Opdracht 6a
---

Dichtheidskromme, normaal-kwantiel-plot

Een nauwkeurige waarde van de lichtsnelheid is van belang voor ontwerpers van computers, omdat de elektrische signalen zich uitsluitend met de lichtsnelheid voortplanten. De eerste redelijk nauwkeurige metingen van de lichtsnelheid werden iets meer dan 100 jaar geleden verricht door A. A. Michelson en Simon Newcomb. Hieronder volgen 64 metingen verricht door Newcomb tussen juli en september 1882.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 28 | 22 | 36 | 26 | 28 | 28 | 26 | 24 | 32 | 30 | 27 |
| 24 | 33 | 21 | 36 | 32 | 31 | 25 | 24 | 25 | 28 | 36 |
| 27 | 32 | 34 | 30 | 25 | 26 | 26 | 25 | 23 | 21 | 30 |
| 33 | 29 | 27 | 29 | 28 | 22 | 26 | 27 | 16 | 31 | 29 |
| 36 | 32 | 28 | 40 | 19 | 37 | 23 | 32 | 29 | 24 | 25 |
| 27 | 24 | 16 | 29 | 20 | 28 | 27 | 39 | 23 | | |

a. Een histogram kan kenmerken van een verdeling openbaren die overduidelijk niet-normaal zijn, zoals uitschieters, uitgesproken scheefheid of hiaten en clusters. Teken een histogram voor de tijden en combineer deze met een normaalcurve. Verwacht je dat de verdeling normaal is? Wat is de betekenis van de totale oppervlakte onder de kromme? Aan welke waarde is die totale oppervlakte dus gelijk?

Voor iedere waarde van een steekproef kan op twee manieren de plaats in de standaardnormaalverdeling bepaald worden. Bij de eerste manier letten we alleen op de muu en de sigma, terwijl we de frequenties van de waarnemingen en de intervallen tussen de waarnemingen buiten beschouwing laten. Omdat voor de standaardnormaalverdeling geldt dat muu=0 en sigma=1, kunnen we de plaats van iedere x vinden door de bijhorende z te berekenen. De z-waarde zegt hoeveel standaardafwijkingen waarneming x van het gemiddelde verwijderd is, en in welke richting (positief is boven, en negatief is onder het gemiddelde). Deze z kunnen we localiseren in de standaardnormaalverdeling.

b. Bereken het gemiddelde en de standaardafwijking.

c. Bepaal voor iedere x de bijbehorende z. Er geldt:

$$z = \frac{X - muu}{sigma}$$

Wat is de z-waarde voor x=37?

d. Maak een spreidingsdiagram van de gegevens. Plaats de x-waarden op de x-as en de z-waarden op de y-as.

Bij de tweede manier letten we alleen op de frequenties van de waarnemingen en de intervallen tussen de waarnemingen, terwijl we muu en sigma buiten beschouwing laten. We bepalen voor iedere waarneming x de kans dat een waarde voorkomt die kleiner is dan of gelijk is aan x. Die kans is terug te vinden als een deel van de oppervlakte onder de kromme van de standaardnormaalverdeling. Die oppervlakte begint links, en hoe groter de kans is, hoe verder naar rechts die oppervlakte eindigt. Ter hoogte van de positie waar die oppervlakte eindigt ligt z.

e. Sorteer de tabel zodanig dat de x-waarden gerangschikt zijn in grootte, in stijgende volgorde. Voeg vervolgens een nieuwe kolom toe, en plaats daarin de rangnummers van 1 t/m 64. Welk rangnummer heeft x=37?

f. Bereken de kwantielen door alle rangnummers te delen door het totale aantal waarnemingen (64). Welk kwantiel heeft x=37?

g. Een kwantiel corresponderend met een z-waarde (of ongenormaliseerd: x-waarde) geeft de relatieve frequentie van alle waarden die kleiner dan of gelijk zijn aan z (of: x). Oftewel: de kans op een waarde die kleiner is dan of gelijk is aan z (of: x). In formule: P(Z<=z) (of: P(X<=x)). Zoek voor het kwantiel (oftewel de P(Z<=z)) van x=37 de z-waarde op in tabel A en geef die waarde.

