

# Q-Q plots, goodness-of-fit toetsen, machtstransformaties

## 2.1 Inleiding

In de kansrekening en statistiek zijn enkele standaard-kansmodellen die vaak gehanteerd worden om een kansexperiment te beschrijven. We sommen een aantal van de tot nu toe behandelde modellen nog even op (zie dictaat Statistiek & kansrekening en de appendix):

- discrete modellen:    binomiale verdeling  
                          hypergeometrische verdeling  
                          Poissonverdeling  
                          geometrische verdeling
- continue modellen:    normale verdeling  
                          exponentiële verdeling  
                          Erlang verdeling  
                          gamma verdeling  
                          Weibull verdeling  
                          lognormale verdeling  
                          uniforme verdeling  
                           $t$ -verdeling met  $n$  vrijheidsgraden.

Incidenteel zijn we ook andere verdelingen tegengekomen. Veelal zijn de data die we willen onderzoeken in dit dictaat van een continu karakter zodat we ons voornamelijk met continue modellen zullen bezighouden. Naast bovengenoemde verdelingen zijn bij Statistiek & kansrekening ook nog de chi-kwadraat verdeling en de  $F$ -verdeling geïntroduceerd. Deze verdelingen zullen we als model niet vaak hanteren, maar komen we wel tegen als kansverdeling van toetsingsgrootheden.

Hoe kunnen we nu zien welke van de bekende verdelingen als modelverdeling gehanteerd kan worden (zo er al één is)? Nadat we de samenvattingen gemaakt hebben is er al een eerste indruk ontstaan bijv. over de symmetrie. Zijn de data redelijk symmetrisch dan komt één van de symmetrische verdelingen in

## II.2

aanmerking. In eerste instantie proberen we de normale verdeling; is er sprake van een dunnere staart, probeer dan de uniforme verdeling, bij dikkere staart neem dan  $t$ -verdelingen met  $n$  (= # vrijheidsgraden) niet te groot (zie ook de tabel met de kurtosis van diverse verdelingen in hoofdstuk 1). Bij een scheve verdeling denken we eerst aan een exponentiële verdeling, vervolgens een lognormale, een Erlang, een gamma-verdeling of een Weibullverdeling.

Hoe gaat dit “proberen” nu in zijn werk. We bespreken twee methoden:

- een informele, snelle methode in de exploratieve fase
- formeel toetsen voor de confirmatieve fase.

In de exploratieve fase hebben we nog geen helder idee voor ogen welk kansmodel gehanteerd zou moeten worden. Op grond van de data willen we daar zicht op krijgen. Is dit inzicht verkregen, dan belanden we in de confirmatieve fase waar we de inmiddels verworven kennis, gewoonlijk met nieuwe data, gaan toetsen.

**Voorbeeld 2.1.1** Michelson’s waarnemingen van de lichtsnelheid (opg.6, hoofdstuk 1) wijzen door de steekproefsheffheidscoëfficiënt ( $-0.018$ ) en de steekproefkurtosis (3.26) op een normale verdeling. De EDA-samenvatting van de waarnemingen verminderd met 299000 is:

steekproefgrootte:	100				
	diepte	laag	hoog	centrum	afstand
steekproefmediaan	50.5	850	850	850	0
steekproefkwartiel	25.5	805	895	850	90
extreem	1	620	1070	845	450

De samenvatting wijst op symmetrie. Uit de EDA-samenvatting volgt dat waarnemingen  $\leq 805 - 1.5 \times 90 = 670$  of  $\geq 895 + 1.5 \times 90 = 1030$  uitschieters zijn. Uit de lijst van alle gegevens (hier niet gepresenteerd) blijkt dat er 3 uitschieters zijn: 620, 650 en 1070 (zie ook de figuur bij voorbeeld 2.2.1).  $\square$

## 2.2 Q-Q plots

Hoe gaan we nu te werk om normaliteit te onderzoeken in de exploratieve fase? We maken daartoe een zogenaamde normale Q-Q plot (Quantile-Quantile plot). We noemen de waarnemingen  $X_1, \dots, X_n$  en we ordenen deze:  $X_{(1)} \leq X_{(2)} \dots \leq X_{(n)}$ . We geven dus met  $X_{(1)}$  de kleinste waarneming aan, met  $X_{(2)}$  de op een na kleinste enz. We maken nu een plaatje, waarbij we

$X_{(i)}$  uitzetten in de  $y$ -richting tegen  $\Phi^{-1}\left(\frac{i}{n+1}\right)$  op de  $x$ -as.

Hierbij is  $\Phi$  de standaardnormale verdelingsfunctie en  $\Phi^{-1}$  zijn inverse. We noemen  $\Phi^{-1}\left(\frac{i}{n+1}\right)$  het  $\frac{i}{n+1}$ -de kwantiel van de standaardnormale verdeling, terwijl  $X_{(i)}$  het  $i$ -de steekproefkwantiel heet.

**Opmerking** Wanneer we langs een as een stochastische variabele zetten, zoals hier  $X_{(i)}$  langs de verticale as, bedoelen we dat in deze richting de waargenomen waarden van deze variabelen worden getekend. We gebruiken deze aanduiding om te benadrukken dat de getallen die langs die as staan realisaties zijn van stochastische variabelen.

Als  $X_1, \dots, X_n$  een aselechte steekproef is uit een normale verdeling dan liggen de aldus getekende punten ongeveer op een rechte lijn. De redenering die hier achter zit is in grote lijnen als volgt.

Stel dat we een aselechte steekproef  $X_1, \dots, X_{10}$  hebben uit de standaard normale verdeling. Kijk eens naar de kleinste waarneming,  $X_{(1)}$ . Dan kan bewezen worden (zie verderop) dat  $\Phi(X_{(1)})$  zich gedraagt als de kleinste waarneming uit een uniforme verdeling op  $(0, 1)$ . Merk op dat inderdaad  $\Phi(x)$  waarden tussen 0 en 1 geeft. Als we 10 waarnemingen uit een uniforme verdeling nemen op  $(0, 1)$  dan delen die het interval  $(0, 1)$  op in 11 ongeveer gelijke stukken. Dus  $\Phi(X_{(1)})$  is dan ongeveer  $1/11$ . Maar dan is dus  $X_{(1)}$  ongeveer  $\Phi^{-1}(1/11)$ . Zo is  $X_{(2)}$  ongeveer  $\Phi^{-1}(2/11)$  enz. Zetten we dus uit  $x = \Phi^{-1}(1/11)$  tegen  $y = X_{(1)}$  en  $x = \Phi^{-1}(2/11)$  tegen  $y = X_{(2)}$  enz. dan zijn  $x$  en  $y$ -coördinaat ongeveer even groot, oftewel  $y \approx x$ . In het plaatje zien we dan dus punten rond de lijn  $y = x$ .

Gaat het nu niet om standaard normaal verdeelde waarnemingen, maar om waarnemingen uit een  $N(\mu, \sigma^2)$ -verdeling, dan is  $\frac{X-\mu}{\sigma}$  standaard normaal verdeeld en dus gaat bovenstaand verhaal voor deze gestandaardiseerden op. Dus bij 10 waarnemingen is  $\frac{X_{(1)}-\mu}{\sigma}$  ongeveer  $\Phi^{-1}(1/11)$ ,  $\frac{X_{(2)}-\mu}{\sigma}$  ongeveer  $\Phi^{-1}(2/11)$  enz. Maar dat betekent dat  $X_{(1)}$  ongeveer gelijk is aan  $\mu + \sigma\Phi^{-1}(1/11)$ ,  $X_{(2)}$  ongeveer gelijk is aan  $\mu + \sigma\Phi^{-1}(2/11)$  enz. Zetten we dus uit  $x = \Phi^{-1}(1/11)$  tegen  $y = X_{(1)}$  en  $x = \Phi^{-1}(2/11)$  tegen  $y = X_{(2)}$  enz. dan zijn  $y$  en  $\mu + \sigma x$  ongeveer even groot. In het plaatje zien we dan dus punten rond de lijn  $y = \mu + \sigma x$ .

De meer precieze en algemene redenering ziet er zo uit.

1. Veronderstel dat  $X_1, \dots, X_n$  o.o. en  $N(0, 1)$ -verdeeld zijn, dan geldt dat  $\Phi(X_1), \dots, \Phi(X_n)$  o.o. en  $U(0, 1)$ -verdeeld zijn, want

$$P(\Phi(X) \leq y) = P(X \leq \Phi^{-1}(y)) = \Phi(\Phi^{-1}(y)) = y$$

voor  $y \in (0, 1)$ , m.a.w.

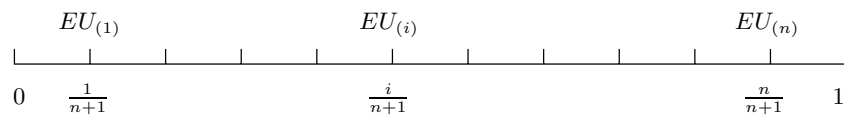
$$X \sim N(0, 1): \Phi(X) \sim U(0, 1).$$

## II.4

2. Als  $U_1, \dots, U_n$  o.o. en  $U(0, 1)$ -verdeeld zijn en we schrijven  $U_{(1)} < U_{(2)} < \dots < U_{(n)}$  voor de geordende waarnemingen, dan geldt

$$EU_{(i)} = \frac{i}{n+1} \quad i = 1, \dots, n.$$

We zullen dit resultaat niet formeel bewijzen. Een intuïtieve verklaring is als volgt. In verwachting liggen de  $n$  geordende waarnemingen op gelijke afstanden van elkaar verspreid over het interval  $(0,1)$ : deze  $n$  verwachtingen delen het interval  $(0,1)$  in  $(n+1)$  stukjes van lengte  $\frac{1}{n+1}$  :



3. Combinatie van 1 en 2 geeft: als  $X_1, \dots, X_n$  o.o. en  $N(0, 1)$ -verdeeld zijn dan is  $E\Phi(X_{(i)}) = \frac{i}{n+1}$  en dus is

$$X_{(i)} \approx \Phi^{-1}\left(\frac{i}{n+1}\right).$$

4. Veronderstel nu dat  $X_1, \dots, X_n$  o.o. en  $N(\mu, \sigma^2)$ -verdeeld zijn, dan zijn  $\frac{X_1 - \mu}{\sigma}, \dots, \frac{X_n - \mu}{\sigma}$  o.o. en  $N(0, 1)$ -verdeeld en dus geldt volgens 3:

$$\frac{X_{(i)} - \mu}{\sigma} \approx \Phi^{-1}\left(\frac{i}{n+1}\right),$$

oftewel

$$X_{(i)} \approx \mu + \sigma \Phi^{-1}\left(\frac{i}{n+1}\right).$$

Zetten we nu  $X_{(i)}$  in de  $y$ -richting uit tegen  $\Phi^{-1}\left(\frac{i}{n+1}\right)$  op de  $x$ -as, dan resulteert een plaatje met punten ongeveer op de lijn

