Elementaire statistiek

Bachelor in de informatica

Project – 2020-2021

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, samen met je code, aan Valérie De Witte (valerie.dewitte@uantwerpen.be), ten laatste op vrijdag 11 juni 2021 om 8u.

2 Dataset Drugs

Het doel van de studie was om behandelingsprogramma's met verschillende geplande duur te vergelijken in functie van de tijd tot herval na de behandeling. De dataset kan je terugvinden onder drugs in de map Project onder Studiemateriaal op Blackboard en bevat de volgende variabelen:

1. **id**: identificatiecode
2. **age**: de leeftijd van de patiënt, in jaren
3. **becktota**: de depressiescore van de patiënt
4. **hercoc**: het gebruik van heroïne/cocaïne voor de inschrijving: '1' = heroïne en cocaïne, '2' = alleen heroïne, '3' = alleen cocaïne, '4' = geen van beide
5. **ivhx**: druggebruik voor de inschrijving: '1' = nooit, '2' = vroeger, '3' = recent
6. **ndrugtx**: het aantal vorige drugbehandelingen
7. **treat**: het type van de behandeling: '0' = kort (gezondheidsopvoeding en preventie van terugval), '1' = lang (zeer gestructureerde levensstijl in een gemeenschappelijke leefomgeving)
8. **site**: de plaats van de behandeling: '0' = A, '1' = B
9. **los**: de duur van de behandeling, in dagen
10. **time**: de tijd tot het herval van de patiënt na de behandeling, in dagen.

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 **age**. 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 efficiëntie af van het type van het programma? M.a.w. is het volgen van een kort programma meer effectief dan het volgen van een lang programma? Voer een gepaste test uit.

3. Ga na of er een verband is tussen het druggebruik voor de inschrijving en het type behandeling. Voer opnieuw een gepaste test uit.

4. Kan je uit de duur van de behandeling de tijd tot het drugherval voorspellen? Beschrijf uitvoerig.