Bij c. berekenden we voor iedere x de z die we zouden krijgen als de gegevens perfect normaal verdeeld zouden zijn, en bij d. maakten we hiervan een spreidingsdiagram. Bij g. berekenden we voor iedere x de z op basis van de werkelijke verdeling. Zouden we hiervan een spreidingsdiagram maken, dan heet dit een normaal-kwantiel-plot.

h. Teken een grafiek waarbij door de punten op basis van een perfecte normaalverdeling een groene lijn loopt, en de punten op basis van de werkelijke verdeling in rood worden weergegeven. Is de verdeling normaal?

## Opdracht 6a - S-PLUS

Plaats de gegevens in de tabel. Noem de kolom met de metingen van Newcomb 'tijd'.

a. De bovenste toolbar is de Standard Toolbar. Klik in deze toolbar op het icoontje dat precies onder Help van de menubalk zit. Als je de cursur op dit icoontje plaatst, moet als bijschrift verschijnen: Commands Window. Na op dit icoontje te hebben geklikt verschijnt een window met de naam Commands. In dit window knippert achter de prompt (weergegeven door '>') de cursor (weergegeven door een verticaal streepje). Achter die cursor kunnen commando's ingetypt worden die uitgevoerd worden nadat je op Enter (of Return) hebt gedrukt.

   We maken eerst het histogram. Binnen het commando 'hist' voor het tekenen van het histogram moeten we met het commando 'ylim' het bereik van de y-as opgeven. Deze begin altijd bij 0, en moet eindigen bij een waarde die iets hoger is dan de relatieve frequentie van de langste staaf. Omdat je vooraf niet weet hoe breed S-PLUS de staven neemt, weet je vooraf ook niet wat de relatieve frequentie van de langste staaf zal zijn. Je komt hier achter door met een waarde te beginnen, het resultaat te bekijken, en dan eventueel deze waarde te verlagen of te verhogen. In ons geval is 0.1 een geschikte waarde.

   Verder geven we binnen het commando 'hist' met 'probability=TRUE' aan dat S-PLUS een histogram op basis van de relatieve frequenties moet tekenen. Stel dat de naam van de tabel die we gebruiken 'data' heet, dan wordt het commando nu:

```
hist(data$TIJD,ylim=c(0,0.1),probability=TRUE)
```

   Voor het tekenen van een normaalcurve in het histogram moeten we x-waarden en y-waarden definieren. De x-waarden in het histogram lopen van het minimum (15) tot het maximum (40). Als stapgrootte kunnen we daarbij 0.5 kiezen. Voor de definitie van de x-waarden geven we het volgende commando:

```
x <- seq(15,40,by=0.5)
```

Op basis van deze x-waarden definieren we de y-waarden. Voor de definitie van de y-waarden geven we het volgende commando:

```
y <- dnorm(x,mean(data$TIJD),stdev(data$TIJD))
```

Op basis van de x- en y-waarden wordt de normaalcurve nu getekend door het volgende commando te geven:

```
lines(x,y)
```

b. Kies >Statistics >Data Summaries >Summaries Statistics. Selecteer onder Data en achter Variables de variabele 'tijd', en controleer of onder Summaries by Group en achter Group Variables voor (None) is gekozen. Klik nu bovenaan dit venster op >Statistics. Zorg dat alleen >Mean en >Std. Deviation geselecteerd zijn. Klik op >OK.

c. Kies >Data >Transform. Geef onder Data en achter Target Column als kolomnaam 'z'. Vul achter Expression de formule in voor de berekening van de z-waarden. In ons geval wordt die formule: (TIJD - 27.75) / 5.08. Klik op >OK. In de tabel is nu een nieuwe kolom 'z' aangemaakt met daarin de z-waarden.

d. Kies Graph >2D Plot. Kies onder Axes voor Linear, en onder Plot Type voor Scatter Plot (x, y1, y2, ...). Klik op >OK. Selecteer onder Data Columns en achter x Columns de variabele 'z', en achter y Columns de variabele 'tijd'. Klik op >OK.

e. Sorteer de tabel op basis van de variabele 'tijd'. Kies >Data >Restructure >Sort. Selecteer achter Sort By Columns de variabele 'tijd'. Klik op >OK.