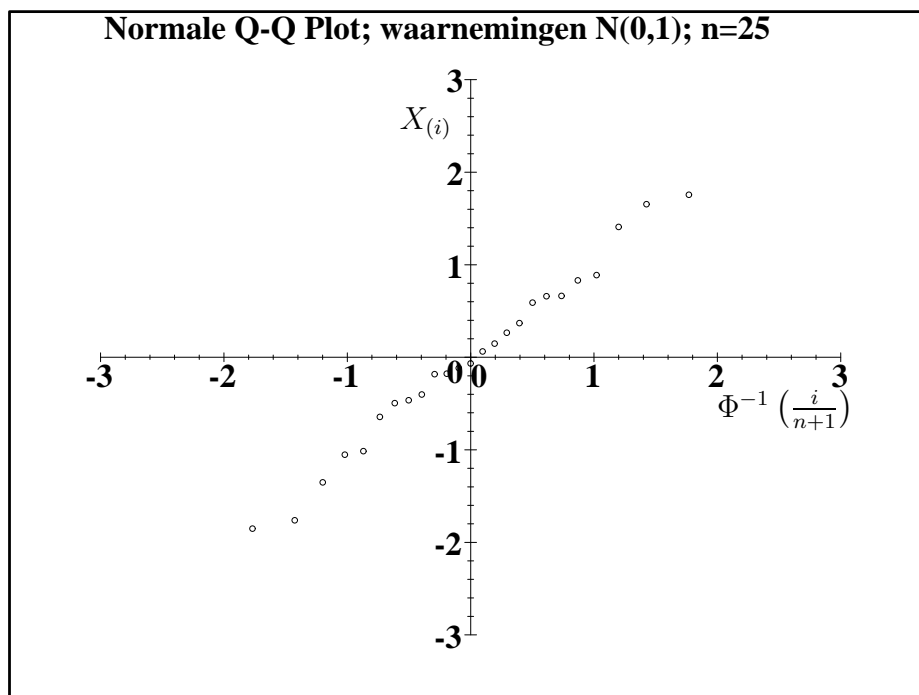
$$y = \mu + \sigma x.$$

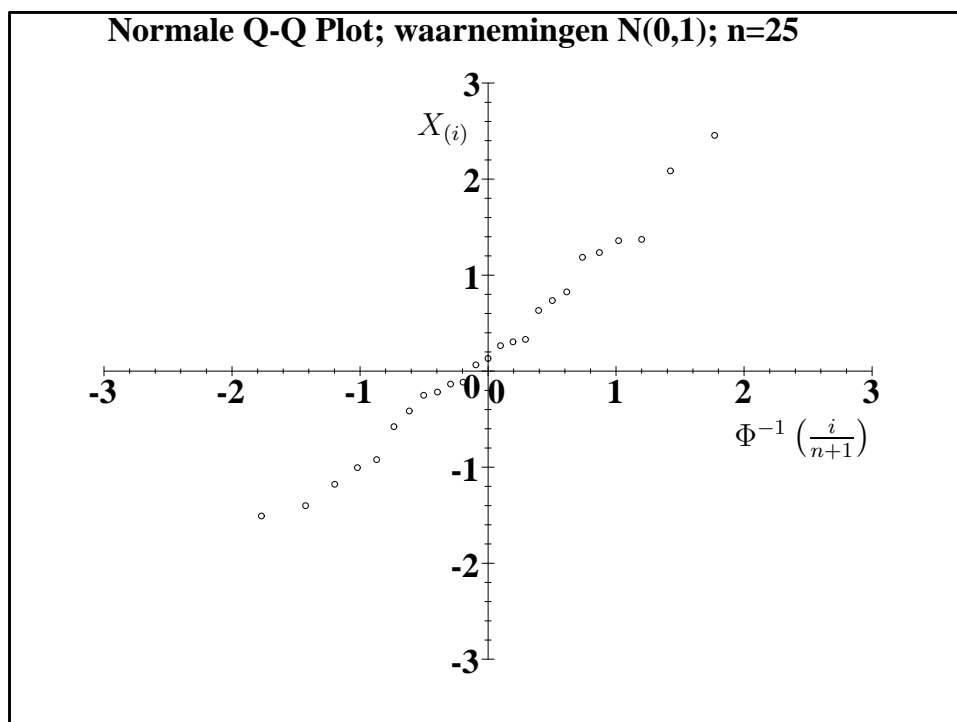
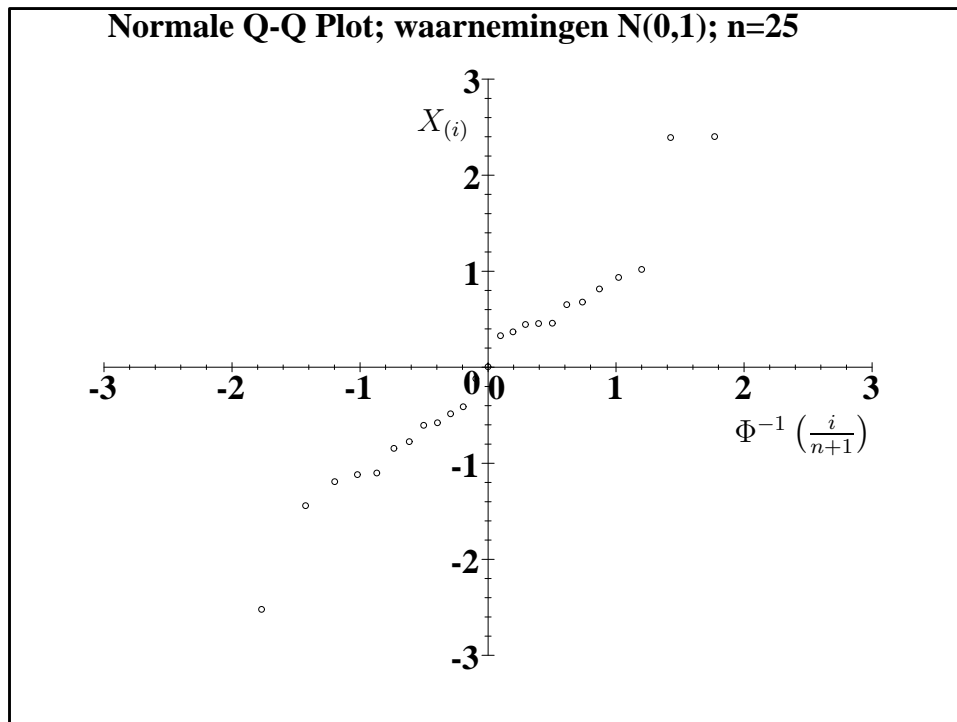
**Conclusie** Ziet het plaatje van  $X_{(i)}$  uitgezet tegen  $\Phi^{-1}\left(\frac{i}{n+1}\right)$  er uit als een rechte lijn, dan wijst dit er op dat de waarnemingen normaal verdeeld zijn. De richtingscoëfficiënt is een schatter voor  $\sigma$ , de afsnijding op de  $y$ -as is een schatter voor  $\mu$ .

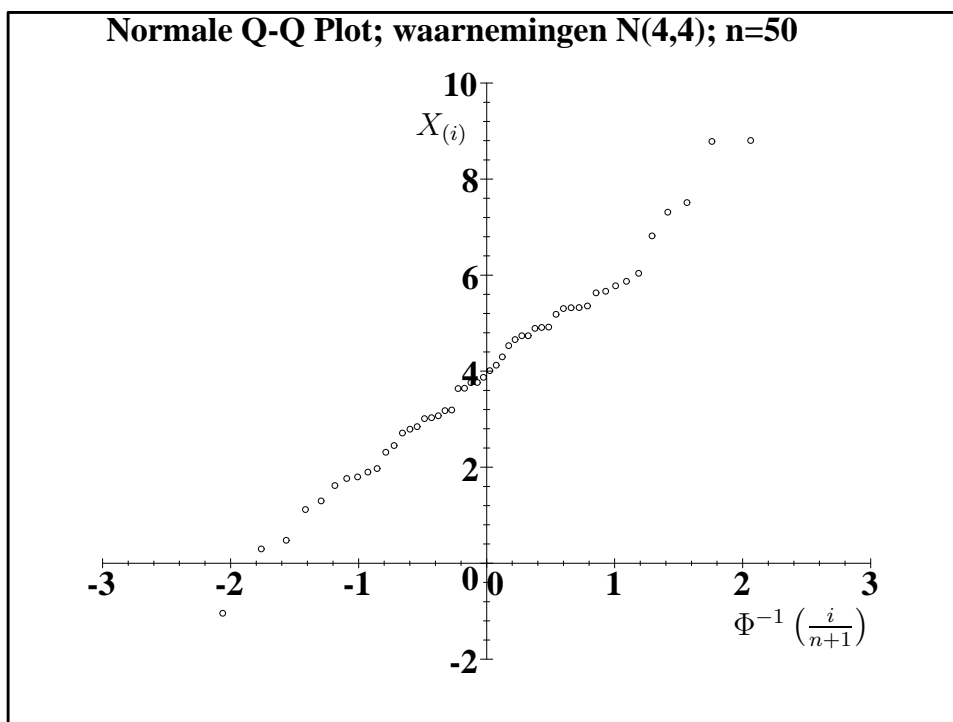
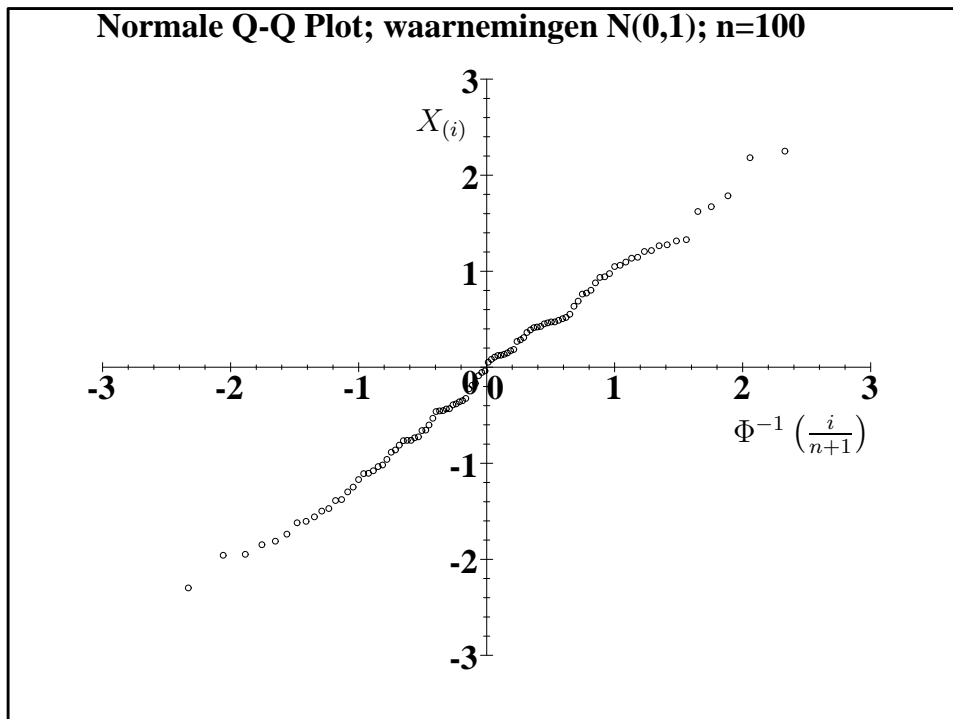
Om een indruk te krijgen wat voor plaatjes we kunnen verwachten staan in de volgende plaatjes enkele voorbeelden van normale Q-Q plots. De eerste 4 plaatjes betreffen trekkingen uit de  $N(0,1)$ -verdeling. Dit is dus wat men zich moet voorstellen bij “punten die ongeveer op de lijn  $y = x$  liggen”. Te zien is dat enige afwijkingen van de lijn  $y = x$  zich wel voor doen! Let ook op het verschil tussen  $n = 25$  en  $n = 100$ . Bij  $n = 100$  komt de lijn  $y = x$  veel sterker naar voren.

Het vijfde plaatje betreft 50 waarnemingen uit de  $N(4,4)$ -verdeling. De punten liggen rond de lijn  $y = 2x + 4$ . Let er op dat de richtingscoëfficiënt correspondeert met  $\sigma$  en niet met  $\sigma^2$ .

Bij het aflezen van de richtingscoëfficiënt moet men goed op de getallen langs de assen letten: de eenheden zijn vaak niet gelijk. Evenzo gaat het assenstelsel niet altijd door  $(0,0)$ , hetgeen van belang is bij het aflezen van de afsnijding op de  $y$ -as.

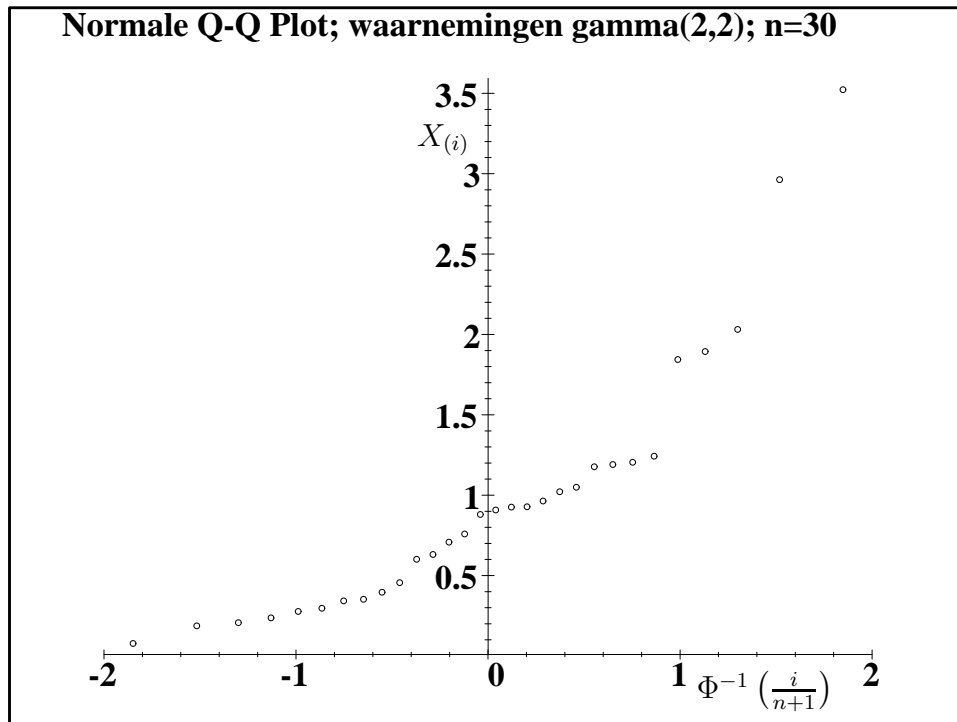






## II.8

Het volgende plaatje illustreert een (geïdealiseerde) situatie van wat er gebeurt als de waarnemingen niet normaal verdeeld zijn. Het betreft een normale Q-Q plot van 30 waarnemingen uit de gamma (2,2)-verdeling. Duidelijk is te zien dat deze punten niet op een rechte lijn liggen door de scheefheid van de gamma-verdeling.



In de praktijk is natuurlijk niet bekend uit welke verdeling de waarnemingen komen. We maken juist Q-Q plots om er achter te komen welke verdeling een geschikte keuze is! We moeten daarbij beoordelen of punten ongeveer op een rechte lijn liggen. De hier gepresenteerde plaatjes met waarnemingen uit normale verdelingen laten zien dat ook in de geïdealiseerde situatie van werkelijk normaal verdeelde waarnemingen nog behoorlijke afwijkingen van de rechte lijn mogelijk zijn. Dit maakt het beoordelen er niet eenvoudiger op, hoewel het laatste plaatje met trekkingen uit de gamma (2,2)-verdeling laat zien dat zo'n heel andere verdeling echt wel een afwijkend resultaat geeft.

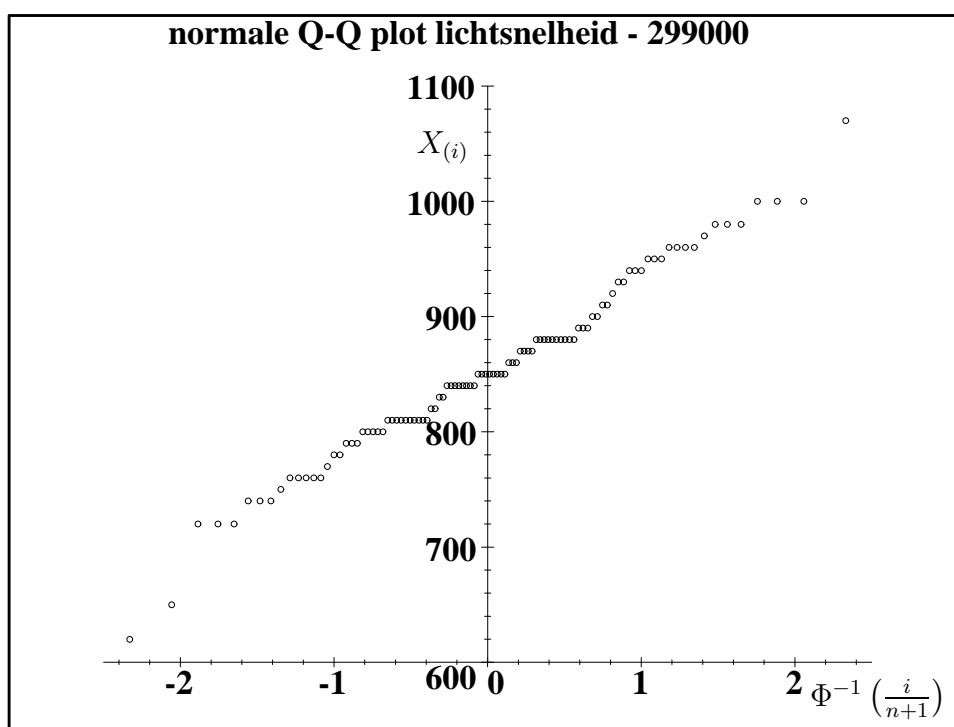
De vraag komt op of er niet een criterium is waarmee we kunnen beoordelen of de punten ongeveer op een rechte lijn liggen. Dit is de goede vraag op de verkeerde plaats! Een beoordelingscriterium (in getalsvorm) is in feite niet anders dan een **toetsingsgrootte**. De toets van Shapiro-Wilk, die in 2.4 aan bod komt, is te beschouwen als een beoordelingscriterium van de rechtlijnigheid in een normale Q-Q plot.

