Activitat 2: Anàlisi descriptiva i inferencial

Enunciat

Estadística Avançada - Semestre 2023.2

Índice

Introducció

En aquesta activitat aplicarem una anàlisi descriptiva i inferencial sobre el conjunt de dades Pisa que hem preprocessat a l'activitat anterior.

El Programa per a l'Avaluació Internacional d'Estudiants (PISA) és una prova que s'aplica cada tres anys a estudiants de 15 anys de tot el món per avaluar-ne el rendiment en matemàtiques, lectura i ciències. Aquesta prova proporciona una forma quantitativa de comparar el rendiment acadèmic dels estudiants de diferents parts del món.

El conjunt de dades pisa_clean.csv conté informació sobre la demografia i les escoles dels estudiants nord-americans que fan l'examen, derivada dels arxius de dades d'ús públic PISA del 2009 distribuïts pel Centre

Nacional d'Estadístiques Educatives (NCES) dels Estats Units. Cada fila del conjunt de dades conté la següent informació d'un estudiant:

- grade: El curs que realitza l'estudiant (la majoria dels estudiants de 15 anys als Estats Units són al desè curs)
- male: Si l'estudiant és home (1/0)
- raceeth: La composició de raça/ètnia de l'estudiant
- preschool: Si l'estudiant va assistir a preescolar (1/0)
- expectBachelors: Si l'estudiant espera fer un grau universitari (1/0)
- motherHS: Si la mare de l'estudiant va completar l'escola secundària (1/0)
- motherBachelors: Si la mare de l'estudiant va obtenir una llicenciatura (1/0)
- motherWork: Si la mare de l'estudiant té feina a temps parcial o complet (1/0)
- fatherHS: Si el pare de l'estudiant va completar l'escola secundària (1/0)
- fatherBachelors: Si el pare de l'estudiant va obtenir una llicenciatura (1/0)
- fatherWork: Si el pare de l'estudiant té feina a temps parcial o complet (1/0)
- selfBornUS: Si l'estudiant va néixer als Estats Units (1/0)
- motherBornUS: Si la mare de l'estudiant va néixer als Estats Units (1/0)
- fatherBornUS: Si el pare de l'estudiant va néixer als Estats Units (1/0)
- englishAtHome: Si l'estudiant parla anglès a casa (1/0)
- computerForSchoolwork: Si l'estudiant té accés a un ordinador per fer les tasques escolars (1/0)
- read30MinsADay: Si l'estudiant llegeix per plaer durant 30 minuts/dia (1/0)
- minutesPerWeekEnglish: El nombre de minuts per setmana que l'estudiant dedica a classe d'anglès.
- studentsInEnglish: El nombre d'estudiants a classe d'anglès d'aquest estudiant a l'escola.
- schoolHasLibrary: Si l'escola d'aquest estudiant té una biblioteca (1/0)
- publicSchool: Si aquest estudiant assisteix a una escola pública (1/0)
- urban: Si l'escola d'aquest estudiant està en una àrea urbana (1/0)
- schoolSize: El nombre d'estudiants a l'escola.
- readingScore: puntuació de lectura de l'estudiant, a una escala de 1000 punts.

A tenir en compte per fer l'activitat:

- Cal 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 l'execució del codi (pas a pas).
- Per facilitar la correcció, els fitxers s'han de lliurar per separat a l'aula. És a dir, es recomana no pujar un fitxer comprimit sinó cada fitxer per separat.
- Cal respectar la mateixa numeració dels apartats que l'enunciat.
- No es poden realitzar llistats complets del conjunt de dades a la solució. Això generaria un document amb centenars de pàgines i dificulta la revisió del text. Per comprovar les funcionalitats del codi sobre les dades, es poden fer servir les funcions **head** i **tail** que només mostren unes línies del fitxer de dades.
- Es valora la precisió dels termes utilitzats (cal utilitzar de manera precisa la terminologia de l'estadística).

- Es valora també la concisió a la resposta. No es tracta de fer explicacions gaire llargues o documents molt extensos. Cal explicar-ne el resultat i argumentar la resposta a partir dels resultats obtinguts de manera clara i concisa.

---

1 Estadística Descriptiva

En primer lloc, fem una anàlisi descriptiva d'algunes variables d'interès i la relació amb readingScore. Seguiu els passos que s'indiquen a continuació.

1.1 Distribució de variables

En primer lloc, mostreu visualment la distribució de gènere a la població, així com la proporció dels estudiants que parlen anglès a casa en relació als que no parlen anglès. Mostreu un gràfic per a cada cas.

1.2 Anàlisi descriptiva de readingScore

Mostreu visualment la distribució de la variable readingScore. Interpreteu el gràfic.

1.3 Distribució de readingScore segons variables d'interès

Mostreu visualment la distribució de la variable readingScore en funció de les variables següents:

- male
- EnglishAtHome
- read30MinsADay
- urban
- selfBornUS
- minutesPerWeekEnglish
- schoolSize

Escolliu el tipus de gràfic que sigui més apropiat en cada cas i interpreteu el resultat.

---

2 Interval de confiança de readingScore

Calculeu l'interval de confiança del valor mitjà de readingScore al 95% i al 97%. Interpreteu el resultat.

Requisits:

- Implementeu una funció que calculi l'interval de confiança i que pugueu utilitzar per obtenir l'IC per al nivell de confiança de 95% i 97% respectivament.
- No podeu fer servir funcions d'R que calculin l'interval de confiança. Sí podeu utilitzar funcions per calcular els valors de la distribució corresponent, com *qt*, *qnorm*, *pt*, *pnorm*.

