Statistiek voor Informatiekunde (I00099)

Bernd Souvignier

voorjaar 2005

Inhoud

Aanbevolen literatuur

- Larray Gonick, Woollcott Smith: The Cartoon Guide to Statistics. HarperResource, 1993, 240 p., ISBN: 0-06-273102-5
  nederlandse versie hiervan:
  Larray Gonick, Woollcott Smith: Het stripverhaal van de statistiek. Epsilon Uitgaven 32, 2004, 240 p., ISBN: 90-5041-037-5

- Murray R. Spiegel, Larry J. Stephens: (Schaum's Outline of Theory and Problems of) Statistics. McGraw-Hill Companies, 1999, 512 p., ISBN: 0-07-060281-6.

Les 1 Beschrijvende statistiek

In de statistiek gaat het erom, vanuit waargenomen gegevens een model te ontwikkelen dat de gegevens goed kan verklaren. Meestal houdt het model een kansverdeling in, daarom bestaat er een grote overlap tussen de methoden van de statistiek en van de kansrekening. Het verschil ligt erin dat men in de kansrekening een proces veronderstelt dat volgens een kansverdeling waarden met zekere kansen produceert, terwijl men in de statistiek van gegevens uitgaat die een zekere frequentieverdeling hebben en probeert conclusies over een hier achter liggende kansverdeling te trekken. In zekere zin bekijken dus kansrekening en statistiek dezelfde vraagstukken uit verschillende invalshoeken.

1.1 Representatie van gegevens

In de statistiek gaat het vooral om het onderzoeken van gegevens die op een of ander manier verzameld zijn, bijvoorbeeld door een of meerdere metingen of door een enquête. Om uitspraken over de gegevens te kunnen doen en structuren erin te kunnen herkennen, is het belangrijk om overzicht over de gegevens te krijgen.

Voorbeeld: We zullen in deze les vaker naar het volgende voorbeeld van gegevens kijken (resultaten bij een zekere toets):

$$54, 41, 59, 45, 34, 49, 58, 30, 61, 47, 43, 48, 80, 27, 56, 45.$$

Meestal is het niet zo handig, de gegevens gewoon op een rij te zetten, omdat de structuur dan verborgen blijft. Daarom worden verschillende manieren toegepast om gegevens grafisch te representeren.

We gaan er nu van uit dat we het over gegevens hebben, die numerieke waarden voor een eigenschap van zekere individuën zijn. Denk hierbij aan de uitslagen van studenten bij een tentamen, de lengte van kinderen op tienjarige leeftijd of iets dergelijks. Het is duidelijk dat het beschrijven van de type van de gegevens afhangt, deze kunnen discrete waarden, zo als aantallen hebben, maar ook continue waarden, waar in principe elke waarde mogelijk is. Natuurlijk zijn er ook gegevens die niet numeriek zijn, zo als eigenschappen, hobbies etc., maar deze kunnen we als gegevens met discrete waarden behandelen, door bijvoorbeeld de verschillende mogelijkheden te nummeren.

Maar eigenlijk bestaan er in de praktijk bijna nooit gegevens met echt continue waarden. Als je bijvoorbeeld naar de resultaten van een competitie in het verspringen kijkt, dan zijn die altijd op centimeters nauwkeurig aangegeven, terwijl we toch ook makkelijk millimeters zouden kunnen meten. Hetzelfde geldt voor tijden, die worden bijvoorbeeld bij het zwemmen in honderdste seconden aangegeven, ook al worden ze nauwkeuriger gemeten (namelijk minstens op duizendsten).

> Bij de olympische spelen van München 1972 hadden er over de 400m wisselslag bij het zwemmen de zweed Gunnar Larsson en de amerikaan Tim McKee een tijd van 4:31,98 minuten. Maar er werden ook duizendsten seconden gemeten en de preciezere tijden waren 4:31,981 voor

> Larsson en 4:31,983 voor McKee. Men heeft toen Larsson de gouden en McKee de zilveren medaille toegekend. Maar sindsdien is er besloten, om de metingen achter de honderdste seconden gewoon te negeren en bij een *dead race* twee gouden medailles uit te reiken.