Toetsen behoort echter tot de **confirmatieve** fase. Hier zijn we nog pas in de **exploratieve** fase en proberen slechts een eerste indruk te krijgen. Vanuit dat

gezichtspunt past de precisie van een formele methode als een beoordelingscriterium niet. Tegelijkertijd is het de bedoeling bij de eerste indruk een **globaal** beeld te krijgen. Een beoordelingscriterium legt zich vast op één aspect. In de exploratieve fase willen we ons daar niet toe beperken: de Q-Q plot geeft een totaal beeld.

Na deze theoretische plaatjes gaan we een normale Q-Q plot bekijken, gebaseerd op “echte” data.

**Voorbeeld 2.2.1** (vervolg van voorbeeld 2.1.1) Michelson’s waarnemingen van de lichtsnelheid geven de volgende Q-Q plot.



De normale Q-Q plot van Michelson’s waarnemingen vertoont heel aardig een rechte lijn, hetgeen ons vermoeden bevestigt dat deze waarnemingen normaal verdeeld verondersteld mogen worden. De richtingscoëfficiënt van de bijbehorende rechte lijn is ongeveer 79, de afsnijding met de  $y$ -as is 850. Dit komt aardig overeen met de eerder verkregen schattingen van  $\mu$  en  $\sigma$  (zie opgave 6 van hoofdstuk 1). Met nadruk vermelden we dat we niet concluderen dat  $\sigma$  gelijk is aan 79. **De parameter  $\sigma$  blijft onbekend.** De realisatie van de richtingscoëfficiënt in de normale Q-Q plot (hier 79) geeft ons (slechts) een schatting van  $\sigma$ . Zie ook de discussie over het verschil tussen parameters, schatters en schattingen in 1.4.  $\square$

Om zelf een normale Q-Q plot te maken, moeten we  $\Phi^{-1}\left(\frac{i}{n+1}\right)$  berekenen voor  $i = 1, \dots, n$ . De waarden  $\Phi^{-1}\left(\frac{i}{n+1}\right)$  kunnen gevonden worden in de tabel van de

$N(0, 1)$ -verdeling door terug te zoeken.

**Voorbeeld 2.2.2** Als  $n = 100$  en  $i = 57$  is  $\frac{i}{n+1} = 0.5644$ . We zoeken  $z$  zodat  $\Phi(z) = 0.5644$ . In tabel 1 staan waarden van  $\Phi(z)$ . We lezen af  $\Phi(0.16) = 0.5636$  en  $\Phi(0.17) = 0.5675$ . Lineaire interpolatie geeft  $z = 0.16 + \{(0.5644 - 0.5636)/(0.5675 - 0.5636)\} \times 0.01 = 0.162$ . Dus  $\Phi^{-1}\left(\frac{i}{n+1}\right) = \Phi^{-1}\left(\frac{57}{101}\right) = \Phi^{-1}(0.5644) = 0.162$ .

Als  $n = 25$  en  $i = 2$  is  $\frac{i}{n+1} = 0.0769$ . We zoeken  $z$  zodat  $\Phi(z) = 0.0769$ . De waarde  $0.0769$  komt als uitkomst voor  $\Phi(z)$  in tabel 1 niet voor: er staan alleen getallen  $\geq 0.5$ . We maken gebruik van de symmetrie. We zoeken terug bij  $1 - 0.0769 = 0.9231$  en vinden dat  $\Phi(1.426) = 0.9231$  (interpoleren tussen  $1.42$  en  $1.43$ !). Vanwege de symmetrie geldt  $\Phi(-1.426) = 1 - 0.9231 = 0.0769$ . Derhalve is  $\Phi^{-1}\left(\frac{i}{n+1}\right) = \Phi^{-1}\left(\frac{2}{26}\right) = \Phi^{-1}(0.0769) = -1.426$ .  $\square$

Op soortgelijke manier als bij de normale verdeling kunnen we te werk gaan als we andere verdelingen willen onderzoeken. In het algemeen zetten we tegen elkaar uit

$$X_{(i)} \text{ in de } y\text{-richting en } F^{-1}\left(\frac{i}{n+1}\right) \text{ op de } x\text{-as}$$

als  $F$  de te onderzoeken verdeling is. Immers, als  $X_1, \dots, X_n$  een aselechte steekproef is met verdelingsfunctie  $F$  dan is  $F(X_1), \dots, F(X_n)$  een aselechte steekproef uit een  $U(0, 1)$ -verdeling, want

$$P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y,$$

mits  $F$  continu is. Derhalve is weer

$$EF(X_{(i)}) = \frac{i}{n+1}$$

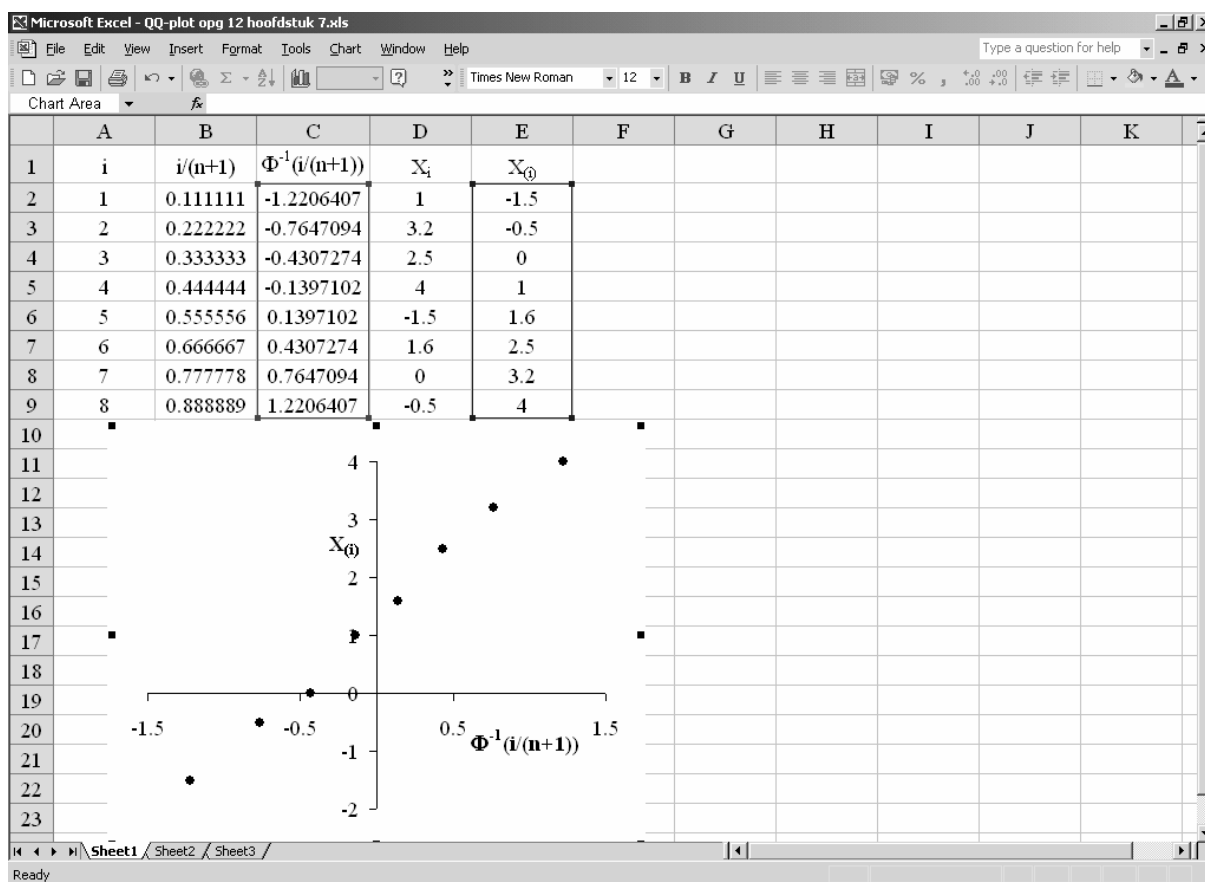
en dus  $F(X_{(i)}) \approx \frac{i}{n+1}$  en  $X_{(i)} \approx F^{-1}\left(\frac{i}{n+1}\right)$ .

Om een Q-Q plot van de te onderzoeken verdeling  $F$  te maken, moeten we dus over de inverse van  $F$  beschikken. Als  $F = \Phi$  (**standaardnormale** verdeling) gebruiken we hiervoor de tabel van de standaardnormale verdeling met terugzoeken, zie voorbeeld 2.2.2. Als  $F(x) = x$ ,  $0 \leq x \leq 1$ , (**uniforme**  $U(0, 1)$ -verdeling) dan is  $F^{-1}(y) = y$ ,  $0 \leq y \leq 1$ . Als  $F(x) = 1 - e^{-x}$ ,  $0 \leq x < \infty$  (**exponentiële**  $E(1)$ -verdeling) dan is  $F^{-1}(y) = -\ln(1 - y)$ ,  $0 \leq y < 1$ , want

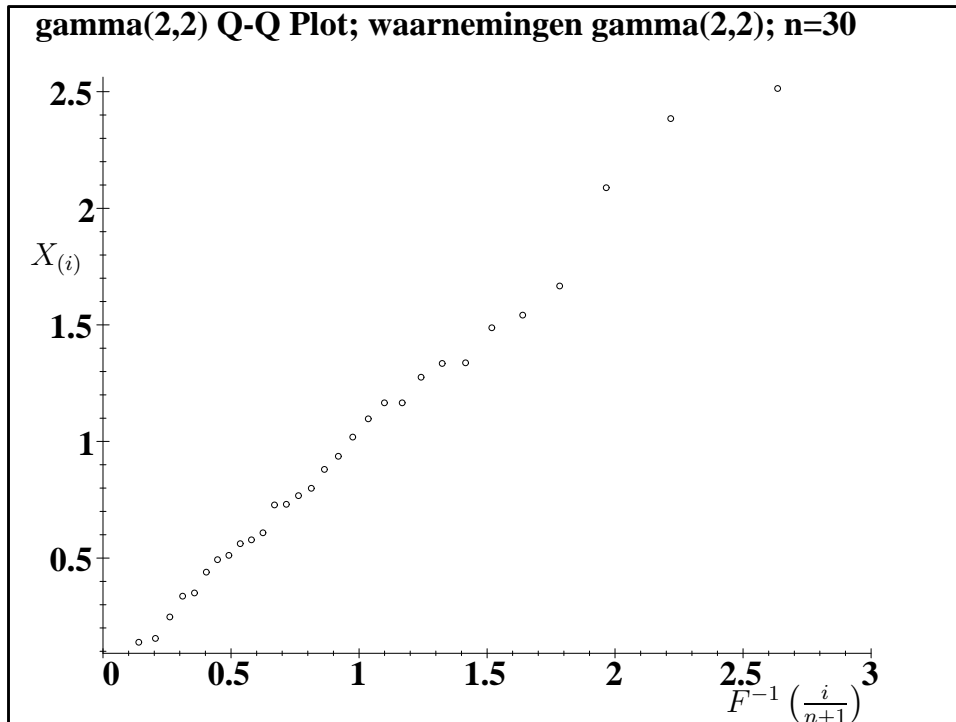
$$\begin{aligned} F(-\ln(1 - y)) &= 1 - \exp(-\{-\ln(1 - y)\}) \\ &= 1 - \exp(\ln(1 - y)) = 1 - (1 - y) = y. \end{aligned}$$

Voor de meeste andere verdelingen is het lastiger  $F^{-1}$  te bepalen. Om **lognormaliteit** te onderzoeken maken we van  $\ln(X_1), \dots, \ln(X_n)$  een normale Q-Q plot.

Q-Q plots kunnen makkelijk m.b.v. Excel gemaakt worden. Excel bevat  $F^{-1}$  met voor  $F$  de standaard normale verdeling (nodig bij de normale Q-Q plot), met voor  $F$  de gammaverdeling (kijk in de Help van Excel voor de definitie van de parameters) en met voor  $F$  de  $t$ -verdeling.



Zagen we eerder de normale Q-Q plot van de gamma (2,2)-verdeling, hier staat de gamma (2,2) Q-Q plot van de gamma (2,2)-verdeling.  $F$  is hier de verdelingsfunctie van de gamma (2,2)-verdeling. Inderdaad liggen de punten aardig rond de lijn  $y = x$ .



Als we kortweg spreken over de normale Q-Q plot, bedoelen we dat  $F = \Phi$ . We zouden eigenlijk over de standaard normale Q-Q plot moeten spreken. Evenzo bedoelen we met de uniforme Q-Q plot de Q-Q plot met  $F$  de verdelingsfunctie van de  $U(0, 1)$ -verdeling en met de exponentiële Q-Q plot de Q-Q plot met  $F$  de  $E(1)$ -verdeling.

In de normale Q-Q plot onderzoeken we in eerste instantie alleen of de waarnemingen standaardnormaal verdeeld zijn. Echter, dit levert tegelijkertijd een onderzoek naar alle andere normale verdelingen op. Immers, zijn de waarnemingen  $N(\mu, \sigma^2)$ -verdeeld, dan zien we dat in de (standaard)normale Q-Q plot de punten liggen rond de lijn  $y = \mu + \sigma x$ . Met de (standaard)normale Q-Q plot kunnen we dan in één keer **normaliteit** in het algemeen onderzoeken.

Omdat lognormaliteit van  $X$  niet anders is dan normaliteit van  $\ln(X)$ , geldt voor onderzoek naar lognormaliteit eveneens dat een (standaard)normale Q-Q plot van  $\ln(X_1), \dots, \ln(X_n)$  automatisch uitsluitel geeft over normaliteit van  $\ln(X_i)$  in het algemeen en dus lognormaliteit van  $X_i$ .

