Practica 1

Estadística descriptiva amb fulls de càlcul

Els full de càlcul són una eina pel tractament de dades a l'abast de tothom. Entre els més utilitzats hi trobem MS Excel, LibreOffice Calc i OpenOffice Calc. En aquesta pràctica veurem com analitzar i descriure dades amb LibreOffice Calc. La mateixa anàlisi es pot fer amb els altres fulls de càlcul seguint instruccions semblants. Amb Excel s'ha d'activar el complement *Anàlisi de Dades*.

En les primeres seccions treballarem amb l'arxiu `edatestatura.ods` que trobeu al Moodle. El full `dades` conté quatre variables d'interès, *sexe*, *edat*, *estatura* i *grup.edat*, més una columna d'identificador del cas (*id*). Les variables *edat* i *estatura* són **numèriques**, amb un nombre elevat de valors diferents; *sexe* i *grup.edat* són **categòriques**, és a dir, classifiquen els individus en un reduït nombre de grups. El full `etiquetes` explica quin són aquests grups i la nomenclatura utilitzada.

1 Funcions

Comencem preparant les dades del nostre arxiu per a la seva anàlisi. Per transformar les variables utilitzem les funcions pròpies del full de càlcul. Podem accedir a la llibreria de funcions des de la barra de fórmules (símbol $\mathbf{f_x}$) o escriure el símbol = seguit del nom de la funció i els seus paràmetres.

Exercici 1. Creeu un nou full anomenat `dades2` amb la mateixa informació del full `dades`. Substituïu *sexe* i *grup.edat* per dues variables noves, *genere* i *grup*, respectivament, que continguin els valors alfanumèrics (les etiquetes) corresponents.
Indicació: utilitzeu la funció lògica `SI` *i tingueu en compte que els valors no numèrics van entre cometes.*

Si el output de l'operació que realitzem és en vàries cel·les, parlem de **funció matricial**. Observeu que si executem una operació d'aquest tipus només s'omple una casella. Per resoldre aquest problema, seleccioneu tot el rang de sortida, escriviu la funció a la barra de fórmules i apliqueu la combinació de tecles: `Control+Shift+Enter`.

Exercici 2. Obriu un nou fitxer amb LibreOffice Calc.

a) Escriviu en 4 caselles una matriu 2×2 invertible qualsevol i calculeu la seva inversa amb la funció `MINVERSA`.

b) Calculeu la transposada de la matriu inicial amb la funció `TRANSPONER`.

c) Multiplique la matriu inicial per la seva inversa amb la funció `MMULT`.

2 Estadística descriptiva univariant

L'objectiu de l'estadística descriptiva és la recollida, classificació, representació i resum de les dades. Es divideix en estadística univariant, que descriu singularment les variables, i bivariant, que analitza les relacions entre elles.

2.1 Resums numèrics

Considerem un conjunt de n dades $x_1, \ldots, x_n$. A continuació trobeu un resum dels estadístics més freqüents per descriure i analitzar una variable.

- Mesures de centre:
  - La **mitjana** és el valor que s'obté dividint la suma de totes les dades entre el nombre d'observacions:
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
  - La **mediana** és el valor que queda al centre una vegada ordenat el conjunt de dades de més petit a més gran. Si n és parell, s'agafen els dos valors centrals i es calcula la seva mitjana.
  - La **moda** és el valor que té més freqüència al conjunt de dades.
- Mesures de dispersió:
  - La **variància** es defineix com la mitjana de les distàncies al quadrat entre les dades i la seva mitjana:
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$
  - La **dispersió típica** és l'arrel de la variància:
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$
  - La **variància corregida**:
$$\widetilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$
    i la corresponent **dispersió típica corregida**: $\widetilde{s} = \sqrt{\widetilde{s}^2}$.
  - El **coeficient de variació**:
$$CV = \frac{s}{\bar{x}}.$$
    S'acostuma a demanar $\bar{x} > 0$. Serveix per comparar la dispersió de dues variables i no té unitats.
- Mesures de posició:
  - Donat un $p\%$, el **p-percentil** és el valor numèric tal que almenys un $p\%$ de les dades observades són inferiors o iguals a aquest valor i almenys el $(100 - p)\%$ de les dades observades són superiors o iguals a aquest valor.
  - El percentil per a $p = 25$, s'anomena **quartil inferior (Q1)**.
    El percentil per a $p = 50$ és la mediana (**Q2**).
    El percentil per a $p = 75$, s'anomena **quartil superior (Q3)**.
- Mesures de forma:
  - El **coeficient d'asimetria** mesura el grau de simetria del conjunt de dades i es defineix com
$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3}.$$

