A2: Analítica Descriptiva i Inferencial

Enunciat

Contents

1 Introducció

Les dades que s'analitzen en aquesta activitat corresponen a l'anàlisi de satisfacció laboral treballat en l'activitat 1. L'objectiu és investigar la satisfacció laboral de la població en relació a les altres variables.

L'arxiu es denomina *rawData_clean.csv*, i conté 1920 registres i 12 variables. Aquestes variables són: city, sex, educ_level, job_type, happiness, age, seniority, sick_leave, sick_leave_b, work_hours, Cho_initial, Cho_final

Les dades recollides per a cada treballador són:

1. La ciutat del lloc de treball
2. El sexe del treballador
3. El nivell d'estuis del treballador (1: sense estudis, 2: estudis primaris, 3: educació secundària o educació professional, 4: universitaris)

4. El tipus de treball (C: qualificat, PC: poc qualificat)
5. La satisfacció laboral del treballador (compresa entre 0 i 10)
6. L'edat (anys)
7. L'antiguitat (en anys)
8. Els dies de baixa en el darrer any laboral
9. Si el treballador va estar de baixa o no en el darrer any (variable binària)
10. Les hores que treballa a la setmana en promig
11. El nivell de colesterol en la penúltima revisió mèdica (mmol/L)
12. El nivell de colesterol en la darrera revisió mèdica (mmol/L).

Nota important a tenir en compte per a lliurar l'activitat:

- És necessari lliurar el fitxer Rmd i el fitxer de sortida (PDF o html). El fitxer de sortida ha d'incloure el codi i el resultat de la seva execució (pas a pas). S'ha de respectar la numeració dels apartats de l'enunciat.
- No realitzeu llistat dels conjunts de dades, donat que aquests poden ocupar varies pàgines. Si voleu comprovar l'efecte d'una instrucció sobre les dades, podeu usar la funció head que mostra les 10 primeres files del conjunt de dades.

2 Analítica descriptiva

2.1 Càrrega del fitxer

Carregueu l'arxiu de dades en R. Independentment del fitxer que vàreu obtenir en l'activitat 1, useu el fitxer "rawData_clean.csv". Un cop carregat el fitxer, valideu que els tipus de dades són els correctes. Si no és així, feu les conversions de tipus oportunes.

2.2 Anàlisi descriptiva visual

Representeu de manera gràfica les variables del conjunt de dades, de manera que es pugui copsar de manera visual i resumida els valors i distribucions de les variables del conjunt de dades. Escolliu els gràfics més apropiats en cada cas.

3 Estadística inferencial

3.1 Intervals de confiança

Representeu, si no ho heu fet en la secció anterior, la distribució de valors de la variable happiness en un histograma. Interpreteu el gràfic.

A continuació, calculeu l'interval de confiança del 90% de la satisfacció laboral (happiness) dels treballadors. També, expliqueu quina és la interpretació d'interval de confiança de la variable happiness.

Nota: S'han de realitzar els càlculs manualment. No es poden usar funcions d'R que calculin directament l'interval de confiança com t.test o similar. Sí que podeu usar funcions com qnorm, pnorm, qt i pt.

3.2 Satisfacció en el treball en relació al sexe

Ens preguntem si existeixen diferències significatives en el grau de felicitat (happiness) dels homes en relació a les dones. Per a respondre a aquesta pregunta, seguiu els passos que es detallen a continuació.

Nota: S'han de realitzar tots els càlculs manualment. No es poden usar funcions R que calculin directament el contrast com t.test o similar. Sí que podeu usar funcions com: **qnorm**, **pnorm**, **qt** i **pt**.

3.2.1 Boxplot

Mostreu la distribució de happiness d'homes i dones per separat en un boxplot. Expliqueu si considereu que hi ha diferències apreciables visualment.

3.2.2 Escriure la hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i alternativa tenint en compte que es vol testejar si el nivell de satisfacció (happiness) és diferent en homes i dones.

3.2.3 Mètode

Indiqueu quin és el mètode més apropiat per a fer aquesta anàlisi, en funció de les característiques de la mostra i l'objectiu de l'anàlisi.

3.2.4 Calcular l'estadístic de contrast, el valor crític i el valor p.

Realitzeu els càlculs amb un nivell de confiança del 95%.