Voor andere verdelingen liggen de zaken wat ingewikkelder. Stel dat we de uniforme Q-Q plot maken, d.w.z.  $X_{(i)}$  uitzetten tegen  $\frac{i}{n+1}$  en hierbij een plaatje krijgen met punten dichtbij de lijn  $y = 3x + 2$ ; dan betekent dit dus dat  $(X_{(i)} - 2)/3 \approx \frac{i}{n+1}$  en dat we derhalve voor  $(X_i - 2)/3$  aan een  $U(0, 1)$ -verdeling kunnen denken. Laat

$U$  een stochastische grootheid zijn met een  $U(0, 1)$ -verdeling. Als  $(X_i - 2)/3$  een  $U(0, 1)$ -verdeling heeft, dan geldt

$$\begin{aligned} P(X_i \leq x) &= P((X_i - 2)/3 \leq (x - 2)/3) = \\ P(U \leq (x - 2)/3) &= (x - 2)/3 \end{aligned}$$

als  $(x - 2)/3 \in (0, 1)$ , d.w.z.  $x \in (2, 5)$ . Voor  $x \leq 2$  geldt uiteraard  $P(X_i \leq x) = 0$  en voor  $x \geq 5$  geldt  $P(X_i \leq x) = 1$ . De kansdichtheid van  $X_i$  wordt nu gegeven door

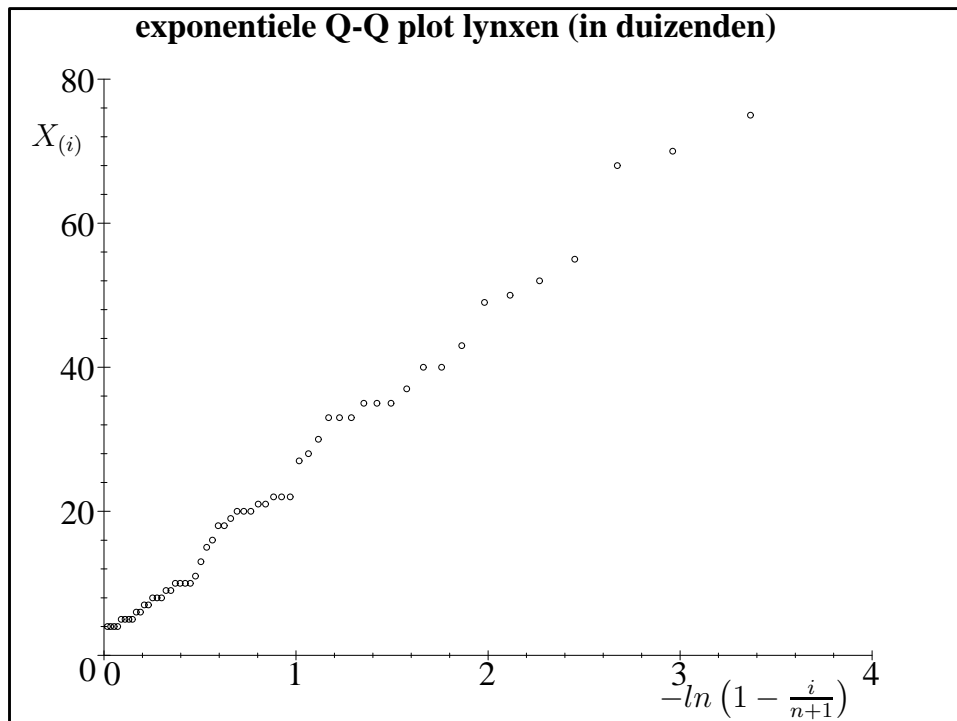
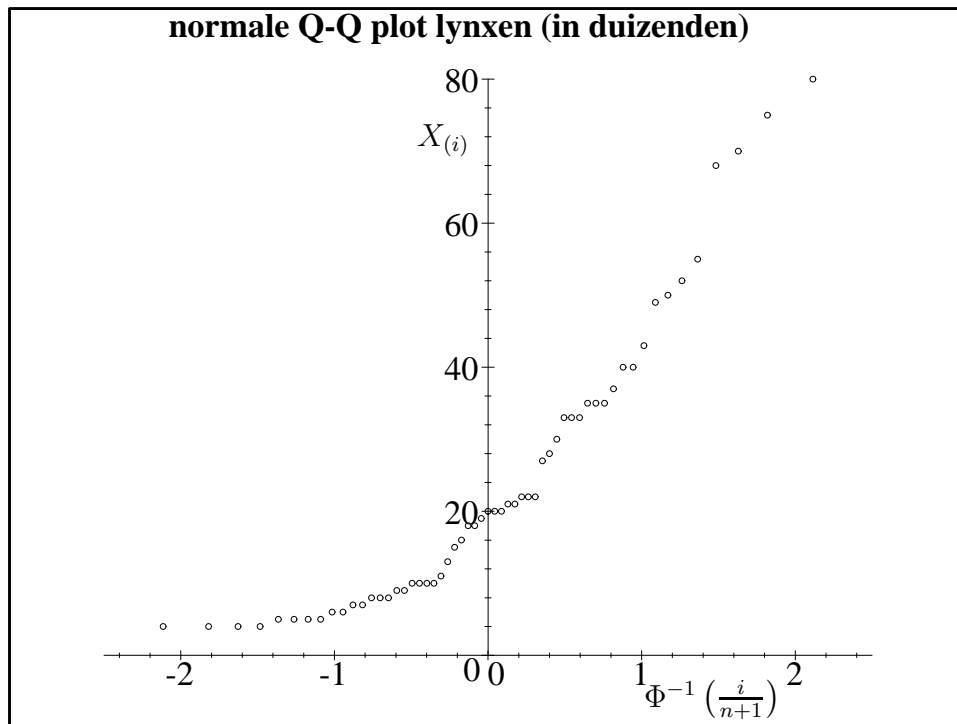
$$f_{X_i}(x) = \frac{d}{dx} P(X_i \leq x) = \begin{cases} \frac{1}{3} & 2 < x < 5 \\ 0 & \text{elders.} \end{cases}$$

De s.v.  $X_i$  is dan dus  $U(2, 5)$ -verdeeld. We kunnen nu zeggen dat de uniforme Q-Q plot er op wijst dat de waarnemingen uniform verdeeld zijn op  $(2, 5)$  oftewel dat ze dezelfde verdeling hebben als  $3U + 2$ , een over een afstand 2 verschoven en met een factor 3 herschaalde  $U(0, 1)$ -verdeling. Merk op dat  $E(3U + 2) = 3.5$  en  $\sqrt{\text{var}(3U + 2)} = 3\sqrt{1/12}$ . De afsnijding op de  $y$ -as (2) en de richtingscoëfficiënt (3) corresponderen hier dus **niet** met verwachting en standaardafwijking.

In het algemeen geldt: als punten in een Q-Q plot van een te onderzoeken verdeling  $F$ , rond de lijn  $y = ax + b$  liggen en  $U$  een s.v. met verdelingsfunctie  $F$  is, dan is de verdeling van  $aU + b$  een geschikte keuze voor de verdeling van de waarnemingen. De richtingscoëfficiënt  $a$  is dus een **herschaling**, de afsnijding op de  $y$ -as  $b$  is een **verschuiving**.

In de normale Q-Q plot is  $U \sim N(0, 1)$  en dus  $aU + b \sim N(b, a^2)$ . De verschuiving  $b$  en herschaling  $a$  in de normale Q-Q plot geven dus juist verwachting en standaardafwijking. Bij andere Q-Q plots zijn verschuiving en herschaling niet noodzakelijk gelijk aan verwachting en standaardafwijking.

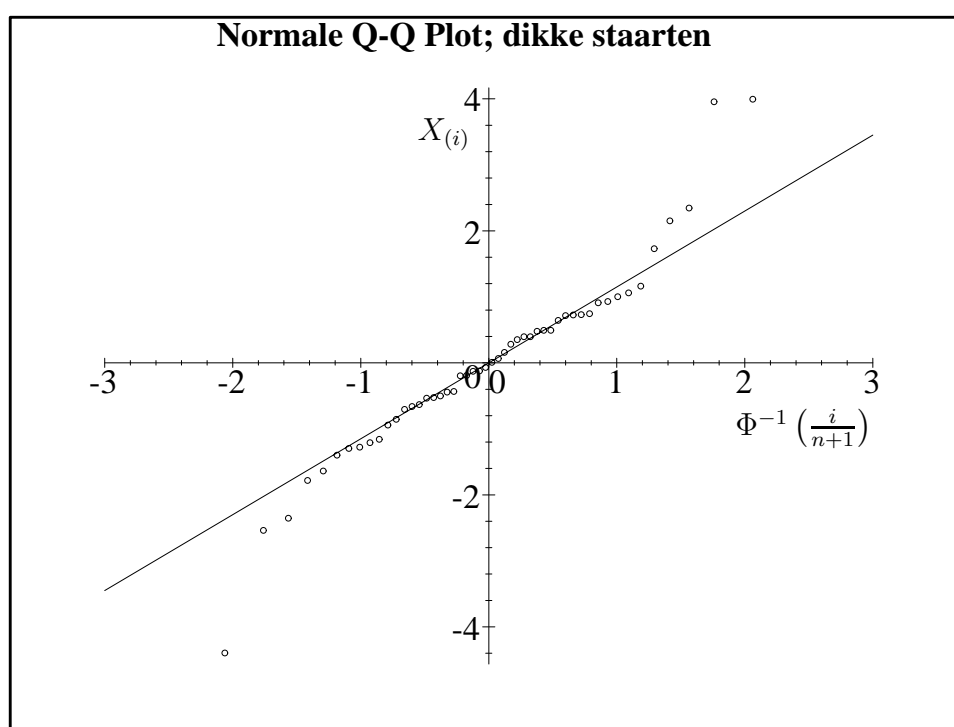
Zijn de waarnemingen exponentieel  $E(\lambda)$ -verdeeld, dan liggen de punten bij een exponentiële Q-Q plot langs een rechte lijn door de oorsprong met richtingscoëfficiënt  $1/\lambda$ . (Immers, als  $U \sim E(1)$  dan is  $(1/\lambda)U \sim E(\lambda)$ .)



**Voorbeeld 2.2.3** Afgebeeld zijn een normale en een exponentiële Q-Q plot van het aantal lynxen in Canada gedurende een aantal jaren. De normale Q-Q plot laat zien dat een normale verdeling niet erg past bij deze waarnemingen. Daarentegen geeft de exponentiële Q-Q plot een redelijk rechte lijn. Deze lijn gaat echter niet door de oorsprong, zodat er aan een model van een verschoven exponentiële verdeling gedacht moet worden.  $\square$

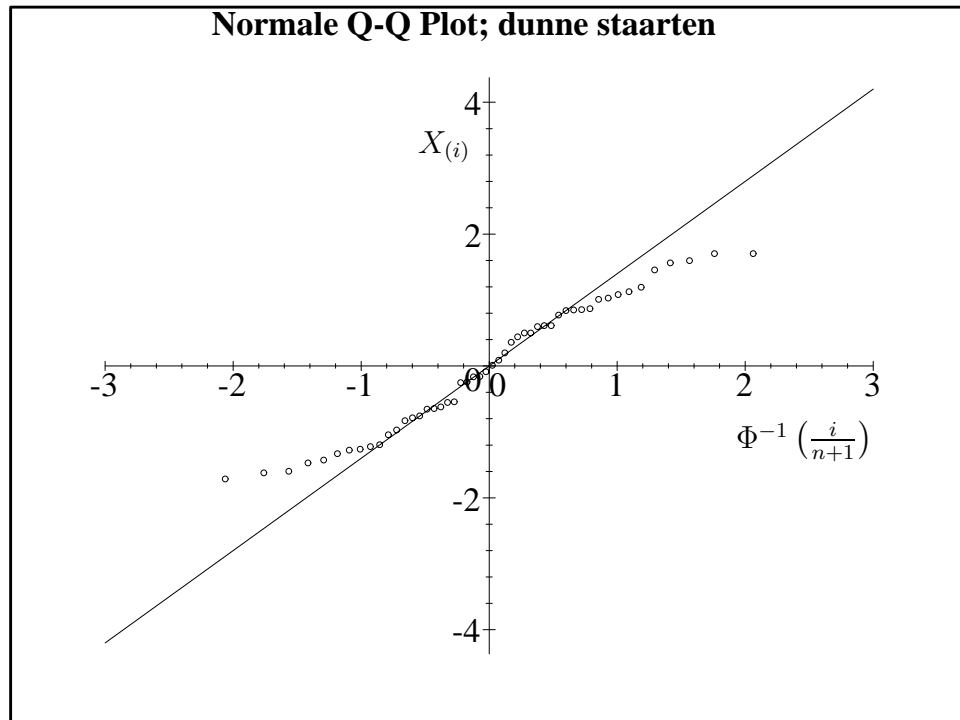
Als een normale Q-Q plot geen rechte lijn oplevert kan zich nogal eens een van de volgende situaties voordoen:

### 1. Dikke staarten



Rechts aan de buitenkant liggen de waarnemingen hoger dan je volgens de rechte lijn zou verwachten: in het gebied van de grote waarden zijn de feitelijke waarden groter dan je op grond van veronderstelde normaliteit zou verwachten; er is een dikke rechterstaart. De waarnemingen wijzen op een verdeling, die meer kansmassa aan grote waarden toekent dan de normale verdeling doet. Links aan de buitenkant zijn de waarnemingen juist kleiner dan je op grond van normaliteit zou verwachten. De waarnemingen wijzen op een kansverdeling die meer kansmassa toekent aan kleine waarden dan de normale verdeling doet, d.w.z. de waarnemingen wijzen op een verdeling met een dikke linkerstaart. Omdat alles er redelijk symmetrisch uitziet kun je een  $t$ -verdeling proberen (mits je in staat bent de betreffende Q-Q plot te maken).

2. **Dunne staarten** Nu liggen de waarnemingen juist rechts onder en links boven de rechte lijn. Er zijn dunne staarten. Probeer de uniforme verdeling.



## 2.3 Pearson's chi-kwadraat

Hebben we een idee gekregen in de exploratieve fase van wat een redelijke modelveronderstelling zou zijn, dan kunnen we in de **confirmatieve fase** op grond van **nieuwe data** formeel toetsen of deze modelveronderstelling hout snijdt. Dit toetsingsprobleem heet het **goodness-of-fit** probleem. We spreken in het Nederlands wel van **toetsen voor aanpassing**. We bespreken drie methoden (van de zeer vele die er zijn). De eerste methode is de oudste, de **chi-kwadraat toets voor aanpassing van Pearson**.

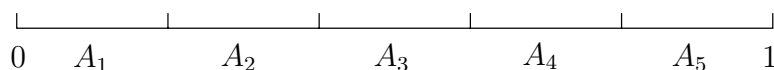
Het toetsingsprobleem dat we hiermee kunnen aanpakken is als volgt. We toetsen de nulhypothese dat  $X_1, \dots, X_n$  een aselechte steekproef is uit een **bekende** verdeling met verdelingsfunctie  $F_0$ . We delen daartoe het mogelijke waardegebied van de waarnemingen (vaak  $\mathbb{R}$ ,  $(0,1)$  of  $(0, \infty)$ ) op, in een aantal klassen, die onder  $F_0$  (ongeveer) gelijke kans hebben. Voor een aantal waarnemingen tussen 25 en 200 is 5 à 7 klassen een goede keuze.