– El **coeficient de curtosi** indica el grau d'apuntament de les dades i es defineix com

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4} - 3.$$

Exercici 3. Al full `dades2` calculeu els següents estadístics per la variable *edat* utilitzant funcions bàsiques no estadístiques:

a) La mitjana.

b) La variància i la desviació típica.

c) La variància utilitzant la fórmula $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$.
Indicació: Utilitzeu la funció `SUMA.PRODUCTO` *que dóna el producte escalar de dos vectors (columnes).*

d) La variància corregida i la corresponent desviació típica.

Exercici 4. Calculeu novament els estadístics de l'exercici anterior utilitzant les funcions estadístiques següents: `PROMEDIO, VAR, VARP, DESVEST`.

Exercici 5. Calculeu el coeficient de d'asimetria i el coeficient de curtosi de la variable *edat*. Utilitzeu les funcions `COEFICIENTE.ASIMETRIA` i `CURTOSIS`, respectivament.
Observació: El coeficient d'asimetria que calcula el full de càlcul és una modificació de γ*:* $\frac{\sqrt{n(n-1)}}{n-2}\gamma$*. El coeficient d'apuntament que calcula el full de càlcul és una versió modificada i normalitzada de* K*.*

Una manera més ràpida d'obtenir el resum dels estadístics més importants és seguint les instruccions

`Datos` $\rightsquigarrow$ `Estadísticas` $\rightsquigarrow$ `Estadísticas descriptivas...`

Exercici 6. Obteniu el resum dels estadístics de la variable *edat*.

2.2 Histograma

Quan tractem amb una variable numèrica amb gran varietat de valors, té sentit organitzar les dades en intervals. Entre les gràfiques més apropiades hi trobem l'histograma, que representa les freqüències dels intervals (classes).

Exemple 1. Representem l'histograma de la variable *edat*.
Al resum numèric podem observar que el mínim d'edat és 15 i el màxim 64. Considerarem els intervals següents: 15-24, 25-34, 35-44, 45-54 i 55-64. Observeu que en realitat, fent la correcció de continuïtat, els intervals són: $[15, 25)$, $[25, 35)$, $[35, 45)$, $[45, 55)$ i $[55, 65)$, i tenen amplitud 10. Aquestes són les nostres classes.
Creeu un nou full anomenat `dades.edat` i copieu-hi la variable *edat*.
Creeu dues columnes auxiliars: una que anomeneu *extrsup* i que conté els extrems superiors dels intervals d'edat que volem, i una que anomeneu *intervals* que conté el nom de les classes. Calculeu la freqüència de cada classe utilitzant la funció `FRECUENCIA(Datos, Clases)`, on `Datos` són les dades de la variable edat i les `Clases` estan definides pel extrems superiors dels intervals. Observeu que és una funció matricial.
Seleccioneu la columna de freqüències i seguiu les següents instruccions per crear l'histograma:

`Insertar` $\rightsquigarrow$ `Diagrama...` $\rightsquigarrow$ `Tipo de diagrama: Columna`

A la pestanya `Series de datos` escriviu a `Categorias` les cel·les que contenem el nom de les classes i a la pestanya `Elementos del diagrama` afegiu el títol i les etiquetes que trobeu oportunes.
Un cop creat el gràfic, cliqueu a sobre de les barres amb el botó dret, seleccioneu `Formato de series de datos` i reduïu l'espai entre columnes a 0.

2.3 Taules dinàmiques

Una manera d'analitzar dades categòriques amb els fulls de càlcul és utilitzar les taules dinàmiques. Després de seleccionar les dades que volem estudiar, seguiu les instruccions:

`Insertar` $\rightsquigarrow$ `Tabla dinámica...`

Us apareixerà un requadre que conté les següents pestanyes:
`Campos disponibles`: Variables preses en consideració.
`Campos de fila/columna`: Variables categòriques amb les quals volem filtrar les dades.
`Campos de datos`: Variables numèriques i estadístics que volem estudiar.

Exercici 7. Considereu les dades contingudes al full `dades2`. Calculeu el mínim, el màxim i la mitjana de la variable *estatura* reagrupant les dades segon el *genere*.

