Elementaire Statistiek Project

E. Nabil, 20215256

2023-06-19

1 Praktische Opdracht

Schrijf een verslag van maximaal 10 pagina's waarin je de volgende vragen zo volledig mogelijk beantwoordt. Geef duidelijk aan welke veronderstellingen worden gemaakt, wat de nul- en alternatieve hypotheses zijn, de conclusie, enzovoort. Geef ook de teststatistiek en de waargenomen waarde van de teststatistiek. Controleer of de voorwaarden (veronderstellingen) die nodig zijn om de gekozen techniek toe te passen, zijn voldaan. Voer de toetsen uit met een significantieniveau van $\alpha = 0,05$. Voor dit project zullen we gebruikmaken van een dataset. De dataset bevat gegevens over eikensoorten die voorkomen in de USA en is te vinden in het bestand "eik.csv". 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 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)

Om ervoor te zorgen dat ik met een unieke dataset werk, moet ik enkele observaties verwijderen op basis van mijn studentnummer. Mijn studentnummer is **20215256**, dus met de laatste 3 cijfers heb ik $\mathbf{i = 2}$, $\mathbf{j = 5}$ en $\mathbf{k = 6}$. Ik zal de volgende rijen selecteren: $k + 1$, $j + 1$, $i + 1$, $jk + 1$, $ij + 1$, $ik + 1$, $ijk + 1$ en $i + j + k + 1$ uit de dataset. Dit betekent dat ik de volgende rijen zal verwijderen: 7, 6, 3, 31, 11, 13, 61 en 14.

2 Vragenlijst

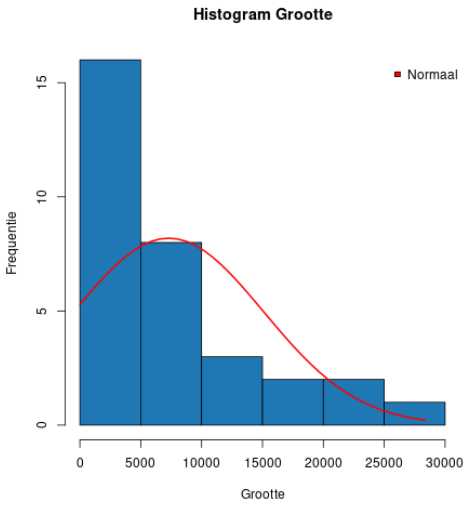
2.1 Vraag 1

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 ?

Om een duidelijk beeld te krijgen van hoe de variabelen verdeeld zijn, heb ik besloten om histogrammen op te stellen voor beide variabelen. Een histogram is een visuele weergave waarbij de gegevens worden verdeeld in verschillende "balken" om de frequentie van waarden in elke balk weer te geven. Hierdoor kunnen we gemakkelijk de verdeling van de waarden observeren en beoordelen of deze lijkt op een normale verdeling.

Om de histogrammen nog informatiever te maken, heb ik ervoor gezorgd dat er ook bijpassende normaalcurven worden geplot. Een normaalcurve, ook wel bekend als een Gaussische curve of een bell curve, is een wiskundige grafiek die de kenmerken van een normale verdeling weergeeft. Door de normaalcurven naast de histogrammen te tonen, kunnen we de verdeling van de variabelen beter beoordelen.

Als de waarden in het histogram mooi samenvallen met de vorm van de normaalcurve en een kenmerkende belvorm hebben, wijst dit op een redelijk normale verdeling van de variabelen. Als daarentegen de waarden in het histogram niet goed passen bij de vorm van de normaalcurve, kan dit erop duiden dat de variabelen afwijken van een normale verdeling. Onderaan ziet u de histogrammen voor de variabelen Volume en Grootte.



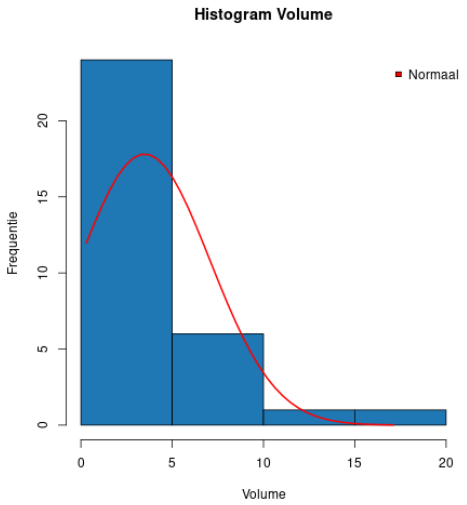


Figure 1: Grootte met normale curve. Figure 2: Volume met normale curve.