**Voorbeeld 2.3.1** Een pseudo-random-number generator heeft de volgende 50 getallen voortgebracht:

0.6345 0.1653 0.7068 0.7768 0.1913 0.9691 0.4232 0.9703 0.6602  
 0.9669 0.9273 0.9078 0.6429 0.9505 0.3038 0.1078 0.4869 0.4401  
 0.7496 0.9845 0.3418 0.9900 0.1038 0.2144 0.1801 0.8822 0.5459  
 0.3612 0.9985 0.9351 0.7306 0.4692 0.3288 0.5863 0.5243 0.5832  
 0.0129 0.2574 0.3996 0.2036 0.6475 0.4085 0.2624 0.0467 0.3433  
 0.5658 0.6907 0.8585 0.0675 0.9554.

We willen onderzoeken of deze getallen beschouwd kunnen worden als realisaties van s.v.-en  $X_1, \dots, X_{50}$ , die o.o. zijn en een  $U(0, 1)$ -verdeling hebben.

We passen Pearson's chi-kwadraat toets toe. We nemen 5 klassen die onder  $F_0$  (verdelingsfunctie van de  $U(0, 1)$ -verdeling) alle gelijke kans hebben. De eerste klasse is dan  $[0, 0.2)$ , de tweede  $[0.2, 0.4)$  enz., want zo is de kans op iedere klasse bij de  $U(0, 1)$ -verdeling gelijk aan  $1/5$ .



In elk van de 5 klassen zullen dan, als de  $U(0, 1)$ -verdeling juist is, ongeveer  $50 \times \frac{1}{5} = 10$  waarnemingen terecht komen. We vergelijken het **werkelijk aantal** waarnemingen, aangetroffen in elk van de klassen, met dit **verwachte aantal** van 10. In de eerste klasse  $[0, 0.2)$  blijken 8 waarnemingen terecht te zijn gekomen, in de tweede klasse en derde klasse precies 10, in de vierde 9 en tenslotte in de laatste klasse 13. Zijn de afwijkingen in de eerste, vierde en vijfde klasse t.o.v. het verwachte aantal 10 zo onrustbarend dat we de  $U(0, 1)$ -verdeling moeten afwijzen? We berekenen daartoe Pearson's chi-kwadraat toetsingsgrootte. Deze ziet er zo uit

$$(2.3.1) \quad \chi^2 = \sum \frac{(\text{gevonden aantal} - \text{verwachte aantal})^2}{\text{verwachte aantal}},$$

waarbij gesommeerd wordt over alle klassen. Hier levert dat op

$$\chi^2 = \frac{(8 - 10)^2}{10} + \frac{(10 - 10)^2}{10} + \frac{(10 - 10)^2}{10} + \frac{(9 - 10)^2}{10} + \frac{(13 - 10)^2}{10} = 1.4.$$

Deze waarde moeten we vergelijken met de kritieke waarde, die we vinden in de tabel van de chi-kwadraat verdeling. Als aantal vrijheidsgraden krijgen we het aantal klassen  $- 1$ , dus  $5 - 1 = 4$ . Nemen we als onbetrouwbaarheidsdrempel  $\alpha = 0.05$ , dan lezen we af in de tabel 9.49 als kritieke waarde. De waarde 1.4 is dus niet uitzonderlijk groot, zodat we de nulhypothese dat de waarnemingen  $U(0, 1)$ -verdeeld zijn, niet verwerpen.  $\square$

We gaan iets dieper in op de algemene gedaante en achtergrond van Pearson's chi-kwadraat toets. Het kansmodel waar we van uit gaan is als volgt. We hebben o.o. s.v.-en  $X_1, \dots, X_n$  met alle dezelfde verdelingsfunctie  $F$ . De nulhypothese is

$F = F_0$ , waarbij  $F_0$  één gegeven verdelingsfunctie is. In voorbeeld 2.3.1 is dat de  $U(0, 1)$ -verdeling. We bepalen (disjuncte) klassen  $A_1, \dots, A_k$  en definiëren voor  $i = 1, \dots, k$

$$\begin{aligned} N_i &= \text{aantal waarnemingen in klasse } A_i \\ \pi_i &= P(X \in A_i), \end{aligned}$$

waarbij  $X$  verdelingsfunctie  $F$  heeft. Omdat  $F$  onbekend is, zijn  $\pi_1, \dots, \pi_k$  onbekende parameters. De vector  $(N_1, \dots, N_k)$  is **multinomiaal** $(n, \pi_1, \dots, \pi_k)$  verdeeld, d.w.z.

$$P(N_1 = n_1, \dots, N_k = n_k) = \frac{n!}{n_1! \dots n_k!} \pi_1^{n_1} \dots \pi_k^{n_k}.$$

Hierin is  $\pi_1^{n_1} \dots \pi_k^{n_k}$  de kans dat we  $n_1$  keer in klasse  $A_1$ ,  $n_2$  keer in klasse  $A_2, \dots, n_k$  keer in klasse  $A_k$  terecht komen bij één specifieke volgorde en

$$(2.3.2) \quad \frac{n!}{n_1! \dots n_k!} = \binom{n}{n_1} \binom{n - n_1}{n_2} \dots \binom{n - n_1 - \dots - n_{k-1}}{n_k}$$

het aantal mogelijke volgordes. Dit laatste kan als volgt ingezien worden. Van de  $n$  waarnemingen kiezen we er  $n_1$  die in  $A_1$  terecht komen. Dit kan op  $\binom{n}{n_1}$  manieren. Van de overige  $n - n_1$  kiezen we er  $n_2$  voor  $A_2$  enz. De gelijkheid van linker- en rechterlid in (2.3.2) volgt door uitschrijven van de binomiaalcoëfficiënten en gebruikmaken van  $n_1 + n_2 + \dots + n_k = n$ .

Onder de nulhypothese geldt  $F = F_0$  en als  $X$  verdelingsfunctie  $F_0$  heeft, noteren we  $p_i$  in plaats van  $\pi_i$ :  $p_i$  is dus de kans  $P(X \in A_i)$  berekend voor het geval dat  $F_0$  de verdelingsfunctie van  $X$  is. Als  $a_i$  en  $b_i$  begin- en eindpunt van het interval  $A_i$  zijn, vinden we  $p_i = F_0(b_i) - F_0(a_i)$ . Omdat  $F_0$  wel bekend is, zijn  $p_1, \dots, p_k$  ook bekend en ook te berekenen. Zo is in voorbeeld 2.3.1  $p_2$  gelijk aan de kans dat een  $U(0, 1)$ -verdeelde s.v. in het interval  $[0.2, 0.4)$  komt en dus is  $p_2 = 1/5$ .

Het kansmodel ziet er nu zo uit. De vector  $(N_1, \dots, N_k)$  is **multinomiaal** $(n, \pi_1, \dots, \pi_k)$  verdeeld en we toetsen

$$H_0 : (\pi_1, \dots, \pi_k) = (p_1, \dots, p_k).$$

Dit is dus een enkelvoudige nulhypothese: de onbekende parameter  $(\pi_1, \dots, \pi_k)$  is onder  $H_0$  volledig gespecificeerd. Dit multinomiale kansmodel is de basis voor Pearson's chi-kwadraat toets. Als toetsingsgrootheid nemen we

$$\chi^2 = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}.$$

Dit is niets anders dan (2.3.1) in formulevorm, want  $N_i$  is het gevonden aantal in de  $i^e$  klasse en  $np_i$  het onder de nulhypothese verwachte aantal in de  $i^e$  klasse.

Is deze toetsingsgroottheid erg groot dan hebben we een behoorlijke afwijking van wat we onder  $H_0$  hadden verwacht, m.a.w. dan zou de nulhypothese dat de waarnemingen verdelingsfunctie  $F_0$  hebben, wel eens onjuist kunnen zijn. Is de toetsingsgroottheid niet zo groot, dan past de verdeling met verdelingsfunctie  $F_0$  kennelijk redelijk goed bij de waarnemingen.

Wanneer zeggen we nu dat we de toetsingsgroottheid te groot vinden om nog in  $F_0$  te kunnen blijven geloven? We verwerpen  $H_0$  als de waarde van de toetsingsgroottheid groter is dan of gelijk aan de kritieke waarde. Zoals gebruikelijk is in de toetsingstheorie, wordt de kritieke waarde zo gekozen dat, indien  $H_0$  juist is, de kans op verwerpen ten hoogste  $\alpha$  (de onbetrouwbaarheidsdrempel) is. Er geldt dat onder  $H_0$ , de grootheid  $\chi^2$  bij benadering (voor  $n \rightarrow \infty$ ) een chi-kwadraatverdeling heeft met  $k - 1$  vrijheidsgraden. (We bewijzen dit niet.) Dus (met  $P$  refererend aan de kansverdeling onder  $H_0$ )

$$(2.3.3) \quad P\left(\sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \geq x\right) \approx P(\chi_{k-1}^2 \geq x),$$

waarbij  $\chi_{k-1}^2$  een stochastische variabele met een chi-kwadraat( $k - 1$ )-verdeling aanduidt. In de tabel van de chi-kwadraat-verdeling vinden we de waarde  $c$  waarvoor  $P(\chi_{k-1}^2 \leq c) = 1 - \alpha$  en dus

$$(2.3.4) \quad P(\chi_{k-1}^2 \geq c) = \alpha.$$

We nemen nu als kritieke waarde deze  $c$ , waarmee vanwege (2.3.3) en (2.3.4)

$$P\left(\sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \geq c\right) \approx \alpha.$$

De toets ziet er dus zo uit:

$\text{verwerp } H_0 \text{ als } \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \geq c \text{ met } P(\chi_{k-1}^2 \leq c) = 1 - \alpha.$
--

Deze toets heeft niet precies onbetrouwbaarheid  $\alpha$ , maar bij benadering onbetrouwbaarheid  $\alpha$ . Gebleken is dat de benadering in het algemeen zeer dicht bij de gewenste waarde  $\alpha$  zit.

Bij Statistiek & kansrekening hebben we toetsingsproblemen gestructureerd in 8 stappen. De 8 stappen staan hier nog eens vermeld.

1. Formuleer het kansmodel.
2. Formuleer nulhypothese en alternatieve hypothese in termen van de parameters van het kansmodel.
3. Formuleer een geschikte toetsingsgrootheid in termen van de voorkomende s.v.-en.
4. Geef de kansverdeling van de toetsingsgrootheid onder (het randpunt van)  $H_0$ .
5. Bereken of geef de waarde van de toetsingsgrootheid.
6. Bepaal de kritieke waarde(n) en geef het kritieke gebied.

of

- 6\*. Bereken de overschrijdingskans.
7. Formuleer de conclusie omtrent het al dan niet verwerpen van  $H_0$  bij de gegeven onbetrouwbaarheid(sdrempel).
8. Vermeld de conclusie in “gewone woorden”.

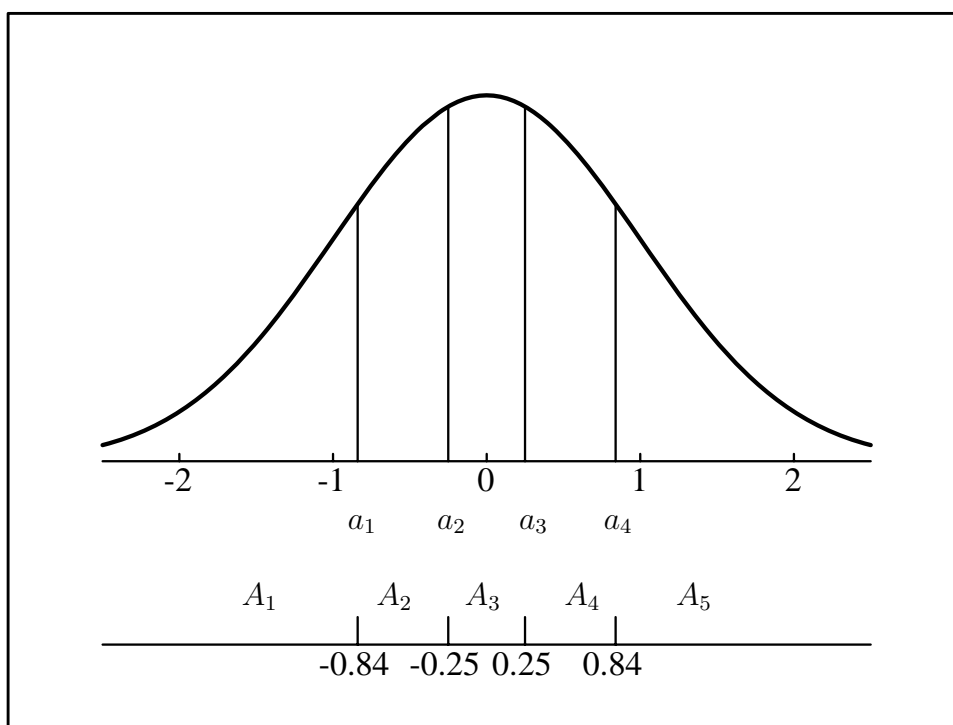
We voeren bij de toetsingsproblemen in dit dictaat veelvoudig deze 8 stappen uit.

**Voorbeeld 2.3.2** (vervolg van voorbeeld 2.3.1) We geven de resultaten van voorbeeld 2.3.1 weer in de vorm van de 8 stappen.

1.  $X_1, \dots, X_{50}$  o.o. s.v.-en met verdelingsfunctie  $F$   
 $N_i$ : aantal waarnemingen in  $A_i$   
 $\pi_i = P(X \in A_i)$  waarbij  $X$  verdelingsfunctie  $F$  heeft,  $i = 1, \dots, 5$   
 $A_1 = [0, 0.2)$ ,  $A_2 = [0.2, 0.4)$ ,  $A_3 = [0.4, 0.6)$ ,  $A_4 = [0.6, 0.8)$ ,  
 $A_5 = [0.8, 1)$ .  
 $(N_1, \dots, N_5) \sim$  multinomiaal( $50, \pi_1, \dots, \pi_5$ ).
2. Onder  $H_0$  geldt  $F = F_0$  met  $F_0$  de verdelingsfunctie van de  $U(0, 1)$ -verdeling implicerend  $H_0 : \pi_i = 1/5 \quad i = 1, \dots, 5$ .  
 $H_1 : (\pi_1, \dots, \pi_5) \neq (1/5, \dots, 1/5)$ .
3.  $\chi^2 = \sum_{i=1}^5 \frac{(N_i - 10)^2}{10}$ .
4. De toetsingsgrootheid  $\chi^2$  heeft onder  $H_0$  bij benadering een chi-kwadratverdeling met 4 vrijheidsgraden.
5. Waarde  $\chi^2 : 1.4$ . (zie voorbeeld 2.3.1)
6. Kritieke waarde: 9.49; kritiek gebied  $[9.49, \infty)$ .
7. Verwerp  $H_0$  niet, want  $1.4 < 9.49$ .
8. Statistisch gezien is er niet voldoende reden om de uniforme verdeling van de pseudo-random-number generator te ontkennen.  $\square$

Ter illustratie laten we in het volgende voorbeeld nog eens zien hoe de klassen berekend worden.