Vaak worden waarden door *afronden* gediscretiseerd, alle waarden die in een zeker interval liggen worden hierbij door dezelfde waarde vervangen. We zouden ons daarom op gegevens met discrete waarden kunnen beperken, maar we zullen zien dat het vaak handig is, een verdeling door een continue functie te beschrijven.

Merk op: Bij het rekenen met afgeronde waarden neemt de nauwkeurigheid (in het algemeen) bij elke bewerking af. Het is daarom verstandig, zo lang mogelijk met hoge nauwkeurigheid te rekenen en pas het uiteindelijke resultaat af te ronden.

Bij het optellen worden de *absolute fouten* bij elkaar opgeteld, want

$$(x + \Delta x) + (y + \Delta y) = (x + y) + (\Delta x + \Delta y).$$

Bij het vermenigvuldigen worden de *relatieve fouten* bij elkaar opgeteld, want uit

$$(x + \Delta x) \cdot (y + \Delta y) = x \cdot y + \Delta x \cdot y + \Delta y \cdot x + \Delta x \cdot \Delta y$$

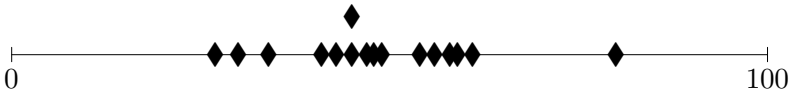
volgt voor $\Delta(x \cdot y) = (x + \Delta x) \cdot (y + \Delta y) - x \cdot y$:

$$\frac{\Delta(x \cdot y)}{x \cdot y} \approx \frac{\Delta x}{x} + \frac{\Delta y}{y}$$

waarbij we de term met twee Δ's hebben weggelaten. Als dus de zijden van een blok met een nauwkeurigheid van 5% gemeten kunnen worden en het volume van de blok als product van de zijden wordt berekend, heeft het volume slechts nog een nauwkeurigheid van 15%.

Stengel-en-blad diagram

Een eenvoudige mogelijkheid om waarden te representeren bestaat erin, de waarden op een lijn te markeren. Dit geeft soms al een overzicht waar de waarden liggen en waar bijvoorbeeld veel punten dicht bij elkaar liggen en hoe ver ze verspreid zijn. Voor ons voorbeeld ziet dit er zo uit:



Natuurlijk is er een probleem als we twee keer dezelfde waarde hebben, wat natuurlijk vooral bij discrete gegevens het geval is. We kunnen dit (zo als in het plaatje) bijvoorbeeld oplossen, door punten voor dezelfde waarde boven elkaar te zetten.

Een representatie die dit idee opneemt is het *stengel-en-blad* diagram, waarbij we alle waarden in een zeker interval naast elkaar schrijven. In het voorbeeld

nemen we het eerste cijfer van een waarde (de tienen) als waarde op de stengel, het laatste cijfer komt dan als blad erachter te staan. Vervolgens worden de bladen die achter een waarde op de stengel staan op volgorde gesorteerd. Voor ons voorbeeld ziet het stengel-en-blad diagram als volgt uit:

2	7
3	0 4
4	1 3 5 5 7 8 9
5	4 6 8 9
6	1
7	
8	0

Deze manier om waarden samen te vatten is al een speciaal voorbeeld voor het vormen van *klassen* die we nu gaan behandelen.

1.2 Klassen

Vaak is het handig om verschillende waarden samen te vatten die op een of ander manier op elkaar lijken. De zo samengevatte waarden noemt men dan een *klasse* van waarden. Als voorbeelden van klassen hebben we al intervallen gezien, waarbij alle waarden tussen zekere grenzen in een pot gegooid worden. Maar er zijn ook heel andere klassen mogelijk, bijvoorbeeld kunnen de woorden in een tekst op totaal verschillende manieren in klassen ingedeeld worden:

- aantal letters in het woord;
- aantal klinkers in het woord;
- syntactische klasse (werkwoord, naamwoord, artikel enz.);
- semantische klasse (wiskundig begrip, kleur, uitdrukking van beweging).

