logística per transformar les probabilitats en valors entre 0 i 1.
- Es basa en el mètode de màxima versemblança per estimar els paràmetres del model.
- S'interpreta la influència de les variables independents a través dels coeficients.

En què es diferencia:

- Maneja més de dues categories en la variable dependent.
- La funció de sortida és un vector de probabilitats, una per a cada categoria.
- La interpretació dels coeficients es refereix a canvis en la probabilitat relativa entre categories, no a canvis absoluts en probabilitats.

Ara, podem veure la taula dels resultats si utilitzem de forma general la regressió multinomial i, amb posterioritat, les imatges (figura 4 i figura 5) dels resultats amb el paquet Lineal Model.

Taula 1: Resultats de regressió multinomial

| Mesures d'ajustament del model |           |        |          |  |  |
|--------------------------------|-----------|--------|----------|--|--|
| Model                          | Desviació | AIC    | R²McF    |  |  |
| 1                              | 8215.5    | 8227.5 | 0.059829 |  |  |

| Coeficients del Model - Garrow1 |           |            |           |          |          |
|---------------------------------|-----------|------------|-----------|----------|----------|
| Garrow1                         | Predictor | Estimador  | EE        | Z        | p        |
| Overweight - Normoweight        | Constant  | -4.2302921 | 0.2667635 | -15.8578 | < .00001 |
|                                 | sysbp1    | 0.0059530  | 0.0024725 | 2.4077   | 0.01605  |
|                                 | diabp1    | 0.0412064  | 0.0046931 | 8.7803   | < .00001 |
| Obese - Normoweight             | Constant  | -8.5380157 | 0.3760676 | -22.7034 | < .00001 |
|                                 | sysbp1    | 0.0076365  | 0.0034580 | 2.2084   | 0.02722  |
|                                 | diabp1    | 0.0743786  | 0.0066900 | 11.1179  | < .00001 |

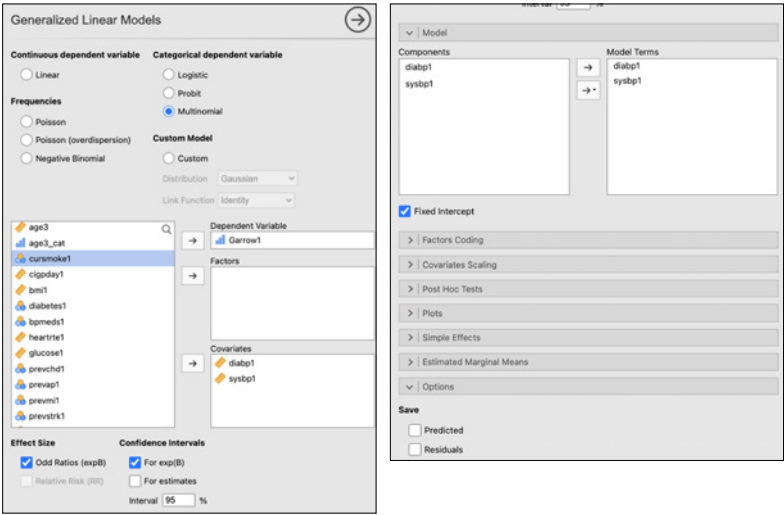


Figura 4. Imatge de Jamovi de les possibilitats de Lineal Models per a regressió múltiple

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 |          |    |         |
|---------------------------|----------|----|---------|
|                           | X²       | df | p       |
| diabp1                    | 151.0451 | 2  | <.00001 |
| sysbp1                    | 7.4806   | 2  | 0.02375 |

| Parameter Estimates      |             |            |           |         |                                |         |           |         |
|--------------------------|-------------|------------|-----------|---------|--------------------------------|---------|-----------|---------|
| Response Contrasts       | Names       | Estimate   | SE        | exp(B)  | 95% Exp(B) Confidence Interval |         | z         | p       |
|                          |             |            |           |         | Lower                          | Upper   |           |         |
| 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 5. Imatge de Jamovi dels resultats de Lineal Models per a regressió múltiple

## Regressió ordinal

En algunes ocasions, encara que el procediment i resultat sigui similar, podem parlar de regressió ordinal. És un mètode estadístic utilitzat per modelar la **relació entre una variable dependent ordinal i una o més variables independents**. A diferència de la regressió lineal i la regressió logística, que s’usen per a variables dependents contínues o binàries respectivament, la regressió ordinal s’empra quan la variable dependent presenta un ordre natural, però les diferències entre les categories no són necessàriament iguals.

Les seves característiques són:

- **Variable dependent:** ordinal (amb categories ordenades, per exemple, nivell de satisfacció: insatisfet, neutral, satisfet).
- **Variables independents:** poden ser quantitatives o categòriques.
- **Objectiu:** estimar la influència de les variables independents en la probabilitat que un cas es classifiqui en cada categoria de la variable dependent ordinal.
- **Funció de sortida:** probabilitats acumulades per a cada categoria.
- **Interpretació de coeficients:** canvi en les probabilitats acumulades per una unitat de canvi en la variable independent, mantenint constants les altres variables.

Com, per exemple, per al cas dels estats d’obesitat en el moment en què el subjecte es va incorporar a la cohort (*Garrow1*) davant les pressions sistòlica (*sysbp1*) i diastòlica (*diabp1*), en el mateix moment. Els resultats els podem veure a la taula següent.

