Project Statistiek

Niels Van den Broeck

s0203844

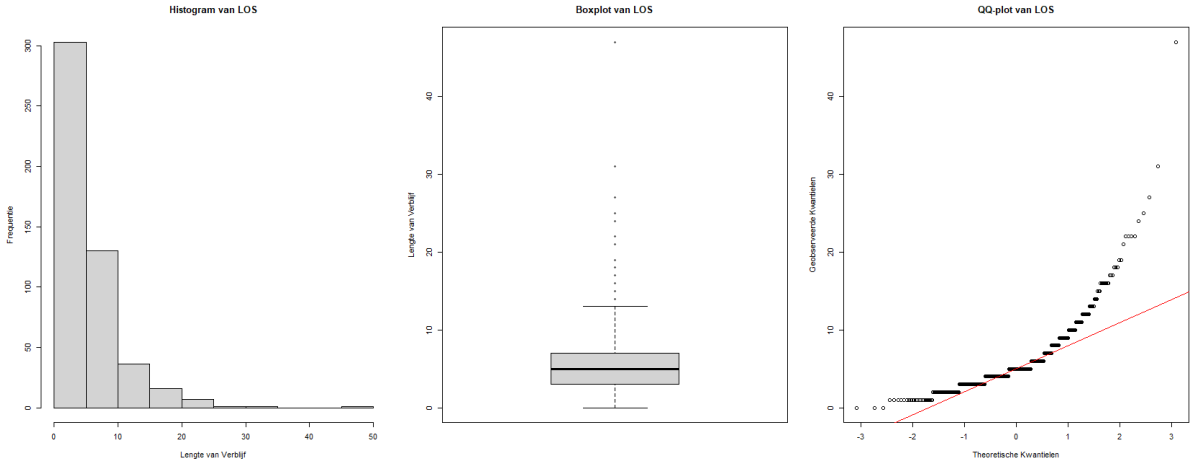
Introductie:

Rijen verwijderd met Ids: 5,9,17,33,129
In dit project wordt het significantieniveau α = 0.05 gebruikt.

Oefening 1:

Bespreking LOS

De dataset beschrijft een steekproef van grootte 495. Er zijn dus 495 observaties en dus ook 495 waarden van LOS. Het minimum is 0 en maximum 47. Als we kijken naar het steekproefgemiddelde 6.111 en steekproefmediaan 5, zien we dat het maximum een grote uitschieter is. Dit is ook duidelijk te zien op de grafieken. Als we enkele spreidingkenmerken bekijken zien we dat de empirische variantie gelijk is aan 22.334 met empirische standaardafwijking 4.726. De empirische interkwartielafstand van deze steekproef is 4.



In het histogram en boxplot zien we dat er duidelijk meer ziekenhuisopnames zijn die maar enkele dagen duren en dat ziekenhuisopnames die langer duren eerder uitzonderlijk zijn. De boxplot geeft daarboven nog een duidelijker beeld over de mediaan (dikke streep), de spreiding (lengte van de doos), de scheefheid (positie van mediaan in doos + de whiskers en de uitschieters). De kwantielplot is handig om op een snelle manier te kunnen veronderstellen of de data een normale verdeling volgt. Dit ziet er in ons geval niet zo uit.

Normaal verdeeld

Om te kunnen bevestigen of de observaties normaal verdeeld zijn, gaan we op een formele manier de Shapiro-Wilk test uit voeren. Hierbij stellen we volgende hypotheses op:

$$H_0\colon\ De\ gegevens\ komen\ uit\ een\ normale\ verdeling$$

$$H_1\colon\ De\ gegevens\ komen\ niet\ uit\ een\ normale\ verdeling$$

Dit geeft als resultaat de p-waarde: $< 2.2 * 10^{-16}$, Wat enorm hard onder het significantieniveau α = 0.05 ligt. We verwerpen de nulhypothese en we kunnen dus constateren dat de gegevens van LOS niet normaal verdeeld zijn.

$$W = 0.76578, p\text{-}value < 2.2e\text{-}16$$

Transformaties

We kunnen transformaties toepassen om te kijken in welke zin de verdeling afwijkt van de normaalverdeling. Als eerste heb ik de inverse genomen van de data, maar dit leverde hetzelfde resultaat voor de p-waarde: $< 2.2 * 10^{-16}$. Dan heb ik geprobeerd de vierkantwortel te nemen. Dit resultaat was niet veel beter: p-waarde $= 2.674 * 10^{-15}$. Vervolgens nam ik de 3de wortel wat al iets meer in de buurt kwam: p-waarde $= 1.056 * 10^{-11}$. Dit is nog steeds niet voldoende om te constateren dat deze transformatie normaal verdeeld is. Ten slotte heb ik het logaritme genomen van de data wat leidde tot een p-waarde van $8.917 * 10^{-8}$. We kunnen dus besluiten dat er geen verband is met de normaalverdeling.

