Elementaire statistiek

Tweede jaar bachelor in de informatica

Project – 2022-2023

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 significatieniveau $\alpha = 0.05$.

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 maandag 19 juni 2023 om 9u.

2 Dataset Eik

Deze dataset bevat gegevens over eikensoorten die voorkomen in de USA en kan je terugvinden in de file eik.csv op Blackboard. De dataset bevat de volgende variabelen:

1. Boom: het volgnummer van de beschouwde boomsoort
2. Regio: de regio waarin de boom voorkomt ('Atlantic' of 'California')
3. Grootte: de grootte van het gebied waarin de soort voorkomt (in 100 km^2)
4. Volume: het volume van de eikel (in cm^3)
5. Hoogte: de hoogte van de boom (in m).

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 i, j en k uit een matrix A verwijderen met het commando $A = A[-c(i, j, k),]$. 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 variabelen Volume en Grootte. 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? Bespreek.

2. Ga na of er een verband is tussen dikke eikels, dit zijn eiken waarvan het volume van de eikel minstens 3 cm^3 is, en het gebied waarin de boom voorkomt. Maak hiervoor een nieuwe variabele 'dikke eikel' aan. Voer dan een gepaste test uit.

3. Kan je uit log(Volume) de Hoogte voorspellen? Beantwoord deze vraag grondig en zo volledig mogelijk.