Elementaire statistiek

Bachelor in de informatica

Project – 2019-2020

1 Praktisch

Schrijf een verslag dat uit maximaal 10 pagina's bestaat, waarin je onderstaande vragen zo volledig mogelijk beantwoordt. Let erop dat je telkens expliciet aangeeft welke veronderstellingen je maakt, wat je nul- en alternatieve hypothese is, wat je besluit is, etc. Geef ook steeds de teststatistiek en de geobserveerde waarde van de teststatistiek. Ga hierbij ook steeds na of de voorwaarden (veronderstellingen), nodig om de gekozen techniek toe te passen, voldaan zijn. Toetsen mag je steeds uitvoeren met significantieniveau $\alpha = 0.05$.

Dit project maakt deel uit van het examen. Het project wordt individueel gemaakt en ook het verslag moet individueel gemaakt worden. Van dit rapport bezorg je een elektronische versie aan Valérie De Witte (valerie.dewitte@uantwerpen.be), ten laatste op vrijdag 19 juni 2020 om 8u.

2 Dataset Hartinfarcten

Het doel van de studie was om factoren te beschrijven geassocieerd met trends in de tijd op overleving na ziekenhuisopname voor acute hartinfarct. De dataset kan je terugvinden onder hartinfarct in de map Project onder Studiemateriaal op Blackboard en bevat de volgende variabelen:

1. **id**: identificatienummer
2. **age**: leeftijd van de patiënt bij ziekenhuisopname
3. **gender**: geslacht van de patiënt: '0' = man, '1' = vrouw
4. **hr**: initiële hartslag: slagen per minuut
5. **bmi**: BodyMassIndex: kg/m^2
6. **cvd**: geschiedenis van hart- en vaatziekten: '0' = nee, '1' = ja
7. **sho**: cardiogene shock: '0' = nee, '1' = ja
8. **mitype**: type van hartinfarct (pathologisch): '0' = geen aanwezigheid van Q-golflengtes, '1' = aanwezigheid van Q-golflengtes
9. **fdate**: datum van de laatste opvolging: dd/mm/jjjj
10. **los**: duur van het ziekenhuisverblijf, in dagen
11. **dstat**: ontslagstatus uit het ziekenhuis: '0' =levend, '1' = dood

12. **lenfol**: totale duur van de opvolging: aantal dagen vanaf ziekenhuisopname tot de datum van de laatste opvolging

13. **fstat**: status bij de laatste opvolging: '0' =levend, '1' = dood.

Opdat elke student met een andere dataset zou werken, verwijder je een aantal observaties op de volgende manier. Beschouw de 3 laatste cijfers ijk van je studentennummer. Verwijder vervolgens de rijen $k+1$, $j+1$, $i+1$, $jk+1$, $ij+1$, $ik+1$, $ijk+1$ en $i+j+k+1$ uit de dataset. In R kan je de rijen o, p en q uit een matrix A verwijderen met het commando $A = A[-c(o,p,q),]$. In je verslag noteer je welke rijen je verwijderd hebt uit de dataset, alsook je studentennummer.

Beantwoord volgende vragen:

1. Bestudeer en bespreek de verdeling van de variabele **los**. Bespreek hiertoe gepaste grafische voorstellingen. Ga ook op een formele manier na of de gegevens normaal verdeeld zijn. Indien dit niet het geval is, in welke zin wijken de gegevens af van normaal verdeelde gegevens. Kan je de gegevens transformeren naar normaal verdeelde gegevens? Bespreek.

2. Hangt de duur van het ziekenhuisverblijf af van het levend of gestorven zijn van de patiënten? M.a.w. is de duur van het ziekenhuisverblijf identiek verdeeld bij de populatie levende en gestorven patiënten? Voer een gepaste test uit.

3. Ga na of er een verband is tussen het type hartinfarct en de ontslagstatus uit het ziekenhuis na opname. Voer opnieuw een gepaste test uit.

4. Kan je uit de leeftijd van de patiënt het BMI voorspellen? Beschrijf uitvoerig.