Conclusie: Bij het analyseren van figuur 1 en 2 is me opgevallen dat beide histogrammen leunen aan de linkerkant. Dit patroon wordt ook weerspiegeld in de vorm van de bijbehorende normale kromme. De piek van de normale kromme verschijnt aan de linkerkant van het histogram en niet in het midden, wat aangeeft dat de gegevens niet normaal verdeeld zijn. Deze observatie suggereert dat de gegevens mogelijk een asymmetrische verdeling hebben, waarbij de waarden zich concentreren aan de linkerkant en uitwaaieren naar rechts. Het kan wijzen op een scheve of skewed verdeling waarin de gemiddelde waarde verschoven is ten opzichte van het midden van de verdeling.

Ik heb ook een Boxplot opgesteld voor beide variabelen. Dit is een grafiek die een snelle samenvatting geeft van een of meer numerieke variabelen. De boxplot toont de verdeling van de gegevens op basis van een vijf getallen samenvatting: minimum, eerste kwartiel, mediaan, derde kwartiel en maximum. Onderaan ziet u de boxplots voor de variabelen Volume en Grootte.

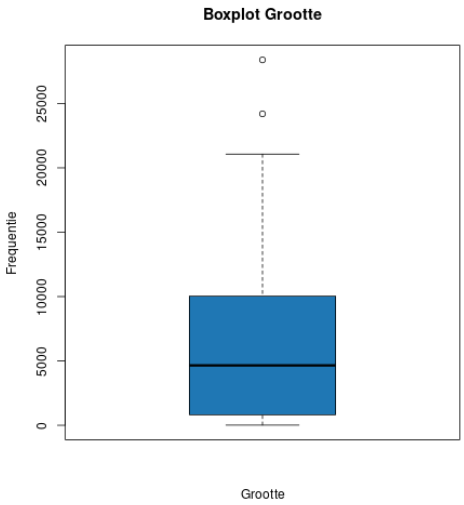


Figure 3: Boxplot voor Grootte.

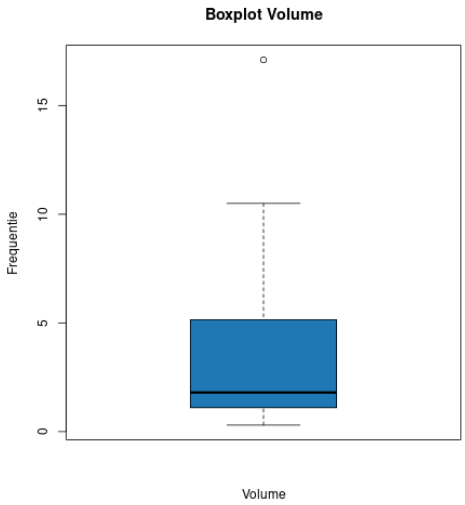


Figure 4: Boxplot voor Volume.

Als het normaal verdeeld is, zal de mediaan (tweede kwartiel) in het midden van de boxplot liggen. Ook zal de "whishker" aan beide kanten van de boxplot ongeveer even lang zijn. Hieronder een voorbeeld van een boxplot met een normaal verdeelde variabele.

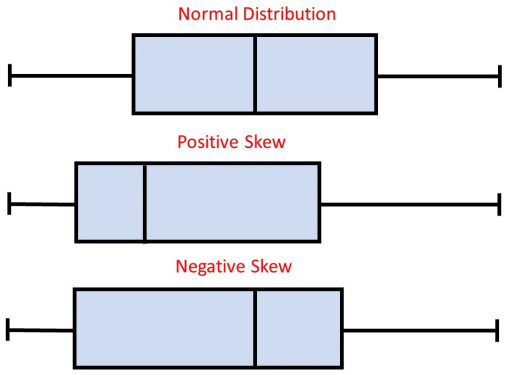


Figure 5: Boxplot referentie.

Hier zien we dus dat het niet het geval is voor beide variabelen.