---

3 Anàlisi de factors que influeixen en readingScore

Realitzarem un contrast per avaluar les diferències en readingScore en relació amb les variables d'interès identificades. En concret, ens preguntem si hi ha diferències significatives en el valor mitjà de readingScore en funció de: gènere, si es parla anglès a casa, si l'estudiant llegeix 30 minuts per plaer, zona urbana, si l'estudiant ha nascut als Estats Units, segons el nombre de minuts per setmana d'anglès i la mida de l'escola.

Per estudiar la influència de la variable del nombre de minuts per setmana que l'estudiant dedica a la classe d'anglès, aquesta variable es convertirà en dos valors: elevat (per a valors superiors a 250) i baix (la resta de valors). Farem el mateix per a la variable grandària de l'escola: valor gran (mides superiors a 1000) i petit (la resta de valors).

Per fer aquesta anàlisi de forma sintètica, s'implementarà una funció de contrast que rebrà dues mostres de dades (i si cal, altra informació addicional) i tornarà els resultats del test (valor p, valor crític, valor observat). S'haurà d'usar aquesta funció per calcular si hi ha diferències en readingScore per a cada variable d'interès esmentada. A continuació, cal resumir la informació en una taula.

Requisits:

- No podeu fer servir funcions que calculin el contrast. Sí podeu fer servir funcions com *qt*, *qnorm*, *pt*, *pnorm*.
- La taula ha de contenir quatre columnes: valor observat, valor crític, valor p i breu comentari sobre si la diferència és significativa. Cada fila serà el resultat del contrast de readingScore segons una variable dinterès. Heu d'indicar a cada fila a quina variable fa referència.

Seguiu els passos que s'indiquen a continuació.

3.1 Hipòtesis

Escriviu quines són les hipòtesis nul · la i alternativa (no cal escriure aquestes hipòtesis per a totes les variables, ja que apliquem el mateix tipus d'hipòtesis per a cada cas).

3.2 Tipus de contrast

Especifiqueu quin tipus de contrast aplicareu i la seva justificació. Si hi ha NA, no inclogueu les dades en el contrast.

3.3 Funció de contrast

Implementeu els càlculs necessaris per realitzar el contrast, seguint el procés indicat més amunt. Calculeu el valor observat, el valor crític i el valor p.

3.4 Càlcul del contrast

Apliqueu els contrastos per avaluar si hi ha diferències significatives en readingScore segons cada variable, usant la funció de contrast implementada. Un cop realitzats els càlculs, resumiu els resultats en una taula, de manera que cada fila correspongui a una variable i a les columnes s'indiqui el valor observat, el valor crític, el valor p obtingut i un breu comentari sobre si la diferència és significativa.

3.5 Interpretació

En funció dels resultats obtinguts a l'apartat anterior i resumits a la taula, interpreteu en quins casos s'observen diferències significatives a readingScore.

---

4 Relació entre parlar anglès a casa i lectura a casa

En aquest apartat ens preguntem si els estudiants que parlen anglès a casa (englishAtHome) tendeixen a llegir per plaer almenys 30 minuts (read30MinsADay). Per això volem investigar si els valors d'aquestes dues variables estan relacionades. Seguiu els passos que s'indiquen a continuació.

4.1 Hipòtesis nul · la i alternativa

Escriviu les hipòtesis nul · la i alternativa.

4.2 Test

Indiqueu quin tipus de test aplicareu i la seva justificació.

4.3 Càlcul del test

Realitzeu els càlculs del test. Detalleu tots els càlculs.

Requisits: No podeu fer servir cap funció que calculi el test directament. Sí que podeu utilitzar funcions per calcular els valors de la distribució corresponent.

4.4 Interpretació del test

Interpreteu el resultat del contrast i responeu a la pregunta plantejada.

---

5 Relació entre parlar anglès a casa i lectura a casa (aproximació 2)

En aquesta secció repetim l'anàlisi anterior fent servir un test de diferència de proporcions. És a dir, ens preguntem si la proporció d'estudiants que llegeixen a casa és diferent entre els estudiants que parlen en anglès a casa en relació amb els que no parlen en anglès a casa. Seguiu els passos que s'indiquen a continuació.

5.1 Hipòtesis nul · la i alternativa

Escriviu les hipòtesis nul · la i alternativa.

5.2 Test

Indiqueu quin tipus de test aplicareu i la seva justificació.

5.3 Càlculs previs

Calculeu les proporcions d'estudiants que llegeixen a casa per a cada mostra.

5.4 Desenvolupament del contrast

Implementeu el codi que calculi aquest contrast.

Nota: No podeu utilitzar funcions ja implementades a R que tornin el resultat d'aquest contrast. Sí podeu fer servir *qnorm*, *qt*, etcètera.

5.5 Interpretació del test

Interpreteu el resultat del contrast i responeu a la pregunta plantejada.

---

6 Conclusions

Escriviu les conclusions de l'anàlisi realitzada (màxim mitja pàgina).

Puntuació de l'activitat

- Apartat 1 (10%)
- Apartat 2 (10%)
- Apartat 3 (30%)
- Apartat 4 (20%)
- Apartat 5 (10%)
- Apartat 6 (10%)
- Qualitat de l'informe dinàmic (10%)