3.2.5 Interpretar el resultat

3.3 Test no paramètric

Quan les assumpcions de normalitat no es compleixen, cal aplicar tests no paramètrics. El **test de suma de rangs de Wilcoxon**, també anomenat **prova U de Mann-Whitney** és l'equivalent no paramètric al test t per a dues mostres independents. En aquest apartat us demanem que apliqueu aquest test. Es realitzarà per a contrastar si hi ha diferències en l'antiguitat (seniority) en funció del tipus de treball (job_type).

Podeu consultar:

- https://www.r-bloggers.com/wilcoxon-mann-whitney-rank-sum-test-or-test-u/

3.3.1 Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i l'alternativa. Recordeu que la pregunta és si hi ha diferències significatives en seniority segons el tipus de treball.

3.3.2 Assumpció de normalitat

Comproveu si es compleix l'assumpció de normalitat en les dades. Per a fer-ho, podeu mostrar els gràfics **Q-Q plots** i aplicar a continuació el test Shapiro-Wilk, interpretant els resultats.

Per a realitzar aquest apartat, podeu consultar la informació següent:

- Quantile-Quantile Plots: https://www.rdocumentation.org/packages/stats/versions/3.5.3/topics/qqnorm
- Interpretació dels Q-Q plots: https://data.library.virginia.edu/understanding-q-q-plots/
- Test Shapiro-Wilk: https://es.wikipedia.org/wiki/Test_de_Shapiro
- Funció shapiro.test d'R: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html

3.3.3 Test U de Mann-Whitney

Apliqueu el test U de Mann-Whitney per a testejar si hi ha diferències significatives en l'antiguitat en funció del tipus de treball. Podeu usar la funció **wilcox.test**.

3.3.4 Càlculs manuals

Per tal d'aprofundir en les operacions del test de suma de rangs de Wilcoxon, realitzeu els càlculs de la suma de ranks del test U de Mann-Whitney i realitzeu el contrast. Podeu consultar l'explicació pas a pas de les fórmules a:

https://www.monografias.com/trabajos106/prueba-u-mann-whitney-dos-muestras-independientes/prueba-u-mann-whitney-dos-muestras-independientes.shtml

3.3.5 Interpretació

Interpreteu el resultat del test.

3.3.6 Reflexió sobre tests paramètrics i no paramètrics

Si els tests no paramètrics no realitzen assumpcions sobre la normalitat de les dades, i per tant, són menys restrictius en la seva aplicabilitat, per què no s'usen els tests no paramètrics com a primera opció (o opció per defecte)?

3.4 Test sobre el nivell de colesterol

A continuació ens preguntem si el nivell de colesterol dels treballadors ha augmentat en l'última revisió en relació a la penúltima revisió. Apliqueu un test adient per a realitzar aquest contrast.

Nota: S'han de realitzar els càlculs manualment. No es poden usar funcions d'R que calculin directament l'interval de confiança com t.test o similar. Sí que podeu usar funcions com qnorm, pnorm, qt i pt.

3.4.1 Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i alternativa, sabent que la pregunta és si el nivell de colesterol en l'última revisió és superior que en la penúltima revisió.

3.4.2 Mètode

Escriviu el mètode que usareu per a realitzar el contrast i les assumpcions sobre les dades que es realitzen per a aplicar el mètode. Apliqueu si s'escau tets per a validar les assumpcions sobre les dades.

3.4.3 Càlculs

Realitzeu tots els càlculs que s'escaiguin: estadístic de contrast, valor crític, valor p...

3.4.4 Interpretació

Interpreteu el resultat.

4 Reflexió final

Elaboreu unes conclusions finals sobre l'anàlisi realitzat. Podeu parlar dels resultats obtinguts en aquesta anàlisi així com una reflexió més general sobre la idoneïtat dels tests i els tipus de tests que hem emprat. No cal excedir la mitja pàgina.

5 Puntuacions dels apartats

- Apartat 2 (10%)
- Apartat 3.1 (10%)
- Apartat 3.2 (20%)
- Apartat 3.3 (20%)
- Apartat 3.4 (20%)
- Apartat 4 (10%)
- Qualitat de l'informe dinàmic (10%)