Om dubbel te controleren of de gegevens normaal verdeeld zijn, heb ik ook een Q-Q plot opgesteld voor beide variabelen. Een Q-Q plot is een grafiek die de kwantielen van een gegevensverzameling vergelijkt met de kwantielen van een referentieverdeling. De kwantielen van een verdeling zijn de waarden

die de verdeling in gelijke delen verdelen.

Als de gegevens normaal verdeeld zijn, zullen de punten op de Q-Q plot ongeveer op een rechte lijn liggen. Als de punten niet op een rechte lijn liggen, kan dit erop duiden dat de gegevens niet normaal verdeeld zijn. Onderaan ziet u de Q-Q plots voor de variabelen Volume en Grootte.

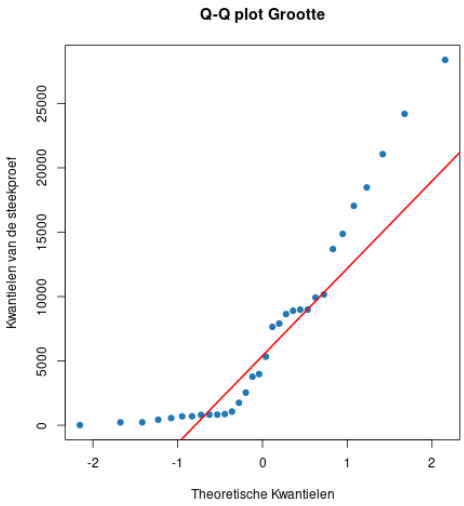


Figure 6: Q-Q plot van Grootte.

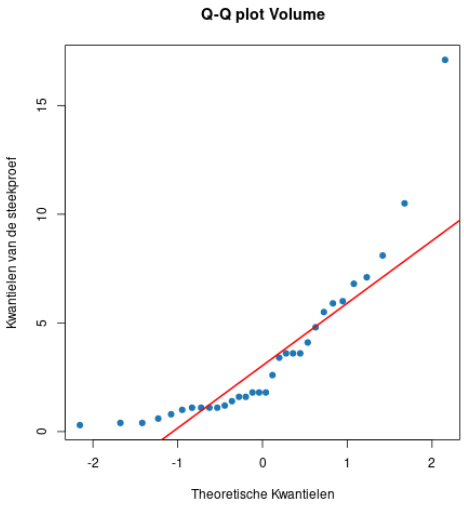


Figure 7: Q-Q plot van Volume.

Conclusie: Bij het analyseren van figuur 6 em 7 is mij opgevallen dat de punten op de Q-Q plot niet op een rechte lijn liggen. Dit suggereert dat de gegevens niet normaal verdeeld zijn. Deze observatie bevestigt mijn eerdere observatie dat de gegevens mogelijk een asymmetrische verdeling hebben, waarbij de waarden zich concentreren aan de linkerkant en uitwaaieren naar rechts.

Om mijn observaties te bevestigen, heb ik ook de skewness en kurtosis van beide variabelen berekend. Skewness is een statistische maatstaf die de symmetrie van de verdeling van een variabele meet. Kurtosis is een statistische maatstaf die de vorm van de verdeling van een variabele meet.

Als de skewness van een verdeling gelijk is aan 0, is de verdeling symmetrisch. Als de skewness van een verdeling groter is dan 0, is de verdeling scheef naar

rechts. Als de skewness van een verdeling kleiner is dan 0, is de verdeling scheef naar links.

Als de kurtosis van een verdeling gelijk is aan 0, is de verdeling normaal. Als de kurtosis van een verdeling groter is dan 0, is de verdeling zwaarstaartig. Als de kurtosis van een verdeling kleiner is dan 0, is de verdeling lichtstaartig.

Onderaan ziet u de skewness en kurtosis van beide variabelen.

	Grootte	Volume
Skewness	1.05	1.92
Kurtosis	0.14	4.24

Conclusie: Op het tabel zie je dat de skewness van beide variabelen groter is dan 0. Dit suggereert dat de verdeling van beide variabelen scheef naar rechts is. Dit komt overeen met figuur 1 en 2

Bij het tabel zie je ook dat de kurtosis van "Grootte" bijna 0 is en de kurtosis van "Volume" groter is dan 0. Dit suggereert dat de verdeling van beide variabelen zwaarstaartig is. Dit komt overeen met wat we eerder hebben gezien.

Als Laatste heb ik ook de Shapiro-Wilk test uitgevoerd. De Shapiro-Wilk test is een statistische test die wordt gebruikt om te bepalen of een gegevensverzameling normaal verdeeld is.