Transformatie	p-waarde
original	$< 2.2 * 10^{-16}$
inverse	$< 2.2 * 10^{-16}$
sqrt	$= 2.674 * 10^{-15}$
cube	$= 1.056 * 10^{-11}$
log	$8.917 * 10^{-8}$

Oefening 2:

Als eerst kijken we naar de contingentietabel van de observaties.

	Q-golflengtes	Geen Q-golflengtes	
Dood	11 = f_{11}	28 = f_{12}	39 = f_{x1}
Levend	140 = f_{21}	316 = f_{22}	456 = f_{x2}
	151 = f_{y1}	344 = f_{y2}	495 = n

Om een verband te zoeken kunnen we de Chi-kwadraat χ^2-test uitvoeren. Hierbij wordt de volgende hypothese getest:

$$H_0\colon\ Het\ type\ hartinfarct\ is\ onafhankelijk\ van\ de\ ontslagstatus$$

$$H_1\colon\ Er\ is\ een\ verband\ tussen\ het\ type\ hartinfarct\ en\ de\ ontslagstatus$$

De test vergelijkt de geobserveerde waarden met de waarden die we zouden verwachten als er geen verband is tussen het type hartinfarct en de ontslagstatus. Die verwachte waarden berekenen we door $f_{ij} = \frac{f_{xi} f_{yi}}{n}$. Deze waarden zijn in R ook op te vragen via de Chi-kwadraat test door "test$expected" uit te voeren.

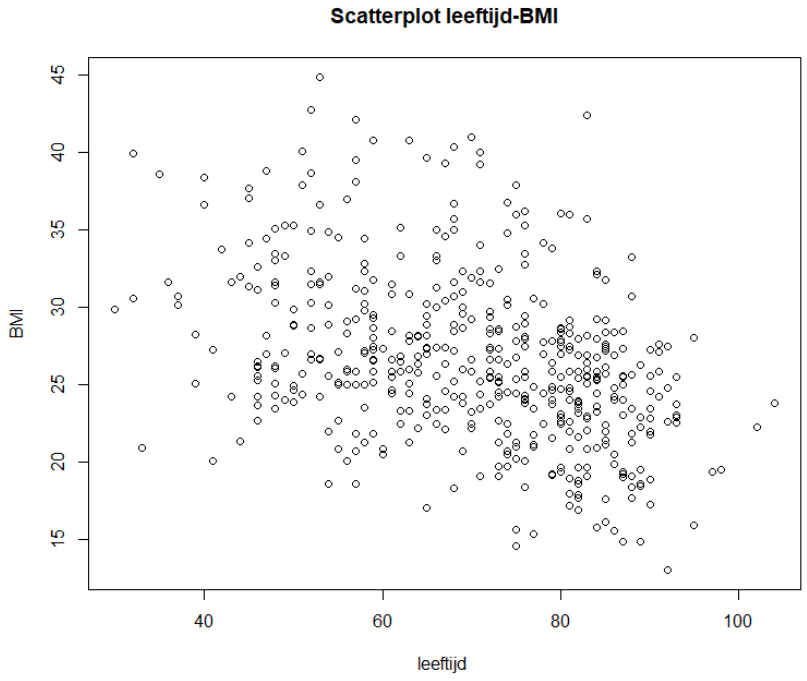
	Q-golflengtes	Geen Q-golflengtes
Dood	11.897	27.103
Levend	139.103	316.897

Bij het uitvoeren van deze test bekomen we een p-waarde van 0.886. Dit ligt hoger dan het significantieniveau 0.05 en dus zullen we de nulhypothese behouden. Het type hartinfarct heeft waarschijnlijk geen invloed op hoe patiënten uit het ziekenhuis worden ontslagen. Dit kunnen we ook zien aan het feit dat de verwachte waarden quasi overeenkomen met de geobserveerde waarden.

$$X\text{-}squared = 0.02069,\ df = 1,\ p\text{-}value = 0.8856$$

Oefening 3:

Om uit de leeftijd van een patiënt het BMI te kunnen voorspellen moeten we nagaan of er een correlatie bestaat tussen deze twee variabelen. We kunnen dit grafisch bekijken met behulp van een scatterplot. Op het eerste gezicht zien we dat de punten erg gespreid zijn en er geen stijgende of dalende trend aanwezig is, maar na een diepere inspectie zien we dat er linksonder en rechtsboven aanzienlijk minder punten bevinden. Dit kan een aanleiding zijn tot een negatief lineair verband tussen de BMI en leeftijd.



We zullen nu op een formele manier aantonen of er al dan niet een correlatie bestaat. We doen dit door gebruik te maken van de correlatiecoëfficiënt tussen de twee variabelen. Eerst zullen we nagaan of ze bivariaat normaal verdeeld zijn. Daaruit zullen we een test afleiden om de correlatie te bepalen.