   Kies >Data >Fill. Geef onder Data en achter Target Column als kolomnaam 'nummer'. Onder Fill Options en achter Context moet gekozen zijn voor Sequence, achter Start moet als startwaarde de waarde 1 gegeven zijn, en achter Increment moet als stapgrootte de waarde 1 gegeven zijn.

f. Kies >Data >Transform. Geef onder Data en achter Target Column als kolomnaam 'kwantiel'. Vul achter Expression de formule in voor de berekening van de kwantielen. In ons geval wordt die formule: nummer/64. Klik op >OK. In de tabel is nu een nieuwe kolom 'kwantiel' aangemaakt met daarin de kwantielen.

g. We zoeken in tabel A naar het element dat het dichtste bij 0.97 ligt. Dit blijkt 0.9699 te zijn. De bijbehorende z-waarde is 1.88.

h. Kies Graph 2D Plot. Kies onder Axes voor Linear, en onder Plot Type voor QQ Normal with Line (x). Klik op >OK. Selecteer onder Data Columns en achter y Columns de variabele 'tijd'. Klik op >OK.

## Opdracht 6a - SPSS

Plaats de gegevens in de tabel. Noem de kolom met de metingen van Newcomb 'tijd'.

a. Kies >Graphs >Histogram. In het window 'Histogram' verplaats je 'tijd' naar het veld bij >Variable door op > (pijl naar rechts) te klikken. Zet 'display normal curve' aan. Klik daarna op >OK. In het output-window verschijnt het histogram.

b. Kies >Statistics >Summarize >Descriptives. Plaats ’tijd’ onder Variable(s). Klik op >Options. Zorg dat >Mean en >Std. Deviation aan staan. Klik op >Continue. Klik op >OK. In het output-window vind je de resultaten.

c. Kies >Transform >Compute. Klik op >Reset. Geef als Target Variable ’z’. Vul onder Numeric Expression de volgende formule in: (tijd-27.75)/5.08. (hierbij is 27.75 het gemiddelde en 5.08 de standaarddeviatie, beide waarden had je berekend bij b.). Klik op >OK. De kolom ’z’ bevat nu de z-waarden.

d. Kies nu >Graphs >Scatter. Kies in het window ’Scatterplot’ voor ’Simple’. Klik vervolgens op ’Define’. Nu kom je in het window ’Simple Scatterplot’. Verplaats ’tijd’ naar XAxis en ’z’ naar YAxis. Klik nu op >OK. In het output-window verschijnt nu het spreidingsdiagram.

e. Kies >Window en kies vervolgens het window dat de tabel bevat. Kies >Data >Sort Cases. Plaats ’tijd’ onder Sort by. Kies onder Sort Order voor >Ascending. Klik op >OK. Vul in de derde kolom de waarden 1 t/m 64 in. Definieer de kolom door te dubbelklikken op de tekst ’var’ in de kolom. Nu verschijnt het dialoogkader ’Define Variable’. Voer bij >Variable name de naam ’nummer’ in. Kies vervolgens >Type. Kies als type ’Numeric’. Vul bij >With 2 (het aantal posities) en bij >Decimal places 0 in (aantal cijfers achter de komma). Klik vervolgens op >Continue en in ’Define Variable’ op >OK.

f. Kies >Transform >Compute. Klik op >Reset. Geef als Target Variable ’kwantiel’. Vul onder Numeric Expression de volgende formule in: nummer/64 (er zijn immers totaal 64 waarnemingen). Klik op >OK. De kolom kwantiel bevat nu de kwantielen.

g. We zoeken in tabel A naar het element dat het dichtste bij 0.97 ligt. Dit blijkt 0.9699 te zijn. De bijbehorende z-waarde is 1.88.