De nulhypothese van de Shapiro-Wilk test is dat de gegevens normaal verdeeld zijn. De alternatieve hypothese van de Shapiro-Wilk test is dat de gegevens niet normaal verdeeld zijn.

Onderaan ziet u de resultaten van de Shapiro-Wilk test voor beide variabelen.

	Grootte	Volume
Shapiro Wilk	0.0003 (not normal)	1.6729e-05 (not normal)

Conclusie: Hier merk je dat de p-waarde van beide variabelen kleiner is dan 0.05. Dit suggereert dat de gegevens niet normaal verdeeld zijn. We kunnen de nullhypothese verwerpen. Alles duidt erop dat de gegevens niet

normaal verdeeld zijn. We kunnen dus nu vanuit gaan dat de gegevens **niet** normaal verdeeld zijn.

Extra: Wat maakt de gegevens nu wel normaal verdeeld? Als we de gegevens transformeren, kunnen we proberen de gegevens normaal verdeeld maken.

Voor de variabele "Grootte" heb ik de vierde machtswortel genomen. Want dit zorgde voor de grootste Shapiro-Wilk p-waarde. Voor de variabele "Volume" heb ik de logaritme genomen. Ook hier zorgde dit voor de beste resultaat.

Onderaan ziet u tabel voor de skewness, kurtosis en Shapiro-wilk test van beide variabelen na de transformatie.

NA transformatie	Grootte	Volume
Skewness	-0.09	-0.05
Kurtosis	-1.24	-0.86
Shapiro Wilk	0.138 (normal)	0.714 (normal)

Conclusie: Zoals je kan zien is de skewness van beide variabelen bijna 0. Dit betekent dat de verdeling van beide variabelen ongeveer symmetrisch is. Voorheen was het voor "Grootte" 1.05 en voor "Volume" 1.92. Ten slotte voor de Shapiro-wilk test is de p-waarde van "Volume" en "Grootte" groter dan 0.05. Voor "Volume" is de logaritme een goede transformatie en is de verdeling normaal verdeeld. Voor "Grootte" is de vierde machtswortel een goede transformatie en is de verdeling normaal verdeeld.

2.2 Vraag 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.

Om na te gaan of er een verband is tussen dikke eikels en het gebied waarin de boom voorkomt, heb ik een kruistabel met de verwachte waarde berekent en een chi-kwadraat test uitgevoerd. Op basis van de kruistabel kan je een beeld

krijgen of het al dan niet verband heeft met elkaar en via de chi-kwadraat test kan je met p-waarde bepalen of er statistisch significant bewijs is voor een verband tussen dikke eikels en het gebied waarin de boom voorkomt. Een kleine p-waarde (typisch onder een vooraf bepaalde significantieniveau zoals 0,05) wijst op een significant verband. Hieronder zie je de kruistabel van de variabelen "Dikke Eikels" en "Regio".

Gekregen waarden	Atlantic	California	Totaal
Geen Dikke Eikels	14	4	**18**
Dikke Eikels	8	6	**14**
Totaal	**22**	**10**	**32**

Dan heb ik de verwachte waarden berekent.

Verwachte verwaarden	Atlantic	California	Totaal
Geen Dikke Eikels	12.375	5.625	**18**
Dikke Eikels	9.625	4.375	**14**
Totaal	**22**	**10**	**32**

Zoals je kan zien zijn er kleine verschillen we gaan dus de chi-kwadraat test uitvoeren om een besluit te nemen.

Onderaan ziet u de resultaten van de chi-kwadraat test.

Variable	Chi-kwadraat test	P Value
Eikels en Regio	0.748	0.387

Conclusie: Bij het analyseren van de resultaten van de chi-kwadraat test merk je dat de p-waarde groter is dan 0.05. Dit suggereert dat er geen verband is tussen dikke eikels en het gebied waarin de boom voorkomt en we kunnen de nullhypothese niet verwerpen.

Er is dus geen verband tussen dikke eikels en het gebied waarin de boom voorkomt.

2.3 Vraag 3

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

Hier heb ik een scatter plot gemaakt om het verband tussen log(Volume) en Hoogte te visualiseren. Onderaan ziet u de scatter plot.