Taula 2. Resultats de regressió ordinal

| Mesures d'ajustament del model                                                              |           |           |          |          |
|---------------------------------------------------------------------------------------------|-----------|-----------|----------|----------|
| Model                                                                                       | Desviació | AIC       | R ^2McF  |          |
| 1                                                                                           | 8213.0    | 8221.0    | 0.060114 |          |
| Nota. La variable dependent 'Garrow1' té el següent ordre: Normoweight   Overweight   Obese |           |           |          |          |
| Coeficients del model - Garrow1                                                             |           |           |          |          |
| Predictor                                                                                   | Estimador | EE        | Z        | p        |
| sysbp1                                                                                      | 0.0057758 | 0.0021349 | 2.7054   | 0.00682  |
| diabp1                                                                                      | 0.0487134 | 0.0040414 | 12.0537  | < .00001 |

## Contrastos no paramètrics

Finalment, comentarem de forma molt breu els contrastos de regressió no paramètrica, que són proves estadístiques emprades per comparar models o avaluar la significança de predictors en un context no paramètric. A diferència dels mètodes paramètrics, que assumeixen una forma específica per a la relació entre les variables (per exemple, lineal), els mètodes no paramètrics no suposen una forma particular per a aquesta relació, cosa que permet una flexibilitat més elevada en modelar dades complexes i no lineals.

Les característiques principals dels contrastos de regressió no paramètrica:

- Flexibilitat: no s'assumeix una forma funcional específica (com lineal o quadràtica) per a la relació entre les variables. S'utilitzen tècniques com la suavització de *kernel*, *splines* o arbres de decisió per modelar la relació.
- Proves d'hipòtesis: les proves no paramètriques solen ser més robustes davant la violació de supòsits comuns en mètodes paramètrics, com la normalitat dels residus o l'homoscedasticitat.

Un exemple és la prova de permutació, on es calculen totes les possibles reordenacions de les dades per avaluar la significança de l'efecte observat.

Els avantatges que té són que no requereix suposicions estrictes sobre la forma funcional de la relació, sol ser més robust davant *outliers* i dades no lineals. Els seus desavantatges són que pot ser computacionalment intensiu, de vegades menys interpretable que els mètodes paramètrics.

En resum, els contrastos de regressió no paramètrica proporcionen una forma flexible i robusta d'avaluar la significança de predictors en models que no imposen suposicions estrictes sobre la relació entre les variables, essent especialment útils en situacions en què les relacions són complexes o no lineals.

Com a exemple, podem veure el cas d'ANOVA d'un factor no paramètric (figura 6 i 7).



Figura 6. Imatge de Jamovi de les possibilitats d’ANOVA

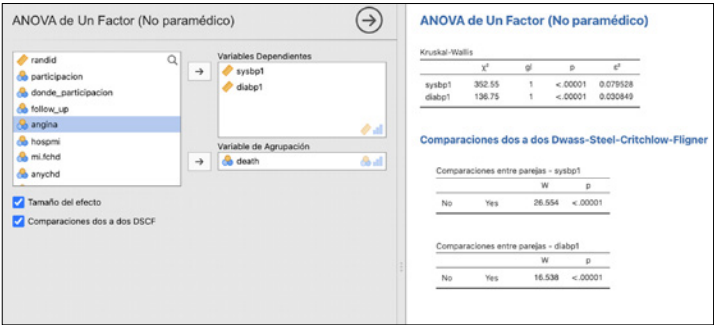


Figura 7. Imatge de Jamovi dels resultats d’ANOVA no paramètric.

## Referències

- Cahusac, P. M. B. (2020). *Evidence-Based Statistics*. John Wiley & Sons. <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119549833>
- Clyde, M. A. (2017). *BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging*. [R package]. Recuperat de <https://cran.r-project.org/package=BAS>.
- Clyde, M. A., Ghosh, J., i Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20, 80-101.
- 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](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)
- Epskamp, S., et al. (2012). *qgraph: Network Visualizations of Relationships in Psychometric Data*. [R package]. Recuperat de <https://CRAN.R-project.org/package=qgraph>.
- Gallucci, M. (2019). *GAMLj: General analyses for linear models*. [Jamovi module]. Recuperat de <https://gamlj.github.io/>.
- Glover, S. (2018). *Likelihood Ratios; A Tutorial*. <https://osf.io/preprints/metaarxiv/g3j2k/>
- JASP Team (2018). *JASP*. [Computer Software]. Recuperat de <https://jasp-stats.org>.
- R Core Team (2021). *R: A Language and environment for statistical computing*. (Versió 4.1) [Computer Software]. Recuperat de <https://cran.r-project.org>. (R packages recuperat de MRAN snapshot 2022-01-01).
- Revelle, W. (2019). *psych: Procedures for Personality and Psychological Research*. [R package]. Recuperat de <https://CRAN.R-project.org/package=psych>.
- Ripley, B. i Venables, W. (2016). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. [R package]. Recuperat de <https://cran.r-project.org/package=nnet>.
- The Jamovi project (2022). *Jamovi*. (Versió 2.3) [Computer Software]. Recuperat de <https://www.Jamovi.org>.



Aquest manual suposa una bona introducció al treball i a l'anàlisi estadística, així com a la seva comprensió per part del professorat, del personal investigador i, especialment, de l'alumnat, que necessita suport per utilitzar eines estadístiques de manera senzilla i clara. Més enllà de l'anàlisi descriptiva, *Jamovi* proporciona una experiència d'aprenentatge accessible i enriquidora per a persones usuàries de tots els nivells. El seu enfocament intuïtiu i les seves capacitats avançades el converteixen en un aliat indispensable per a l'anàlisi de dades en l'àmbit acadèmic i professional. Tot això és el que ofereix *Jamovi*.

Amb el suport de:



Generalitat de Catalunya  
**Departament de Recerca  
i Universitats**

  
**Universitat  
de Girona**