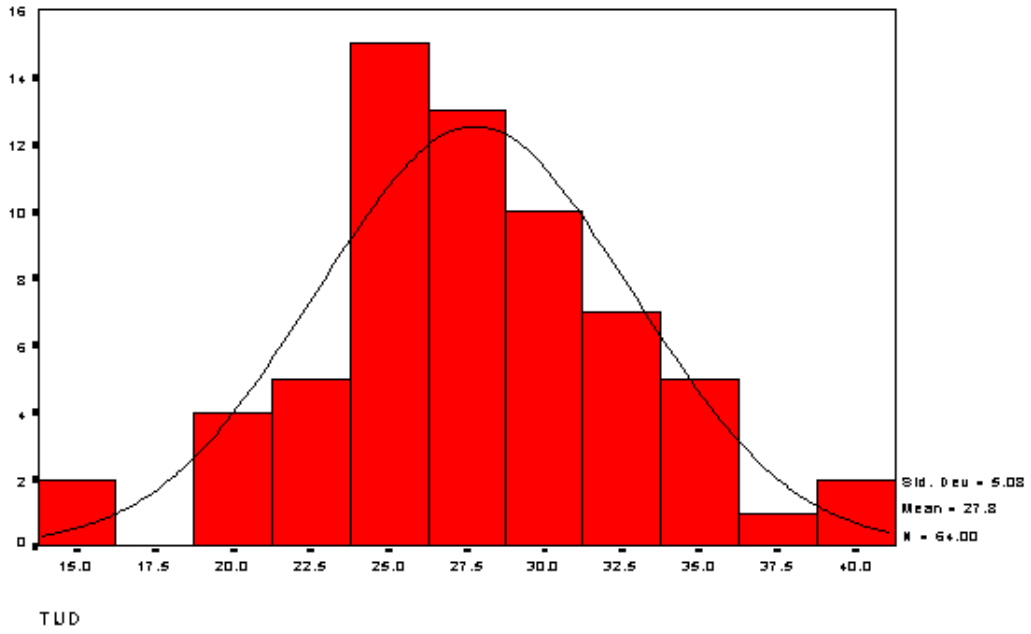
h. Kies >Graphs >Q-Q. Plaats ’tijd’ onder Variables. Klik op OK.

## Opdracht 6a - verslag

### Dichtheidskromme, normaal-kwantiel-plot

Een nauwkeurige waarde van de lichtsnelheid is van belang voor ontwerpers van computers, omdat de elektrische signalen zich uitsluitend met de lichtsnelheid voortplanten. De eerste redelijk nauwkeurige metingen van de lichtsnelheid werden iets meer dan 100 jaar geleden verricht door A. A. Michelson en Simon Newcomb. We gebruiken 64 metingen verricht door Newcomb tussen juli en september 1882.

a. Een histogram kan kenmerken van een verdeling openbaren die overduidelijk niet-normaal zijn, zoals uitschieters, uitgesproken scheefheid of hiaten en clusters. Teken een histogram voor de tijden en combineer deze met een normaalcurve. Verwacht je dat de verdeling normaal is? Wat is de betekenis van de totale oppervlakte onder de kromme? Aan welke waarde is die totale oppervlakte dus gelijk?




De verwachting is dat hier sprake is van een normale verdeling. Er zijn geen uitschieters, er is geen sprake van scheefheid, er is nauwlijks sprake van hiaten en clusters. De totale oppervlakte onder de kromme is gelijk aan de som van de frequenties. De totale oppervlakte onder de kromme is dus gelijk aan 64.