Als we eindig veel gegevens op klassen verdelen, krijgen we een frequentieverdeling voor de klassen, en als we naar de relatieve frequenties van de klassen kijken, voldoen deze aan de eisen van een kansverdeling.

> Merk op dat er een subtiel verschil is tussen een kansverdeling en de frequentieverdeling van klassen: Bij een kansverdeling veronderstellen we een proces die waarden met zekere kansen *produceert*, terwijl de frequentieverdeling gewoon een verzameling van gegevens *beschrijft*. Maar natuurlijk is het vaak nuttig een waargenomen frequentieverdeling met bekende kansverdelingen te vergelijken.

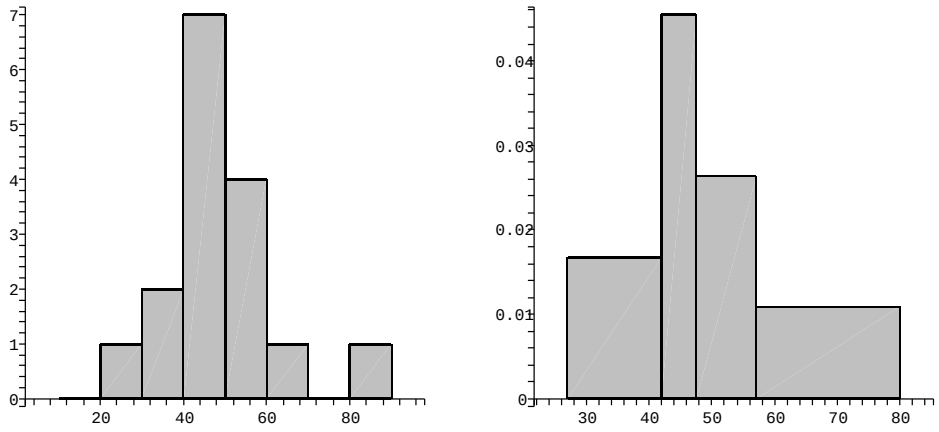
De indeling in klassen is een belangrijke voorwaarde voor de interpretatie van de gegevens. Te veel klassen geven vaak alleen maar versplinterde informatie omdat heel weinig gegevens in een klasse terecht komen, terwijl te weinig klassen geen structuur meer laten herkennen. Soms wordt als vuistregel gehanteerd, bij n gegevens het aantal klassen als $1+{}^2\log(n)$ te kiezen, maar dit is ook niet meer dan een heuristische gok.

Soms kan zelfs de verschuiving van de klassen kritiek zijn, omdat er een duidelijk grootste klasse over twee ongeveer even grote maar veel kleinere klassen verdeeld wordt.

De frequentieverdelingen van klassen laten zich op verschillende manieren grafisch representeren. We zullen de meest belangrijke vormen kort bespreken.

Histogram

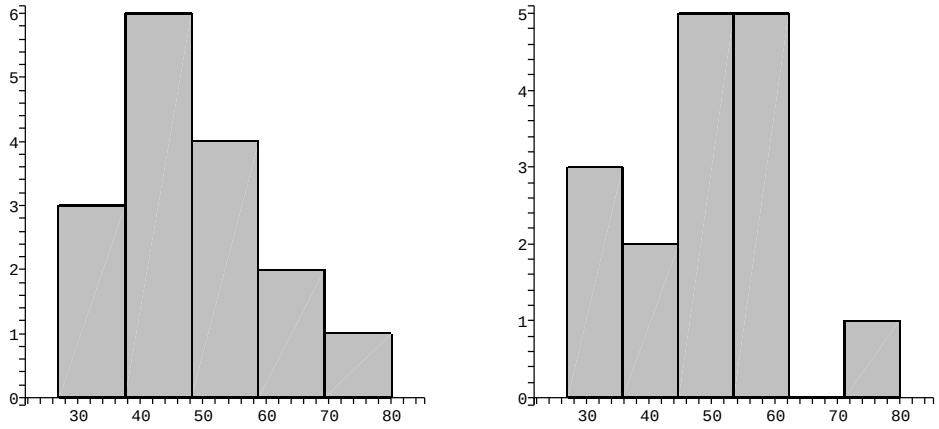
Bij een *histogram* worden de klassen door balken vertegenwoordigd, waarbij de *oppervlakte* van de balken de frequenties representeert. Als de balken ook dezelfde breedte hebben, zijn natuurlijk ook de hoogtes van de balken proportioneel met de frequenties. In Figuur 1 zijn twee histograms voor ons voorbeeld te zien: In het linkerplaatje zijn de klassen intervallen van breedte 10, in het rechterplaatje zijn de klassen automatisch zo gekozen dat elke klasse even veel (in dit geval 4) punten bevat, en de balken dezelfde oppervlakte hebben.