Exercici 8. Repetiu l'exercici anterior, aquest cop reagrupant les dades segon el *genere* i el *grup*.

2.4 Taules de contingència

Les taules de contingència serveixen per analitzar la relació entre dues o més variables, habitualment categòriques. Contenen les freqüències amb que apareixen les diferents categories de les variables. Per obtenir una taula de contingència es segueixen les mateixes instruccions que per crear una taula dinàmica. En aquest cas, es selecciona l'opció `Conteo` al `Campo de datos`.

Exercici 9. Genereu dues taules de contingència per veure la relació entre les variables *genere* i *grup*: una amb les freqüències absolutes i l'altre amb les freqüències relatives.

2.5 Regressió lineal simple

Considerem un conjunt de n dades aparellades: $(x_1, y_1), \ldots, (x_n, y_n)$. La **covariància** es defineix com la mitjana dels productes de les desviacions de les dades respecte a la seva mitjana:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}).$$

Amb un full de càlcul podem calcular la matriu de covariàncies seguint les instruccions:

`Datos` $\rightsquigarrow$ `Estadísticas` $\rightsquigarrow$ `Covarianza...`

Les dades es poden representar amb un núvol de punts o **diagrama de dispersió** (scatterplot). En LibreOffice Calc el gràfic s'obté seguint les següents instruccions:

`Insertar` $\rightsquigarrow$ `Diagrama...` $\rightsquigarrow$ `Tipo de diagrama: XY (dispersión)`

És interessant determinar si hi ha una certa dependència lineal entre les dades. El **coeficient de correlació de Pearson** dóna aquest grau d'ajust lineal. Es defineix com la covariància entre les dues variables dividit pel producte de les seves desviacions típiques:

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y}.$$

És fàcil comprovar que $-1 \leq r_{xy} \leq 1$. El signe de r_{xy} indica si la tendència és creixent (+) o decreixent (-). L'ajust lineal és millor quan més proper a 1 és en valor absolut ($|r_{xy}| \approx 1$). Es calcula seguint les següents instruccions:

`Datos` $\rightsquigarrow$ `Estadísticas` $\rightsquigarrow$ `Correlación...`

Quan les dades presenten una forta correlació lineal es poden calcular els coeficients $a, b \in \mathbb{R}$ de manera que la recta d'equació $y = a + bx$ sigui una bona aproximació al conjunt de punts. La **recta de regressió per mínims quadrats** és la recta que millor aproxima aquests punts en el sentit que minimitza les distàncies entre els punts i la recta. En altres paraules, és la recta que té els paràmetres a i b que fan mínima la funció

$$F(a,b) = \sum_{i=1}^{n} (y_i - (a + bx_i))^2.$$

Els paràmetres es calculen com

$$b = \frac{s_{xy}}{s_x^2} \qquad \text{i} \qquad a = \bar{y} - b\bar{x}.$$

El **coeficient de determinació** R^2, que (només) en regressió simple coincideix amb el quadrat del coeficient de correlació (r_{xy}^2) indica la proporció de variabilitat de y explicada per x.
En el full de càlcul podem afegir la recta de regressió al nostre diagrama de dispersió clicant amb el botó dret a sobre del conjunt de punts i seleccionant l'opció `Insertar línea de tendencia`.
També podem calcular els paràmetres seguint les instruccions:

`Datos` $\rightsquigarrow$ `Estadísticas` $\rightsquigarrow$ `Regresión...`

o amb les funcions `PENDIENTE(DatosY;DatosX)` i `INTERSECCIÓN.EJE(DatosY;DatosX)`

Exercici 10. Considereu les dades del fitxer `vidre.ods` que trobeu al Moodle. Aquest document conté informació sobre l'índex de refracció i la densitat de 18 diferents peces de vidre.

a) Calculeu la covariància entre les variables.

b) Dibuixeu el diagrama de dispersió de la densitat amb respecte de l'índex de refracció.

c) Comproveu que la correlació lineal és força bona.

d) Afegiu la recta de regressió per mínims quadrats al diagrama de dispersió i determineu-ne els paràmetres.

e) Calculeu el coeficient de determinació. Considereu que és un bon model lineal?

f) Feu la predicció de la densitat per a un índex de refracció de 1.521.

Exercici 11. Repetiu l'exercici anterior amb les dades d'estatura versus edat del fitxer `edatestatura.ods`. Què opineu de la qualitat de l'ajust?