Voor iedere waarde van een steekproef kan op twee manieren de plaats in de standaardnormaalverdeling bepaald worden. Bij de eerste manier letten we alleen op de muu en de sigma, terwijl we de frequenties van de waarnemingen en de intervallen tussen de waarnemingen buiten beschouwing laten. Omdat voor de standaardnormaalverdeling geldt dat muu=0 en sigma=1, kunnen we de plaats van iedere x vinden door de bijhorende z te berekenen. De z-waarde zegt hoeveel standaardafwijkingen waarneming x van het gemiddelde verwijderd is, en in welke richting (positief is boven, en negatief is onder het gemiddelde). Deze z kunnen we localiseren in de standaardnormaalverdeling.

b. Bereken het gemiddelde en de standaardafwijking.

gemiddelde = 27.75
standaarddeviatie = 5.08

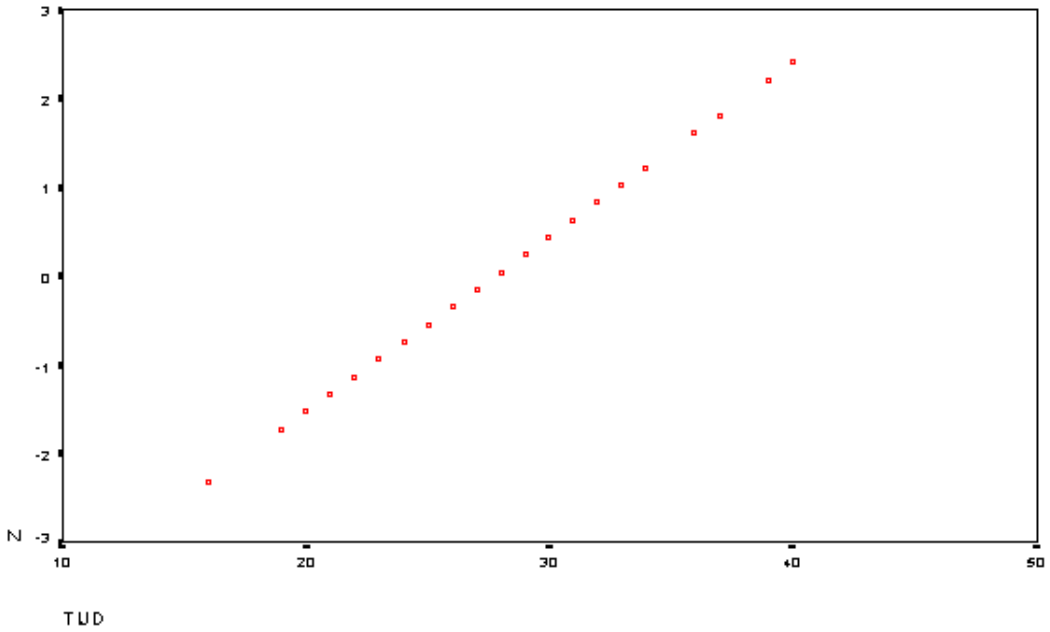
c. Bepaal voor iedere x de bijbehorende z. Er geldt:

$$z = \frac{X - muu}{sigma}$$

Wat is de z-waarde voor x=37?

z-waarde = 1.82

d. Maak een spreidingsdiagram van de gegevens. Plaats de x-waarden op de x-as en de z-waarden op de y-as.




Bij de tweede manier letten we alleen op de frequenties van de waarnemingen en de intervallen tussen de waarnemingen, terwijl we muu en sigma buiten beschouwing laten. We bepalen voor iedere waarneming x de kans dat een waarde voorkomt die kleiner is dan of gelijk is aan x. Die kans is terug te vinden als een deel van de oppervlakte onder de kromme van de standaardnormaalverdeling. Die oppervlakte begint links, en hoe groter de kans is, hoe verder naar rechts die oppervlakte eindigt. Ter hoogte van de positie waar die oppervlakte eindigt ligt z.

e. Sorteer de tabel zodanig dat de x-waarden gerangschikt zijn in grootte, in stijgende volgorde. Voeg vervolgens een nieuwe kolom toe, en plaats daarin de rangnummers van 1 t/m 64. Welk rangnummer heeft x=37?

   rangnummer = 62

f. Bereken de kwantielen door alle rangnummers te delen door het totale aantal waarnemingen (64). Welk kwantiel heeft x=37?