Figuur 1: Histograms met balken van dezelfde en verschillende breedtes.

Als we in ons voorbeeld het aantal klassen volgens de formule $1 +^2 \log(n)$ zouden kiezen, hadden we 5 klassen nodig. De histograms in Figuur 2 laten zien dat een opsplitsing in 5 of 6 klassen een duidelijk kwalitatief verschil in de histograms veroorzaakt: In het eerste geval is er een duidelijk grootste klasse, in het tweede geval zijn er twee grootste klassen en men kan zien dat er een uitschieter is, omdat er een gat tussen de klasse met de maximale waarde en de andere klassen valt.

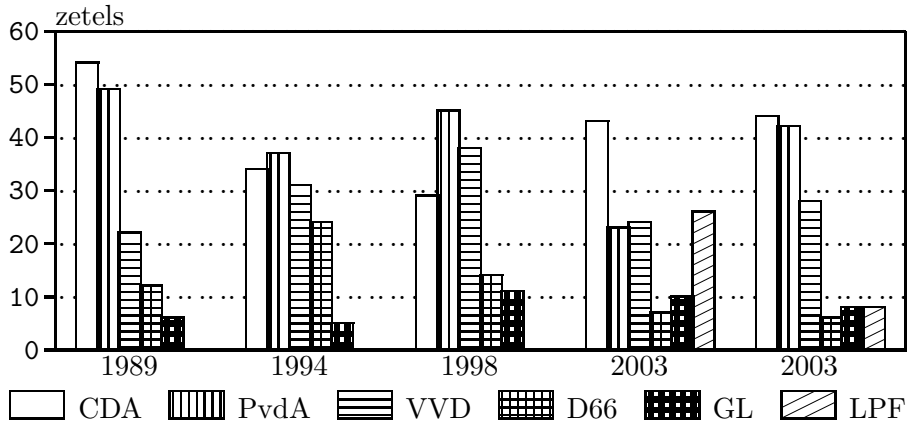
Er kunnen ook histograms van meerdere verzamelingen gegevens in een grafiek gecombineerd worden. Dit wordt vaak gebruikt om de ontwikkeling over de tijd te laten zien. De volgende tabel geeft het aantal zetels in de Tweede Kamer weer voor de verkiezingen sinds 1989 (beperkt tot partijen die in een van de verkiezingen minstens 10 zetels heeft gehaald).



Figuur 2: Histograms met 5 en 6 klassen.

Partij	1989	1994	1998	2002	2003
CDA	54	34	29	43	44
PvdA	49	37	45	23	42
VVD	22	31	38	24	28
D66	12	24	14	7	6
GroenLinks	6	5	11	10	8
LPF	0	0	0	26	8

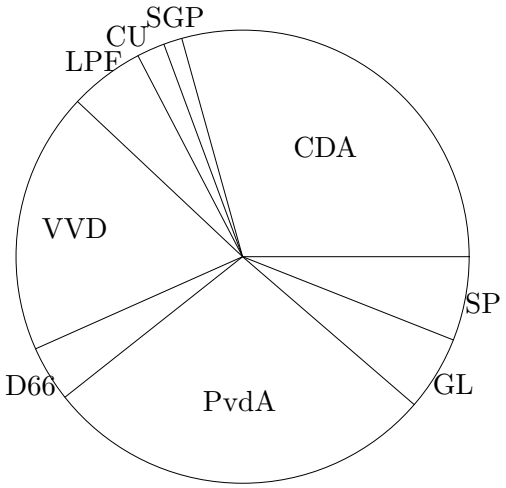
Als we voor elke partij een histogram voor het aantal zetels in de verschillende verkiezingen maken, ziet de combinatie van deze histograms uit als in Figuur 3 te zien. Natuurlijk kan men ook de verdelingen van zetels in een verkiezing als histogram zien, dan worden in deze grafiek gewoon verschillende histograms naast elkaar gezet.



