**Opdracht: Eindopdracht Statistics for Data Science Deel 1**

**Samenvatting opdracht 1:**

Opdracht\_1\_obv\_universiteitsdata\_final.ipynb

1. zie code

lees de dataset in een pandas DataFrame

df = pd.read\_csv(DATA\_PATH, sep=",")

df.columns = [c.replace(".", "\_") for c in df.columns]

1. **a)** Om universiteiten objectief in te delen, wordt een kolom 'Size' toegevoegd, gebaseerd op 'Enroll' (het aantal nieuwe ingeschreven studenten), wat de capaciteit en aantrekkingskracht van de instelling weergeeft. De mediaan van 'Enroll' wordt gebruikt als drempel, omdat deze minder gevoelig is voor uitschieters en zo een eerlijkere scheiding tussen 'Groot' en 'Klein' garandeert.

A diagram of a graph

AI-generated content may be incorrect.

**b)** Er is een positief verband tussen de selectiviteit van een universiteit en de mediane totale kosten. Hoe selectiever de universiteit, hoe hoger de mediaan van de totale kosten.

A graph of a bar graph

AI-generated content may be incorrect.

**c)** Is er een relatie tussen de Student-Faculty Ratio (S\_F\_Ratio) en de Graduation Rate (Grad\_Rate) van de universiteiten, en verschilt deze relatie tussen publieke en private universiteiten?

A diagram of a graph

AI-generated content may be incorrect.

Privé-universiteiten hebben significant hogere afstudeerpercentages en lagere student-faculteitsratios dan publieke universiteiten.

De student-faculteitsratio heeft een zwak negatief effect op de afstudeercijfers, maar de privé/publieke status van een universiteit is een sterkere voorspeller

van het afstudeerpercentage, waarschijnlijk door andere factoren zoals selectiviteit en financiële middelen.

**(3) zie code**

|  |  |
| --- | --- |
| A graph of a graph of a diagram with Ryugyong Hotel in the background  AI-generated content may be incorrect. | A graph with a red line  AI-generated content may be incorrect. |
| AIC Normal: 13086.119473449222  AIC Log-Normal: 13048.702583327864  De AIC geeft aan dat de normale verdeling beter past. (AIC normal: 13086.119473449222) |  |

**Conclusie:**

De verdeling van de Room and Board Costs vertoont een duidelijke scheefheid naar rechts, waarbij de meeste waarden zich concentreren rond de €3.500–€5.000 en enkele uitschieters tot boven de €7.000 zichtbaar zijn. Om te bepalen welke theoretische verdeling het beste aansluit, is zowel een normale als een lognormale verdeling getest en vergeleken met de Akaike Information Criterion (AIC). De resultaten laten zien dat de lognormale verdeling een lagere AIC-waarde heeft (13048.70) dan de normale verdeling (13086.12). Dit betekent dat de lognormale verdeling een betere fit geeft. Daarmee wordt bevestigd dat de Room and Board Costs niet symmetrisch verdeeld zijn, maar beter worden beschreven door een lognormale verdeling die de scheefheid en langere staart naar rechts kan modelleren.

Deze bevinding wordt ondersteund door de Q-Q plot, waarin de data nauw aansluit bij de theoretische lognormale lijn, met slechts kleine afwijkingen in de staarten. Daarmee kan geconcludeerd worden dat de Room and Board Costs het best worden beschreven door een lognormale verdeling, die de scheefheid en lange rechterstaart adequaat kan modelleren.

**(4)**

**(a)** Ontvangen elite scholen een ander aantal aanmeldingen in vergelijking met niet- elite scholen? Definieer “elite-school” als scholen waarvoor geldt dat meer dan 50% van de studenten tot de top 10% van hun high school behoort.

Definitie elite school: Een school waarbij meer dan 50% van de studenten tot de top 10% van hun high school behoort.