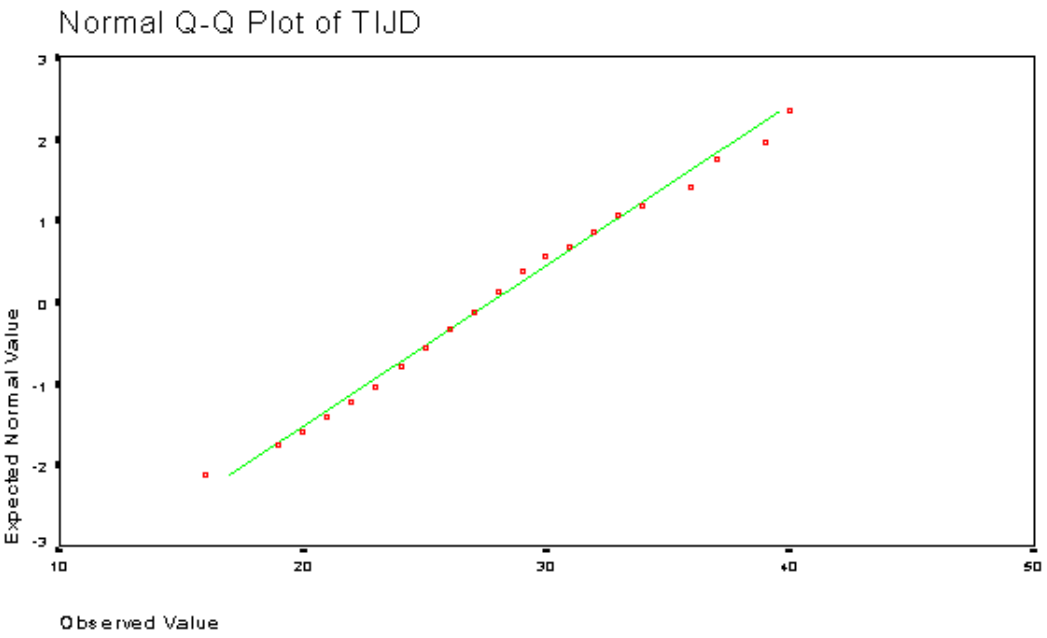
   kwantiel = 0.97

g. Een kwantiel corresponderend met een z-waarde (of ongenormaliseerd: x-waarde) geeft de relatieve frequentie van alle waarden die kleiner dan of gelijk zijn aan z (of: x). Oftewel: de kans op een waarde die kleiner is dan of gelijk is aan z (of: x). In formule: P(Z<=z) (of: P(X<=x)). Zoek voor het kwantiel (oftewel de P(Z<=z)) van x=37 de z-waarde op in tabel A en geef die waarde.

   We zoeken in tabel A naar het element dat het dichtste bij 0.97 ligt. Dit blijkt 0.9699 te zijn. De bijbehorende z-waarde is 1.88.

Bij c. berekenden we voor iedere x de z die we zouden krijgen als de gegevens perfect normaal verdeeld zouden zijn, en bij d. maakten we hiervan een spreidingsdiagram. Bij g. berekenden we voor iedere x de z op basis van de werkelijke verdeling. Zouden we hiervan een spreidingsdiagram maken, dan heet dit een normaal-kwantiel-plot.

h. Teken een grafiek waarbij door de punten op basis van een perfecte normaalverdeling een groene lijn loopt, en de punten op basis van de werkelijke verdeling in rood worden weergegeven. Is de verdeling normaal?




De verdeling is normaal. De rode punten van de normaal-kwantiel-plot liggen allemaal dicht bij de groene lijn z=x.