Figuur 3: Verdeling van zetels in de Tweede Kamer.

Taart-diagram

Bij een *taart-diagram* (*pie chart*) wordt een cirkelschijf zo onderverdeeld dat de oppervlaktes van de sectoren de frequenties van de klassen representeren. Omdat de oppervlakte van een sector evenredig is met de hoek van de sector, geven ook de hoeken van de sectoren de frequenties weer. Voor de verkiezingen van 2003 is dit in Figuur 4 te zien.



Figuur 4: Taart-diagram voor de verdeling van zetels in de Tweede Kamer.

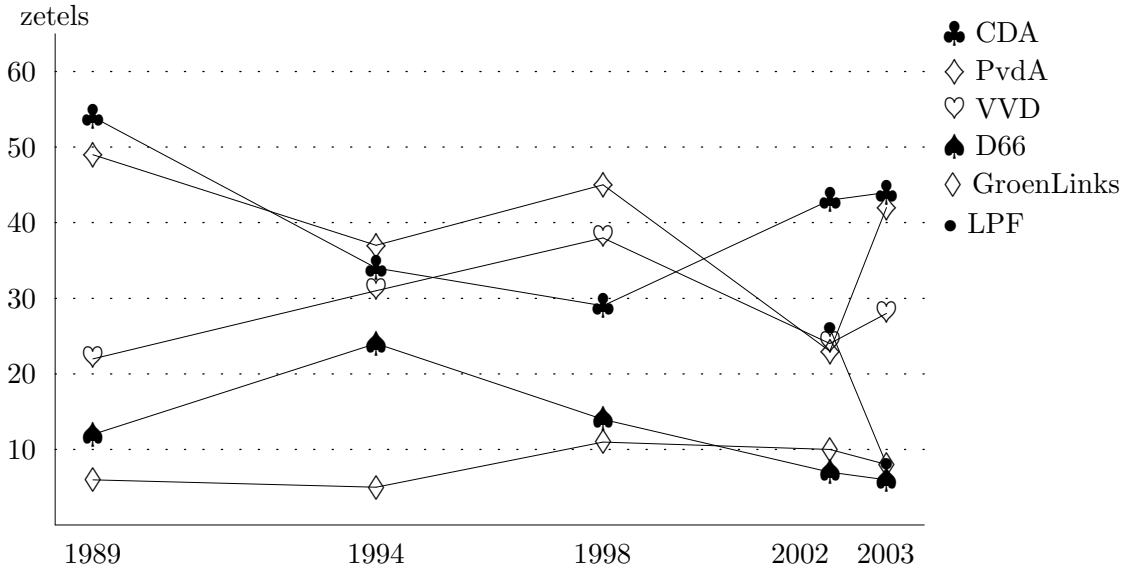
Frequentiepolygoon

In plaats van verschillende histograms in een grafiek te combineren, kan men ook de waarden van verschillende verdelingen over de tijd door *frequentiepolygonen* aangeven. Hierbij worden de waarden voor verschillende tijdstippen (bijvoorbeeld) door lijnstukken verbonden. Merk op dat de tussenwaarden meestal geen betekenis hebben. Ook al kun je op een lijnstuk tussen de verkiezingen van 1994 en 1998 een waarde voor het jaar 1996 aflezen, zegt dat niets over een mogelijke uitslag van verkiezingen in het jaar 1996. De ontwikkeling van het aantal zetels in de Tweede Kamer die in Figuur 3 door een combinatie van histograms beschreven werd, wordt in Figuur 5 door *frequentiepolygonen* gerepresenteerd.