Verdeling Leeftijd

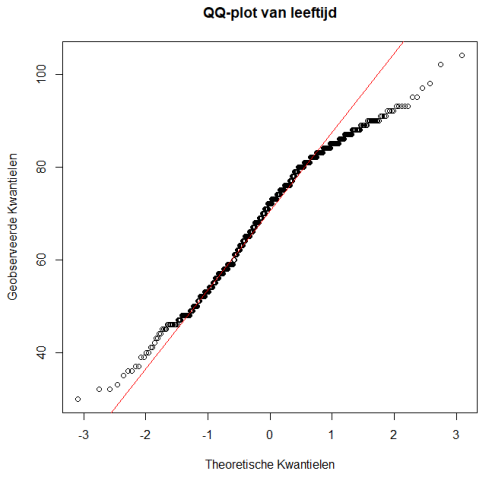
Zoals in de eerste oefening, gaan we de Shapiro-Wilk test uitvoeren om de normaliteit van de leeftijd te bepalen. We gebruiken dus gelijkaardige hypotheses:

$$H_0\text{: } De\ gegevens\ komen\ uit\ een\ normale\ verdeling$$

$$H_1\text{: } De\ gegevens\ komen\ niet\ uit\ een\ normale\ verdeling$$

We krijgen als resultaat een p-waarde $= 6.543 * 10^{-8}$ wat ver onder ons significantieniveau 0.05 ligt. Hieruit besluiten we dat de leeftijd niet normaal verdeeld is.

$$W = 0.97302,\ p\text{-}value = 6.543e\text{-}08$$



Verdeling BMI

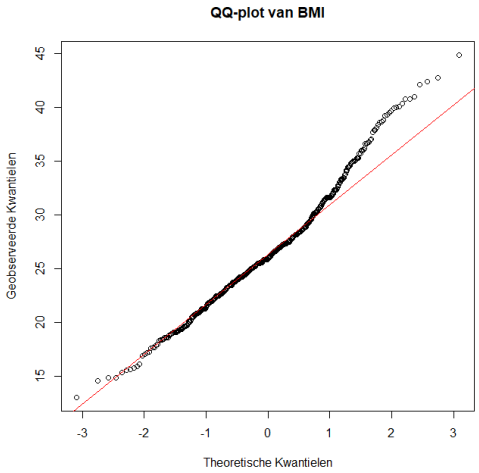
We doen hetzelfde voor het BMI:

$$H_0\text{: } De\ gegevens\ komen\ uit\ een\ normale\ verdeling$$

$$H_1\text{: } De\ gegevens\ komen\ niet\ uit\ een\ normale\ verdeling$$

De p-waarde in deze test is $1.832 * 10^{-6}$ wat opnieuw onder ons significantieniveau 0.05 ligt. Hieruit besluiten we dat het BMI niet normaal verdeeld is.

$$W = 0.9794,\ p\text{-}value = 1.832e\text{-}06$$



Uitvoeren test

Nu we weten dat de leeftijd en BMI niet uit een bivariate normale verdeling komen, gebruiken we volgende hypothese testen met behulp van de Spearman correlatiecöefficiënt.

$$H_0\text{: } Er\ is\ geen\ monotoon\ verband\ tussen\ leeftijd\ en\ BMI$$

$$H_1\text{: } Er\ is\ een\ mate\ van\ monotoon\ verband\ tussen\ leeftijd\ en\ BMI$$

Deze test geeft volgende resultaten:

Spearman's rank correlation rho

data: age and bmi
S = 28497427, p-value < 2.2e-16
sample estimates:
rho
-0.4097532

De p-waarde is veel kleiner dan ons significantieniveau 0.05, dus we verwerpen de nulhypothese. Aangezien de correlatiecöefficiënt rho negatief is, besluiten we dat er een negatief monotoon verband is tussen leeftijd en BMI. Een hoge leeftijd staat dus geassocieerd met een lage BMI en andersom.

Lineaire regressie

Nu we weten dat er een negatief monotoon verband is tussen leeftijd en BMI, gaan we kijken of dit verband ook lineair is. Dit doen we door Lineaire regressie, met volgende hypotheses:

$$H_0\colon\ b = 0$$

$$H_1\colon\ b \neq 0$$

Als we een regressie uitvoeren in R, bekomen we volgende informatie:

```
Call: lm(formula = bmi ~ age)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2405  -3.4506  -0.3728   2.8103  17.7650

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.14015    1.09598  33.888   <2e-16 ***
age         -0.15087    0.01537  -9.818   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.963 on 493 degrees of freedom
Multiple R-squared:  0.1635,       Adjusted R-squared:  0.1619
F-statistic: 96.39 on 1 and 493 DF,  p-value: < 2.2e-16
```

Hier zien we dat de p-waarde kleiner is dan 2.2e-16 en dus ook kleiner dan het significantieniveau 0.05. We verwerpen dus de nulhypothese wat besluit dat er een significant lineair verband is tussen leeftijd en BMI. De R-kwadraat waarde 0.1635 geeft aan dat er ongeveer 16.35% van de variatie in BMI wordt verklaard door de leeftijd, wat suggereert dat er nog andere factoren kunnen zijn die het BMI beïnvloeden.