**Voorbeeld 2.3.3** Laat  $X_1, \dots, X_{48}$  o.o. s.v.-en zijn met verdelingsfunctie  $F$ . We willen toetsen  $H_0 : F = F_0$ , waarbij  $F_0$  de verdelingsfunctie is van de  $N(0, 1)$ -verdeling. Stel dat we kiezen voor 5 klassen met ieder gelijke kans onder  $H_0$ . Voor de bovengrens  $a_1$  van klasse  $A_1$  moet gelden  $F_0(a_1) = \Phi(a_1) = \frac{1}{5} = 0.2$ . Uit de tabel van de  $N(0, 1)$ -verdeling lezen we af (terugzoeken bij 0.8)  $\Phi(0.84) = 0.8$ , zodat  $\Phi(-0.84) = 0.2$  (symmetrie gebruiken) en  $a_1$  gelijk is aan  $-0.84$ . De bovengrens  $a_2$  van klasse  $A_2$  is  $-0.25$ , want  $\Phi(0.25) = 0.6$ . Verder is  $a_3 = 0.25$  en  $a_4 = 0.84$ .



□

## 2.4 Toets voor normaliteit, Shapiro-Wilk

Bij het gebruik van Pearson's chi-kwadraat toets is het noodzakelijk dat  $F_0$  volledig bekend is. Immers, anders kunnen we bij een gegeven klasse-indeling de  $p_i$ 's niet berekenen dan wel bij gegeven  $p_i$ 's de klasse-indeling niet bepalen. We kunnen dus bijv. toetsen of de waarnemingen **standaard** normaal verdeeld zijn door te nemen  $F_0 = \Phi$ , de standaardnormale verdelingsfunctie. Dit is een ernstige beperking. Immers, we zouden veel liever toetsen op **normaliteit**, d.w.z. de nulhypothese toetsen dat  $X_1, \dots, X_n$  een aselechte steekproef is uit een  $N(\mu, \sigma^2)$ -verdeling, waarbij we de waarden van  $\mu$  en  $\sigma^2$  verder in het midden laten.

Nogal eens wordt dit opgelost door voor  $F_0$  de normale verdelingsfunctie te nemen

met  $\mu = \bar{X}$  en  $\sigma^2 = S^2$  en vervolgens Pearson's chi-kwadraat toets toe te passen, waarbij men soms het aantal vrijheidsgraden van de chi-kwadraat limietverdeling met 2 verlaagt. **Deze procedure is echter onjuist**, omdat de limietverdeling van de aldus verkregen toetsingsgrootte geen chi-kwadraat verdeling is.

Het is wel mogelijk Pearson's chi-kwadraat toets te modificeren zodat het interessantere probleem van toetsen op normaliteit in het algemeen i.p.v. toetsen op standaardnormaliteit, opgelost kan worden. Het gaat echter buiten het bestek van dit college om hierop in te gaan.

Om toch het zeer vaak voorkomend toetsingsprobleem van normaliteit op te lossen, bespreken we een speciale toets voor dit toetsingsprobleem: de **toets van Shapiro en Wilk**. Het kansmodel is als volgt. De s.v.-en  $X_1, \dots, X_n$  zijn o.o. en gelijk verdeeld, ieder met verdelingsfunctie  $F$ . We toetsen

$$H_0 : F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \text{ voor alle } x \text{ en zekere } \mu \in \mathbb{R}, \sigma > 0.$$

Deze nulhypothese is samengesteld, want onder  $H_0$  is  $F$  niet één vaste verdelingsfunctie, maar behoort  $F$  tot de klasse van **alle** normale verdelingsfuncties. De toetsingsgrootte is

$$W = \left( \sum_{i=1}^n a_i X_{(i)} \right)^2 / \sum_{i=1}^n (X_i - \bar{X})^2$$

waarin  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  de geordende waarnemingen zijn en de constanten  $a_1, \dots, a_n$  in de toegevoegde tabel staan. Er geldt  $a_{n-i+1} = -a_i$ .

De toetsingsgrootte  $W$  is op een constante factor na het kwadraat van het quotiënt van de beste zuivere lineaire schatter van  $\sigma$  en de steekproefstandaardafwijking. Lineair is hier lineair in de geordende waarnemingen. Onder  $H_0$  wijken teller en noemer van  $W$  daarom niet veel van elkaar af (beide zijn na deling door  $n - 1$  goede schatters van  $\sigma^2$ ). Zijn de waarnemingen niet normaal verdeeld, dan is de steekproefvariantie nog altijd een goede schatter van  $\sigma^2$ , maar de teller zal dan in het algemeen aanzienlijk **lager** uitvallen.

Men kan laten zien dat  $W \leq 1$ . Waarden van  $W$  dichtbij 1 wijzen dus op normaliteit (teller en noemer nagenoeg gelijk). We verwerpen de nulhypothese van normaliteit als  $W$  voldoende ver van 1 afwijkt, d.w.z. als  $W \leq c$ , waarbij  $c$  uit tabel 5 afgelezen kan worden. Samengevat:

$$H_0: \text{teller} \approx \text{noemer}; H_1: \text{teller} \ll \text{noemer}; H_0 \text{ verwerpen als } W \leq c.$$

De toets van Shapiro en Wilk heeft indirect ook te maken met de eerder besproken normale Q-Q plot, de exploratieve techniek om normaliteit te onderzoeken. De toets van Shapiro-Wilk is namelijk verwant aan de zgn. toets van De Wet en Venter. De toets van De Wet en Venter hangt sterk samen met de normale Q-Q plot. De toetsingsgrootte van de Wet en Venter is het kwadraat van de “steekproefcorrelatiecoëfficiënt” van de vectoren  $(\Phi^{-1}(\frac{1}{n+1}), \dots, \Phi^{-1}(\frac{n}{n+1}))$  en  $(X_{(1)}, \dots, X_{(n)})$ . Is deze grootte nagenoeg gelijk aan 1, dan wijst dit op een rechte lijn in de normale Q-Q plot, en dus op normaliteit. De toets van Shapiro-Wilk is dus te beschouwen als een beoordelingscriterium van de rechtlijnigheid in een normale Q-Q plot.

Uiteraard kunnen we met de toets van Shapiro-Wilk ook **log-normaliteit** toetsen. Immers,  $X$  is lognormaal verdeeld  $\Leftrightarrow \ln(X)$  is normaal verdeeld. We passen dan op  $\ln(X_1), \dots, \ln(X_n)$  de toets van Shapiro-Wilk toe.

**Voorbeeld 2.4.1** Bij een onderzoek naar luchtverontreiniging werd de  $SO_2$ -concentratie gemeten in microgrammen per kubieke meter lucht in Antwerpen. Dit zijn de data:

136 161 134 118 284 361 244 150 103 203 124 145 255 177 133  
128 175 219 175 273 235 323 205 205 286 362 269 134 228 528

Als we hiervan de natuurlijke logaritme nemen, krijgen we:

4.913 5.081 4.898 4.771 5.649 5.889 5.497 5.011 4.635 5.313  
4.820 4.977 5.541 5.176 4.890 4.852 5.165 5.389 5.165 5.609  
5.460 5.778 5.323 5.323 5.656 5.892 5.595 4.898 5.429 6.269

Ordenen levert op:

4.635 4.771 4.820 4.852 4.890 4.898 4.898 4.913 4.977 5.011  
5.081 5.165 5.165 5.176 5.313 5.323 5.323 5.389 5.429 5.460  
5.497 5.541 5.595 5.609 5.649 5.656 5.778 5.889 5.892 6.269.

Berekening teller  $W$  (zie tabel 5 en gebruik  $a_{n-i+1} = -a_i$ ):

$$\begin{array}{rclcl}
(6.269 - 4.635) & \times & 0.4254 & = & 0.6951 \\
(5.892 - 4.771) & \times & 0.2944 & = & 0.3300 \\
(5.889 - 4.820) & \times & 0.2487 & = & 0.2659 \\
(5.778 - 4.852) & \times & 0.2148 & = & 0.1993 \\
(5.656 - 4.890) & \times & 0.1870 & = & 0.1432 \\
(5.649 - 4.898) & \times & 0.1630 & = & 0.1224 \\
(5.609 - 4.898) & \times & 0.1415 & = & 0.1006 \\
(5.595 - 4.913) & \times & 0.1219 & = & 0.0831 \\
(5.541 - 4.977) & \times & 0.1036 & = & 0.0584 \\
(5.497 - 5.011) & \times & 0.0862 & = & 0.0419 \\
(5.460 - 5.081) & \times & 0.0697 & = & 0.0264 \\
(5.429 - 5.165) & \times & 0.0537 & = & 0.0142 \\
(5.389 - 5.165) & \times & 0.0381 & = & 0.0085 \\
(5.323 - 5.176) & \times & 0.0227 & = & 0.0033 \\
(5.323 - 5.313) & \times & 0.0076 & = & 0.0001 \\
& & & & \hline
& & & & 2.0926
\end{array} +$$

De teller van  $W$  heeft dus als waarde  $(2.0926)^2$ .

Berekening van de noemer van  $W$  levert op: 4.527. Daarom vinden we als realisatie  $w$  van  $W$  de waarde 0.9673. We toetsen bij onbetrouwbaarheidsdrempel  $\alpha = 0.05$ . Uit de tabel lezen we de kritieke waarde af ( $n = 30$   $\alpha = 0.05$ ): 0.927. Omdat we gevonden hebben  $w = 0.9673 > 0.927$  wordt de nulhypothese van lognormaliteit niet verworpen. In termen van de 8 stappen laat bovengenoemd resultaat zich zo omschrijven.

1.  $X_1, \dots, X_{30}$  o.o. s.v.-en met verdelingsfunctie  $F$ .
2. Zij  $G$  de verdelingsfunctie van  $\ln(X_i)$  en  $\mathcal{N}$  de klasse van alle normale verdelingsfuncties.  $H_0 : G \in \mathcal{N}$ ,  $H_1 : G \notin \mathcal{N}$ .
3.  $W = \left( \sum_{i=1}^{30} a_i X_{(i)} \right)^2 / \sum_{i=1}^{30} (X_i - \bar{X})^2$  met  $a_i$  uit de tabel van de Shapiro-Wilk toets.
4. Shapiro-Wilk verdeling.
5.  $w = 0.9673$ .
6. Kritieke waarde: 0.927; kritiek gebied  $[0, 0.927]$ .
7. Verwerp  $H_0$  niet, want  $0.9673 > 0.927$ .
8. Statistisch gezien is er niet voldoende reden om de lognormaliteit van de waarnemingen te ontkennen. □

## 2.5 Toets voor de exponentiële verdeling, Gini's toets

De toetsingsgrootheid van deze toets is gebaseerd op de afstanden tussen de op volgorde gezette waarnemingen. Laat  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  de geordende

waarnemingen zijn, dan gebruikt de toetsingsgrootheid de afstand van de een na kleinste tot de kleinste,  $X_{(2)} - X_{(1)}$ , de afstand van de twee na kleinste tot de een na kleinste,  $X_{(3)} - X_{(2)}$  enz. Deze afstanden krijgen nog een wegingsfactor mee en de som van dit alles wordt nog gedeeld door  $(n - 1) \sum_{i=1}^n X_i$ ; in formule:

$$G = \frac{\sum_{i=1}^{n-1} i(n-i)(X_{(i+1)} - X_{(i)})}{(n-1) \sum_{i=1}^n X_i}$$

**Voorbeeld 2.5.1** Bekijk de volgende 20 waarnemingen: 0.72, 1.10, 2.32, 0.73, 4.11, 0.34, 0.53, 1.34, 0.82, 1.71, 1.05, 0.16, 0.06, 0.49, 0.99, 0.46, 0.79, 2.57, 2.47, 0.57. We berekenen voor deze waarnemingen de waarde van de toets van Gini. Daartoe ordenen we de waarnemingen van klein naar groot: 0.06, 0.16, 0.34, 0.46, 0.49, 0.53, 0.57, 0.72, 0.73, 0.79, 0.82, 0.99, 1.05, 1.10, 1.34, 1.71, 2.32, 2.47, 2.57, 4.11.

We berekenen nu de teller van  $G$ . De eerste term is  $19 \times (0.16 - 0.06) = 1.9$ . De volgende term is  $2 \times 18 \times (0.34 - 0.16) = 6.48$  enz. De laatste term is  $19 \times (4.11 - 2.57) = 29.26$ . De berekening is nogal bewerkelijk, maar met Excel eenvoudig uit te voeren. De teller wordt  $1.9 + 6.48 + \dots + 29.26 = 200.93$ . De noemer is  $19 \times (0.06 + 0.16 + \dots + 4.11) = 443.27$ . De waarde van de toetsingsgrootheid is dus  $200.93/443.27 = 0.453$ .  $\square$

We passen Gini's toets toe als tweezijdige toets en verwerpen dus zowel voor kleine waarden van de toetsingsgrootheid als voor grote waarden van de toetsingsgrootheid. Om de kritieke waarde te bepalen kunnen we voor  $n$  (= aantal waarnemingen)  $\leq 20$  gebruik maken van de tabel achter in het dictaat. Hierbij gebruiken we tevens dat onder de nulhypothese van exponentialiteit de kansverdeling van  $G$  symmetrisch is rond 0.5. Dat betekent dus dat  $G$  en  $1 - G$  onder de nulhypothese dezelfde verdeling hebben. Voor  $n \geq 21$  wordt gebruikt dat  $(G - \frac{1}{2})\sqrt{12(n-1)}$  bij benadering standaard normaal verdeeld is.

**Voorbeeld 2.5.2** We willen toetsen of de waarnemingen uit 2.5.2 exponentieel verdeeld zijn. We nemen als onbetrouwbaarheid  $\alpha = 0.05$ . We passen de 8 stappen toe.