**1.⁠ ⁠Hypothesen**

Nulhypothese (H₀): Er is geen verschil in het gemiddeld aantal aanmeldingen tussen elite- en niet-elite scholen.

Alternatieve hypothese (H₁): Er is wél een verschil in het gemiddeld aantal aanmeldingen tussen elite- en niet-elite scholen.

**2.⁠ ⁠Keuze van de toets**

Aangezien we de gemiddelden van twee onafhankelijke groepen willen vergelijken (elite versus niet-elite scholen), is een t-toets voor twee onafhankelijke steekproeven geschikt.

**3. Resultaten van de toets**

Levene's Test: statistic=25.731, p=0.000

Welch's T-test (ongelijke varianties aangenomen):

T-statistic: 5.662

P-value: 0.000

Vrijheidsgraden (Welch's): 85.90

Omdat Levene’s test voor gelijke varianties significant is (statistic = 25.731, p = 0.000), kunnen we de aanname van gelijke varianties verwerpen. Daarom is gekozen voor de Welch’s t-toets, die robuust is bij ongelijke varianties.

**4.⁠ ⁠Conclusie**

Omdat de p-waarde kleiner is dan het gekozen significantieniveau (α = 0.05), verwerpen we de nulhypothese. Er is dus een significant verschil in het aantal aanmeldingen tussen elite- en niet-elite scholen. Vrijheidsgraden staat voor het aantal vrijheidsgraden (degrees of freedom). In het geval van een Welch-t-toets worden de vrijheidsgraden aangepast om rekening te houden met ongelijke varianties tussen de twee groepen.In dit geval is het aantal vrijheidsgraden afgerond 86, wat aangeeft dat de toets voldoende informatie heeft om het verschil in gemiddelden met nauwkeurigheid te schatten

(b) Is er een verband tussen acceptance rate en graduation rate?

|  |
| --- |
| A graph with blue dots  AI-generated content may be incorrect. |
| Pearson's correlatiecoëfficiënt: -0.287  P-value: 0.000 |

De Pearson-correlatieanalyse toont aan dat er een significant negatief verband bestaat tussen acceptance rate en graduation rate (r = -0.287, p < 0.001).

Dit betekent dat scholen met een hogere acceptance rate gemiddeld een lagere graduation rate hebben. Hoewel het verband statistisch significant is, is de sterkte van de relatie beperkt (zwak tot matig).

* De correlatie is negatief: hoe hoger de acceptance rate, hoe lager de graduation rate.
* De waarde van -0.287 wijst op een zwak tot matig negatief lineair verband.
* Omdat de p-waarde significant is, verwerpen we de nulhypothese (H₀: geen correlatie).

**(c)**

Is er een significant verschil in het gemiddelde Outstate collegegeld tussen scholen met een hoge en lage Grad\_Rate? (Definieer "hoog" als een afstudeerpercentage boven het gemiddelde)

Reslutaat:

* Levene's Test: statistic=28.707, p=0.000
* Welch's T-test (ongelijke varianties aangenomen):
* T-statistic: 17.315
* P-value: 0.000

**Conclusie:**

Uit de analyse blijkt dat er een significant verschil is in het gemiddelde collegegeld (Outstate) tussen scholen met een hoog afstudeerpercentage en scholen met een laag afstudeerpercentage. Levene’s test toont aan dat de varianties tussen de groepen significant verschillen (p < 0.001), waardoor de aanname van gelijke varianties is geschonden en terecht gebruik is gemaakt van de Welch’s t-test. De resultaten van de Welch’s t-test bevestigen dat het verschil in gemiddelde collegegelden statistisch significant is (Welch’s t = 17.315, p < 0.001). Op basis van de positieve t-waarde kan worden geconcludeerd dat scholen met een hoger afstudeerpercentage gemiddeld ook een hoger collegegeld hanteren.

We verwerpen de nulhypothese. Er is een significant verschil in het gemiddelde collegegeld tussen scholen met een hoog en laag afstudeerpercentage