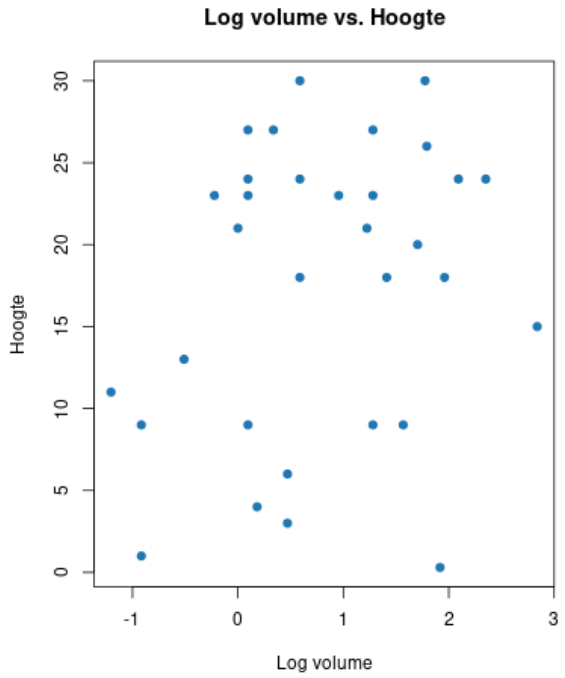


Figure 8: Scatter plot voor log(Volume) en Hoogte.

Conclusie: Hier kunt u zien dat er een **geen** verband is tussen log(Volume) en Hoogte. Omdat bij de scatter plot de punten niet in een rechte lijn liggen, maar willekeurig verspreid zijn, is er geen verband. Om dit te bevestigen heb ik een correlatiecoëfficiënt berekend.

Een positieve correlatiecoëfficiënt betekent een sterke positieve relatie, een negatieve correlatiecoëfficiënt betekent een sterke negatieve relatie, en een correlatiecoëfficiënt dichtbij nul geeft een zwakke of geen relatie weer. Onderaan ziet u de resultaten van de correlatiecoëfficiënt.

Variable	Correlation Coefficient
log(Volume) en Hoogte	0.251

Conclusie: Hier kunt u zien dat de correlatiecoëfficiënt 0.251 is. Dit betekent dat er een **zeer zwak** verband is tussen log(Volume) en Hoogte. Ik heb ook een lineaire regressie uitgevoerd om het verband tussen log(Volume) en Hoogte aan te tonen. Onderaan zult u de waarden terug vinden.

Variable	Intercept	Slope(*log_volume*)
log(Volume) en Hoogte	15.778	2.193

De *log_volume* coëfficiënt geeft de geschatte verandering in het gemiddelde Hoogte voor een toename van één eenheid in de *log_volume* voorspellende variabele. In dit geval is de geschatte coëfficiënt voor *log_volume* 2.193. Dit betekent dat voor elke toename van één eenheid in *log_volume*, het geschatte gemiddelde Hoogte met ongeveer 2.193 eenheden stijgt. In dit geval zou de geschatte vergelijking voor het lineaire regressiemodel zijn: $Hoogte = 15.778 + 2.193 * log\_volume$ De R-kwadraatwaarde is in dit geval 0.063. Dit wilt zeggen dat 6.3% van de variantie in Hoogte verklaard wordt door *log_volume*. Dat is een zeer lage waarde. Het is niet perse dat het model onbruikbaar is, maar het kan erop wijzen dat andere factoren of variabelen een grotere invloed hebben op de variabiliteit van de afhankelijke variabele.

Als laatst heb ik ook de Spearman's rank correlation coefficient berekend. Deze is minder gevoelig voor outliers. Onderaan ziet u de resultaten van de Spearman's rank correlation coefficient.

Variable	Spearman's Rank Correlation Coefficient
log(Volume) en Hoogte	0.2016

De Spearman's rangcorrelatiecoëfficiënt van 0.2016 geeft aan dat er een zwak positief monotoon verband is tussen "*log_volume*" en "Hoogte". Dit betekent dat over het algemeen, wanneer de rangordes van "*log_volume*" toenemen, de rangordes van "Hoogte" ook toenemen, zij het met een zwak verband.

Conclusie: Uit de scatter plot, correlatiecoëfficiënt, lineaire regressie en Spearman's rank correlation coefficient kunnen we besluiten dat er **zeer zwak** verband is tussen log(Volume) en Hoogte. Het is dus niet aangeraden om uit log(Volume) de Hoogte te voorspellen.