1.  $X_1, \dots, X_{20}$  o.o. s.v.-en ieder met verdelingsfunctie  $F$
2. Zij  $\mathbb{E}$  de klasse van alle exponentiële verdelingsfuncties.  $H_0 : F \in \mathbb{E}, H_1 : F \notin \mathbb{E}$ .

3. Toetsingsgrootheid:  $G = \frac{\sum_{i=1}^{19} i(20-i)(X_{(i+1)} - X_{(i)})}{19 \sum_{i=1}^{20} X_i}$

4. Gini-verdeling

5. De waarde  $g$  van  $G$  is 0.453

6. Kritieke waarden: 0.630 en  $1 - 0.630 = 0.370$ ; kritieke gebied:  $G \leq 0.370 \cup G \geq 0.630$ .

7. Verwerp  $H_0$  niet, want  $0.370 < 0.453 < 0.630$ .

8. Statistisch gezien is er geen reden om exponentialiteit te ontkennen.  $\square$

## 2.6 Data driven toetsen voor normaliteit en exponentialiteit

Recentelijk zijn door Neyman in 1937 gepresenteerde toetsen weer in de belangstelling gekomen. Neyman heeft oorspronkelijk toetsen ontwikkeld om te onderzoeken of de waarnemingen uniform verdeeld zijn op het interval  $(0,1)$ . Deze toetsen kunnen ook eenvoudig aangepast worden om te toetsen of de waarnemingen een precies van te voren vastgelegde verdeling hebben (net zoals bij Pearson's chi-kwadraat toets, zie 3.3).

Een volgende stap is deze toetsen zo te modificeren dat ze gebruikt kunnen worden om te toetsen op normaliteit of op exponentialiteit. Tenslotte moet uit de collectie toetsen die Neyman voorstelt nog een geschikt exemplaar gekozen worden. Dat laatste is in de afgelopen 10 jaar uitvoerig bestudeerd en heeft geleid tot zogenaamde data driven toetsen, die zo genoemd worden omdat de waarnemingen vertellen welk van de Neyman toetsen gekozen moet worden.

We bekijken eerst de data driven toets om exponentialiteit te toetsen. Stel dat we waarnemingen  $X_1, \dots, X_n$  hebben. De toetsingsgrootte wordt als volgt bepaald.

1. Bereken  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ .
2. Bereken  $Y_1 = 1 - e^{-\frac{X_1}{\bar{X}}}, Y_2 = 1 - e^{-\frac{X_2}{\bar{X}}}, \dots, Y_n = 1 - e^{-\frac{X_n}{\bar{X}}}$ .
3. Bereken  $Z_1 = \frac{1}{n} \sum_{i=1}^n h_1(Y_i), \dots, Z_6 = \frac{1}{n} \sum_{i=1}^n h_6(Y_i)$ , waarbij
 
$$h_1(y) = \sqrt{3}(2y - 1),$$

$$h_2(y) = \sqrt{5}(6y^2 - 6y + 1)$$

$$h_3(y) = \sqrt{7}(20y^3 - 30y^2 + 12y - 1)$$

$$h_4(y) = 3(70y^4 - 140y^3 + 90y^2 - 20y + 1)$$

$$h_5(y) = \sqrt{11}(252y^5 - 630y^4 + 560y^3 - 210y^2 + 30y - 1)$$

$$h_6(y) = \sqrt{13}(924y^6 - 2772y^5 + 3150y^4 - 1680y^3 + 420y^2 - 42y + 1).$$
4. Bereken
 
$$nZ_1^2 - \ln(n),$$

$$n(Z_1^2 + Z_2^2) - 2\ln(n),$$

$$n(Z_1^2 + Z_2^2 + Z_3^2) - 3\ln(n),$$

$$n(Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2) - 4\ln(n),$$

$$n(Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2) - 5\ln(n),$$

$$n(Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2) - 6\ln(n)$$
 en bekijk welk van deze het grootste is. Als de eerste het grootste is, wordt  $k = 1$ , als de tweede het grootste is wordt  $k = 2$  etc.

5. De toetsingsgrootheid wordt nu  $T_k$  met  $k$  de in het vorige punt verkregen waarde. Hierbij is

$$T_k = n \sum_{j=1}^k Z_j^2 + n \frac{\left\{ \sum_{j=1}^k a_j Z_j \right\}^2}{1 - \sum_{j=1}^k a_j^2}$$

met

$$a_1 = \frac{1}{2}\sqrt{3} = 0.8660, a_2 = \frac{1}{6}\sqrt{5} = 0.3727, a_3 = \frac{1}{12}\sqrt{7} = 0.2205, \\ a_4 = \frac{3}{20} = 0.15, a_5 = \frac{1}{30}\sqrt{11} = 0.1106, a_6 = \frac{1}{42}\sqrt{13} = 0.0858.$$

Bij onbetrouwbaarheidsdrempel  $\alpha = 0.05$  zijn de kritieke waarden als volgt:

$n = 20$	6.77
$n = 30$	5.69
$n = 50$	4.95
$n = 70$	4.61
$n = 100$	4.35
$n = 150$	4.17
$n = 200$	4.09

Voor tussenliggende waarden van  $n$  vinden we de kritiek waarde door interpolatie.

We bekijken vervolgens het toetsen van normaliteit. Stel dat we waarnemingen  $X_1, \dots, X_n$  hebben. De toetsingsgrootheid wordt als volgt bepaald.

- Bereken  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$  en  $S = \sqrt{S^2}$  met  $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ .
- Bereken  $Y_1 = \Phi\left(\frac{X_1 - \bar{X}}{S}\right), Y_2 = \Phi\left(\frac{X_2 - \bar{X}}{S}\right), \dots, Y_n = \Phi\left(\frac{X_n - \bar{X}}{S}\right)$ .
- Bereken  $Z_1 = \frac{1}{n} \sum_{i=1}^n h_1(Y_i), \dots, Z_6 = \frac{1}{n} h_6(Y_i)$ , waarbij
 
$$h_1(y) = \sqrt{3}(2y - 1),$$

$$h_2(y) = \sqrt{5}(6y^2 - 6y + 1)$$

$$h_3(y) = \sqrt{7}(20y^3 - 30y^2 + 12y - 1)$$

$$h_4(y) = 3(70y^4 - 140y^3 + 90y^2 - 20y + 1)$$

$$h_5(y) = \sqrt{11}(252y^5 - 630y^4 + 560y^3 - 210y^2 + 30y - 1)$$

$$h_6(y) = \sqrt{13}(924y^6 - 2772y^5 + 3150y^4 - 1680y^3 + 420y^2 - 42y + 1).$$
- Bereken
 
$$nZ_1^2 - \ln(n),$$

$$n(Z_1^2 + Z_2^2) - 2\ln(n),$$

$$n(Z_1^2 + Z_2^2 + Z_3^2) - 3\ln(n),$$

$$\begin{aligned}
& n(Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2) - 4\ln(n), \\
& n(Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2) - 5\ln(n), \\
& n(Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2) - 6\ln(n)
\end{aligned}$$

en bekijk welk van deze het grootste is. Als de eerste het grootste is, wordt  $k = 1$ , als de tweede het grootste is wordt  $k = 2$  etc.

5. De toetsingsgrootte wordt nu  $T_k$  met  $k$  de in het vorige punt verkregen waarde. Hierbij is

$$T_k = n \sum_{j=1}^k Z_j^2 + n \frac{\left\{ \sum_{j=1}^k a_{1j} Z_j \right\}^2}{1 - \sum_{j=1}^k a_{1j}^2} + n \frac{\left\{ \sum_{j=1}^k a_{2j} Z_j \right\}^2}{2 - \sum_{j=1}^k a_{2j}^2}$$

met

$$\begin{aligned}
a_{11} &= \frac{\sqrt{3}}{\sqrt{\pi}} = 0.9772, \quad a_{12} = 0, \quad a_{13} = 0.1830, \quad a_{14} = 0, \quad a_{15} = 0.0817, \quad a_{16} = 0, \\
a_{21} &= 0, \quad a_{22} = 1.2328, \quad a_{23} = 0, \quad a_{24} = 0.5211, \quad a_{25} = 0, \quad a_{26} = 0.3045.
\end{aligned}$$

Bij onbetrouwbaarheidsdrempel  $\alpha = 0.05$  zijn de kritieke waarden als volgt:

$n = 20$	3.96
$n = 30$	3.85
$n = 50$	3.84
$n = 70$	3.84
$n = 100$	3.84
$n = 150$	3.84
$n = 200$	3.84

Voor tussenliggende waarden van  $n$  vinden we de kritiek waarde door interpolatie.

## 2.7 (Machts)transformaties

In voorbeeld 2.4.1 hebben we niet de data zelf rechtstreeks bestudeerd, maar de natuurlijke logaritme van de data. Het komt nogal eens voor dat het handiger is niet de data zelf te bestuderen maar eerst een transformatie op de data toe te passen. Zo'n transformatie kan leiden tot een informatiever plaatje van de data, effectievere samenvattingen of een minder gecompliceerde analyse. Redenen voor transformatie zijn:

- makkelijker **interpreteerbaarheid** van de data na transformatie, bijv. de logaritme bij exponentiële groei;
- **asymmetrie** in de data, die mogelijk wordt opgelost door een transformatie;
- de data bestaan uit **verschillende steekproeven**, die verschillen in “midden” en “spreiding” ; het doel van de transformatie is dan om min of meer gelijke

spreiding te krijgen, waarmee vergelijking van de middens eenvoudiger wordt (gelijke schaal).

De meest gebruikte transformaties zijn de zgn. **machtstransformaties**. Ze hebben deze vorm

$$(2.7.1) \quad T_p(x) = \begin{cases} ax^p + b & p \neq 0 \\ c \ln(x) + d & p = 0 \end{cases}$$

met  $ap > 0$  en  $c > 0$ .

De logaritmische transformatie ( $p = 0$ ) kan beschouwd worden als een limiet van  $T_p$  voor  $p \rightarrow 0$ , want

$$\lim_{p \rightarrow 0} \frac{x^p - 1}{p} = \ln(x).$$

Bij speciale keuzes van  $a$ ,  $b$ ,  $c$  en  $d$  gaan de krommen door het punt  $(1,0)$  en past  $\ln(x)$ ,  $p = 0$ , mooi bij de rest:

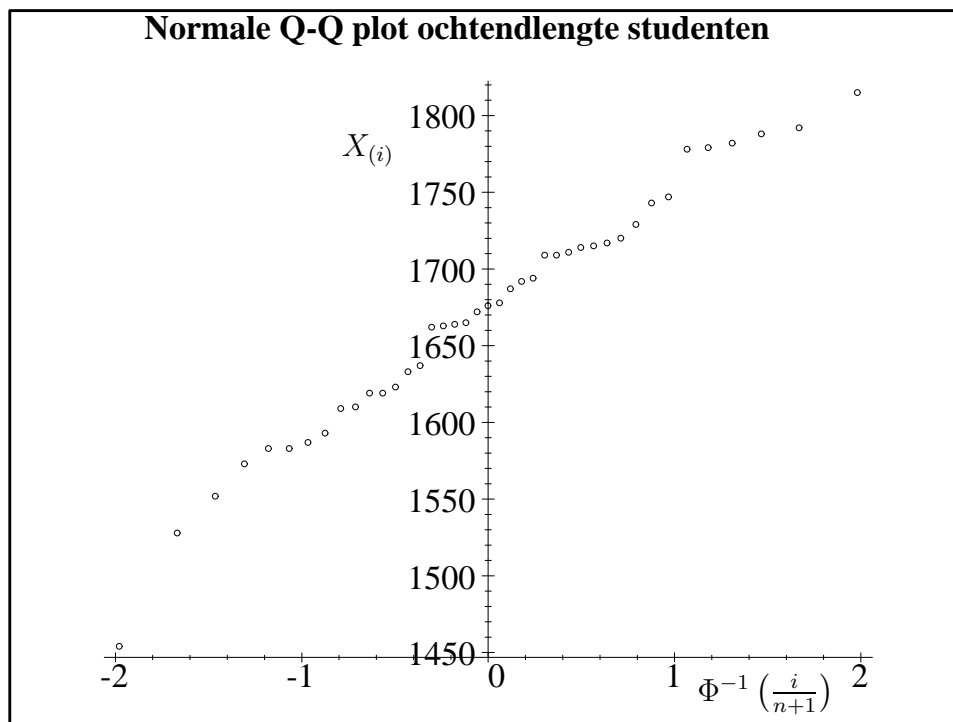
$$T_p^*(x) = \begin{cases} \frac{x^p - 1}{p} & p \neq 0 \\ \ln(x) & p = 0. \end{cases}$$

In het plaatje zien we hoe de transformaties er uit zien voor verschillende waarden van  $p$ .



Bepaal  $x_{(10)}$  en  $x_{(27)}$ . Bij welke punten op de  $x$ -as zijn dit de  $y$ -coördinaten in een normale Q-Q plot?

2. Hieronder staat de normale Q-Q plot voor de gegevens van opg. 1. Geef commentaar. Geef een schatting van  $\mu$  en  $\sigma$  m.b.v. deze Q-Q plot.



3. In de confectie-industrie is men geïnteresseerd in de verdeling van de lengte van volwassen vrouwen. Om te onderzoeken of de lengte normaal verdeeld is (met niet nader gespecificeerde parameters  $\mu$  en  $\sigma^2$ ) heeft men de lengte van 100 aselect gekozen volwassen vrouwen gemeten. De gemiddelde lengte was 168 cm en de steekproefstandaardafwijking bedroeg 6 cm. De chi-kwadraat toets is toegepast. Omdat voor een standaardnormaal verdeelde stochastische variabele  $U$  geldt:

$$P(U \leq -0.97) = P(-0.97 < U \leq -0.43) = P(-0.43 < U \leq 0) = \\ P(0 < U \leq 0.43) = P(0.43 < U \leq 0.97) = P(0.97 < U) = 1/6$$

besloot men de volgende klasse-indeling voor de lengte te maken:

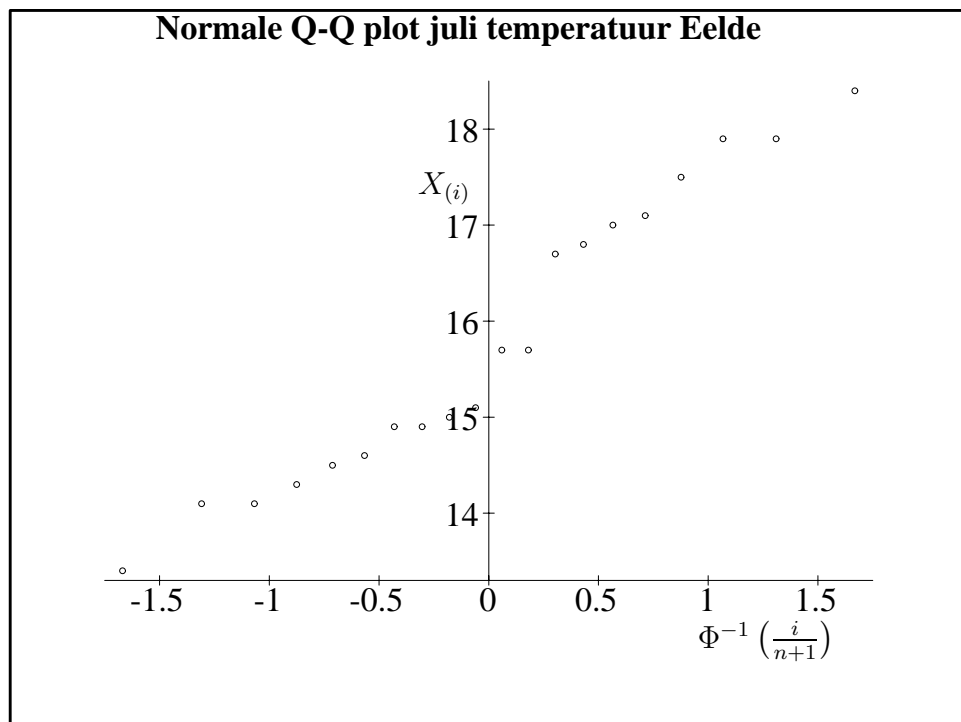
klasse 1	lengte	$\leq$	$168 - 0.97 \times 6 = 162.2$
klasse 2	$162.2 <$	lengte	$\leq 168 - 0.43 \times 6 = 165.4$
klasse 3	$165.4 <$	lengte	$\leq 168$
klasse 4	$168 <$	lengte	$\leq 168 + 0.43 \times 6 = 170.6$
klasse 5	$170.6 <$	lengte	$\leq 168 + 0.97 \times 6 = 173.8$
klasse 6	$173.8 <$	lengte.	

Als toetsingsgrootheid werd genomen

$$\chi^2 = \sum_{i=1}^6 \frac{(N_i - 100 \times 1/6)^2}{100 \times 1/6}$$

met  $N_i$  het aantal waarnemingen in klasse  $i$  voor  $i = 1, \dots, 6$ . De waarde van deze toetsingsgrootheid bleek te zijn 6.2. Omdat men bij onbetrouwbaarheidsdrempel  $\alpha = 0.05$  wenste te toetsen, werd deze waarde vergeleken met 7.81, het 0.95-punt van de  $\chi_3^2$ -verdeling. (Dus  $P(\chi_3^2 \geq 7.81) = 0.05$ ). De conclusie was dat de nulhypothese van normaliteit niet verworpen moest worden bij onbetrouwbaarheidsdrempel  $\alpha = 0.05$ . Bespreek **kort** of de gebruikte methode juist is.

4. Hieronder staat een normale Q-Q plot van de gemiddelde juli temperatuur in Eelde gedurende 20 jaren (1964-1983). Geef commentaar.



5. De numerieke samenvatting van de gegevens van opg. 4 is

steekproefgrootte	20
steekproefgemiddelde	15.8
steekproefvariantie	2.3
steekproefstandaardafwijking	1.5
steekproefscheefheidscoëfficiënt	0.25
steekproefkurtosis	1.8

Welke Q-Q plot zou je, mede gezien opg. 4, nog willen proberen?

6. De gemiddelde juli temperatuur in de Bilt gedurende de jaren 1964 - 1983 bedraagt:

14.9 13.7 14.2 15.3 14.1 18.1 15.7 17.2 17.2 17.0  
15.4 17.8 19.3 17.0 15.4 15.9 15.8 16.3 18.9 20.1

Toets op normaliteit m.b.v. de toets van Shapiro-Wilk. Neem  $\alpha = 0.05$  en voer de 8 stappen om een toetsingsprobleem op te lossen uit.

7. Men wenst te onderzoeken of een normale verdeling een redelijke modelveronderstelling is voor de verdeling van de praktijkgrootte van huisartsen uit opgave 9 van hoofdstuk 1.

- Men maakt daarom een normale Q-Q plot van de gegevens uit deze opgave. Wat zijn de  $x$ - en  $y$ -coördinaat van de waarneming 2173 in deze normale Q-Q plot.
- Men past de toets van Shapiro-Wilk toe op de gegevens van opgave 9 van hoofdstuk 1. De waarde van de toetsingsgrootte  $W$  is 0.9448. Welke conclusie trekt u hieruit omtrent de eventuele normaliteit? Motiveer uw antwoord.

8. Men onderzoekt in de exploratieve fase wat een redelijke modelveronderstelling zou zijn voor de kansverdeling van de waarnemingen. Men heeft daartoe een exponentiële Q-Q plot gemaakt, d.w.z. tegen elkaar uitgezet  $-\ln(1 - i/(n + 1))$  en  $X_{(i)}$  voor  $i = 1, \dots, n$ . Het plaatje geeft punten te zien die vrij behoorlijk rond de lijn  $y = 3x + 2$  liggen. Geef een redelijke modelveronderstelling voor de kansverdeling van de waarnemingen.

9. Men wenst de volgende gegevens te analyseren.

0.1570 -1.9553 -0.8534 2.5127  
-1.3648 0.0727 -1.5813 -0.5948  
1.4583 -0.2997 1.8131 -0.1825  
0.0941 1.1949 -0.1953 -0.3567  
0.3709 0.9371 0.6350 0.5758

Men beschouwt de gegevens als realisaties van de o.o. stochastische variabelen  $X_1, \dots, X_{20}$ . Onderzoek m.b.v. Pearson's chi-kwadraattoets de veronderstelling dat  $X_1, \dots, X_{20}$  alle standaardnormaal verdeeld zijn. Neem als onbetrouwbaarheidsdrempel  $\alpha = 0.05$  en voer de 8 stappen om een toetsingsprobleem op te lossen uit.

10. We willen onderzoeken of een steekproef representatief is voor de populatie waaruit deze steekproef genomen is. Bekend is dat de populatie bestaat uit 7 deelpopulaties, die respectievelijk 27%, 18%, 15%, 14%, 10%, 9% en 7% van de totale populatie bevatten. In de steekproef ter grootte 150 vonden we de volgende aantallen voor de achtereenvolgende deelpopulaties: 43, 27, 31, 20, 11, 10, 8. Ga d.m.v. een geschikte toets na of de in de steekproef gevonden aantallen significant afwijken van de te verwachten aantallen. Neem als betrouwbaarheidsdrempel  $\alpha = 0.05$ . Voer de 8 stappen om een toetsingsprobleem op te lossen uit.
11. Als  $X_1, \dots, X_n$  o.o. stochastische variabelen zijn, ieder met een  $N(\mu, \sigma^2)$ -verdeling, zijn dan ook  $X_{(1)}, \dots, X_{(n)}$  o.o. en  $N(\mu, \sigma^2)$ -verdeeld? En  $\Phi^{-1}\left(\frac{1}{n+1}\right), \dots, \Phi^{-1}\left(\frac{n}{n+1}\right)$ ? Argumenteer uw antwoorden.

12. Read en Cowan (1976) hebben bij 595 windhondenraces gekeken welk startnummer de winnende hond had. Dit zijn de resultaten.

startnummer winnende hond	1	2	3	4	5	6	7	8
aantal races	104	95	66	63	62	58	60	87

Toets de hypothese dat alle 8 startnummers met gelijke kans een winnaar opleveren. Neem als onbetrouwbaarheidsdrempel  $\alpha = 0.05$ . Voer de 8 stappen uit om een toetsingsprobleem op te lossen.

13. De uniforme Q-Q plot gebaseerd op een aselechte steekproef  $X_1, \dots, X_{25}$  van  $X$  laat punten zien die rond de lijn  $y = 2x + 3$  liggen. Geef op basis hiervan een schatting van de verwachting van  $X$ . Argumenteer uw antwoord.
14. a. Laat  $X_1, \dots, X_{53}$  een aselechte steekproef zijn uit een verdeling met verdelingsfunctie  $F$ . We toetsen de nulhypothese dat  $F = \Phi$ , de standaardnormale verdelingsfunctie. We gebruiken Pearson's chi-kwadraat toets met 6 klassen die alle gelijke kans hebben onder de nulhypothese. Bepaal de 6 klassen.
  - b. In de situatie van onderdeel a. vinden we in de 6 klassen respectievelijk 12, 10, 8, 5, 8 en 10 waarnemingen. Bereken de (benaderde) overschrijdingskans. Welke conclusie trekt u?
15. Bekijk de gegevens van opgave 11 van hoofdstuk 1. Bepaal  $x_{(4)}$  voor deze gegevens en de hierbij behorende  $x$ -coördinaat in de normale Q-Q plot van deze gegevens.

16. Enige jaren geleden stond een bericht in de krant dat de lottoballetjes en lottomachine, die wekelijks op de televisie figureren, op het kansspel laboratorium bij het Nederlands Meetinstituut in Delft onderzocht zouden worden. Er wordt 5000 keer gespeeld met 5 ballen, genummerd  $1, \dots, 5$  in de machine (in plaats van de gebruikelijke 45). Bovendien wordt bij elke trekking slechts één bal getrokken en deze wordt vervolgens weer teruggelegd in de lottomachine, waarna de volgende trekking plaatsvindt. Het resultaat van de trekkingen is 988 keer bal 1, 971 keer bal 2, 1058 keer bal 3, 1011 keer bal 4 en 972 keer bal 5. Ga m.b.v. Pearson's chi-kwadraat toets na of de lottomachine elke bal een gelijke kans geeft om te worden geselecteerd. Voer hiertoe de 8 stappen uit om een toetsingsprobleem op te lossen. Neem als onbetrouwbaarheidsdrempel  $\alpha = 0.05$ .

17. a. Maak een uniforme Q-Q plot voor de volgende gegevens

5.11 -0.55 8.47 2.44 5.02 4.62 -0.85 1.00 7.22.

- b. Acht u een uniforme verdeling geschikt als modelveronderstelling? Zo ja, welke (uniforme) verdeling stelt u voor? Zo nee, waarom vindt u de uniforme verdelingen ongeschikt en welke verdeling stelt u dan voor te onderzoeken?

18. In een ziekenhuis worden gedurende 24 uur de tijden (in minuten) tussen opeenvolgende geboortes van baby's geregistreerd. We willen onderzoeken of deze tijden gemodelleerd kunnen worden met een exponentiële verdeling. De 43 tijden zijn als volgt:

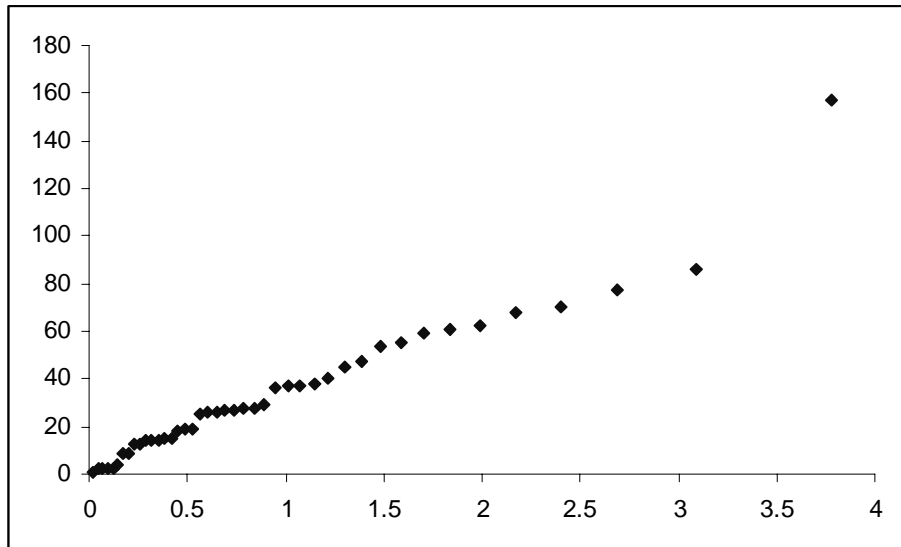
59, 14, 37, 62, 68, 2, 15, 9, 157, 27, 37, 2, 55, 86, 14, 4, 40, 36, 47, 9, 61, 1, 26, 13, 28, 77, 26, 45, 25, 18, 29, 15, 38, 2, 2, 19, 27, 14, 13, 19, 54, 70, 28.

De geordende waarnemingen zijn als volgt.

1, 2, 2, 2, 2, 4, 9, 9, 13, 13, 14, 14, 14, 15, 15, 18, 19, 19, 25, 26, 26, 27, 27, 28, 28, 29, 36, 37, 37, 38, 40, 45, 47, 54, 55, 59, 61, 62, 68, 70, 77, 86, 157.

We maken eerst een exponentiële Q-Q plot.

- a. Bereken de  $x$ -coördinaat behorend bij de waarneming 59.  
 b. De exponentiële Q-Q plot ziet er als volgt uit.



Geef commentaar op het resultaat.

- c. Bereken de eerste en de laatste term uit de teller van de toetsingsgrootheid van Gini.
  - d. Toets op exponentialiteit bij onbetrouwbaarheidsdrempel  $\alpha = 0.05$ . Voer daartoe de 8 stappen uit. De waarde van de toetsingsgrootheid van Gini is: 0.455.
  - e. Geef op grond van de Q-Q plot een schatting van de parameter  $\lambda$  aannemende dat de waarnemingen  $E(\lambda)$ -verdeeld zijn.
  - f. Voor een exponentiële verdeling geldt dat de verwachting gelijk is aan  $1/\lambda$ . Daarom (zie ook Statistiek & kansrekening) schatten we  $\lambda$  met  $\frac{1}{\bar{x}}$ . Er geldt  $\sum x_i = 1430$ . Bereken de waarde van deze schatter en vergelijk die met de waarde verkregen bij e.
19. Onderzoek m.b.v. de data driven toets op exponentialiteit of de waarnemingen van opgave 18 als exponentieel verdeeld mogen worden beschouwd. Neem als onbetrouwbaarheidsdrempel  $\alpha = 0.05$ .
  20. Onderzoek m.b.v. de data driven toets op normaliteit of de waarnemingen van voorbeeld 3.4.1 lognormaal verdeeld zijn. Neem als onbetrouwbaarheidsdrempel  $\alpha = 0.05$ .