Vervalsende representatie

Het kiezen van een vorm van representatie houdt altijd een manipulatie van de gegevens in. Dit hoeft niet per se negatief te zijn, want *een plaatje zegt meer dan duizend woorden*. Maar door een specifieke keuze van representatie kan er wel een zekere tendentie aan de gegevens gegeven worden. Dit leidt soms - bewust of onbewust - tot een vervalsing van de gegevens. Een paar bekende vervalsingen zijn:

- Schaling van de assen. Hierdoor wordt het stijgen of dalen stijler of vlakker en de veranderingen worden versterkt of verzwakt weergegeven.



Figuur 5: Frequentiepolygonen voor de verdeling van zetels in de Tweede Kamer.

- Afbreken van de y-as boven het nulpunt. Hierdoor lijken veranderingen veel extremer dan ze in werkelijkheid zijn.
- 'Slimme' keuze van klassen. Hierdoor kunnen effecten kunstmatig voortgebracht of onderdrukt worden.
- Representeren van de frequentie door een figuur waarvan de hoogte proportioneel met de frequentie is. Omdat de oppervlakte en niet de hoogte van de figuur waargenomen wordt, lijkt een twee keer zo hoge figuur vier keer zo groot.
- Suggereren van een *ontwikkeling* door representatie middels frequentiepolygonen.

1.3 Typische waarden

Om verschillende verzamelingen van gegevens te kunnen vergelijken, is het vaak handig om een *typische waarde* voor een verzameling aan te geven. Er zijn verschillende manieren, om zo'n typische waarde te definiëren, en er is geen *juiste* manier.

Het gemiddelde

Het *rekenkundig gemiddelde* (meestal kort *gemiddelde* genoemd) van waarden $x_1, x_2, \ldots, x_n$ is gedefinieerd door

$$\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

De interpretatie hiervan is dat de gegevens bij elkaar opgeteld worden en vervolgens de som gelijkvormig over de individuën verdeeld wordt.

Een karakterisering van het gemiddelde is de eigenschap dat de verschillen tussen de gegevens en het gemiddelde bij elkaar opgeteld 0 geven, dus dat $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. Maar de belangrijkste eigenschap van het gemiddelde $\overline{x}$ is, dat het juist de waarde x is waarvoor de som van de kwadratische afstanden van de x_i minimaal wordt, dus waarvoor de functie

$$f(x) := \sum_{i=1}^{n}(x_i - x)^2$$

minimaal wordt. Dit wordt vaak zelfs als definitie van de gemiddelde gebruikt. Een minimum van $f(x)$ vinden we als nulpunt van de afgeleide $f'(x)$. Er geldt $f'(x) = \sum_{i=1}^{n}(2x - 2x_i)$ en dus $f'(x) = 0$ voor $n \cdot x = \sum_{i=1}^{n} x = \sum_{i=1}^{n} x_i$. Omdat de functie $f(x)$ een naar boven geopende parabool is, is dus $x = \overline{x}$ het eenduidige minimum van de functie.

We kunnen het gemiddelde ook in samenhang met kansverdelingen interpreteren. Als we ons voorstellen dat de x_i waarden van een stochast X zijn, die met kans p_x het resultaat x oplevert, dan zullen we de waarde x in een verzameling van n waarden ongeveer $p_x \cdot n$ keer verwachten. Maar als we nu bij het gemiddelde $\overline{x}$ niet meer de som over de x_i maar over de waarden x met hun frequenties nemen, zien we dat $\overline{x}$ een benadering van de verwachtingswaarde $E[X] = \sum_x x \cdot p_x$ van de stochast X is.

Met een analoog argument zien we voor een stochast X met continue kansverdeling met dichtheidsfunctie $f(x)$ dat $\overline{x}$ ook hier een benadering van de verwachtingswaarde $E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$ is.

De mediaan

De *mediaan* $\tilde{x}$ van een verzameling gegevens is gedefinieerd als de waarde die in het midden van de geordende waarden ligt. Dit wil zeggen dat er even veel waarden kleiner dan $\tilde{x}$ zijn als er waarden groter zijn. Als we aannemen, dat de waarden opstijgend geordend zijn, dus $x_1 \leq x_2 \leq \ldots \leq x_n$, dan is voor oneven $n = 2m + 1$ de mediaan $\tilde{x}$ juist de middelste waarde x_m. Voor een even aantal $n = 2m$ neemt men gewoon het gemiddelde van de twee middelste waarden, dus $\tilde{x} = \frac{1}{2}(x_m + x_{m+1})$. Voor opstijgende waarden $x_1 \leq x_2 \leq \ldots \leq x_n$ hebben we dus:

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{als } n \text{ oneven} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{als } n \text{ even.} \end{cases}$$

We hebben gezien dat de som van de verschillen tussen de waarden en het gemiddelde $\overline{x}$ nul geeft en dat het gemiddelde de kwadratische afstanden minimaliseert. De mediaan heeft de eigenschap dat hij de gewone afstanden minimaliseert, dus dat $\tilde{x}$ de waarde is waarvoor

$$g(x) := \sum_{i=1}^{n} |x_i - x|$$

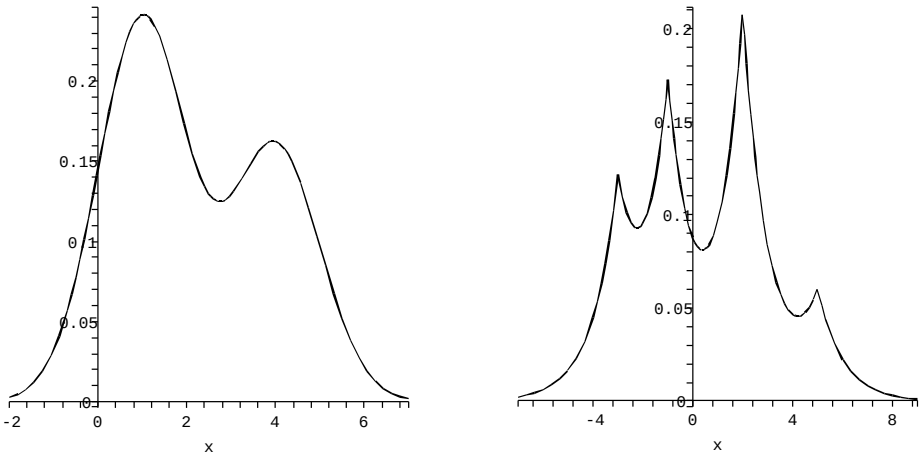
minimaal wordt. Dit ziet men (voor oneven n) als volgt in: Stel we hebben $x > \tilde{x}$, dan liggen er r waarden rechts van x en l waarden links van x en we hebben $l > r$. Als we nu x om Δx naar rechts schuiven, dan neemt $g(x)$ om $\Delta x(l - r)$ toe, als we x om Δx naar links schuiven, neemt $g(x)$ om $\Delta x(l - r)$ af. Dus is $g(x)$ niet minimaal als $l > r$ is. Met hetzelfde argument, toegepast of $x < \tilde{x}$, zien we dat $g(x)$ ook voor $l < r$ niet minimaal is. Dus moet $l = r$ gelden, en hieruit volgt $x = \tilde{x}$.

Voor even $n = 2m$ is $g(x)$ op het interval $[x_m, x_{m+1}]$ horizontaal met minimale waarde. Men neemt daarom het middelpunt van dit interval als mediaan.

De modus

Een verdere mogelijkheid om een typische waarde te definiëren is de *modus* $\hat{x}$ die de waarde met de hoogste frequentie is.

In veel gevallen geeft de modus een goede beschrijving die ook redelijk dicht bij het gemiddelde en de mediaan ligt, maar dit hangt sterk van de situatie af. Het kan bijvoorbeeld zijn, dat een verdeling twee duidelijke spitsen heeft, dan is de modus de hogere van de twee spitsen, maar gemiddelde en mediaan liggen waarschijnlijk tussen de spitsen. Een verdeling met twee spitsen heet *bimodaal*, een verdeling met nog meer spitsen *multimodaal*.



Figuur 6: Bimodale en multimodale verdelingen.

Het linkerplaatje in Figuur 6 laat een bimodale verdeling zien. De modus van deze verdeling is $\hat{x} = 1$, de mediaan is $\tilde{x} \approx 1.92$ en het gemiddelde is $\overline{x} = 2.2$.

In het rechterplaatje van Figuur 6 vinden we een multimodale verdeling met vier spitsen. In dit geval is de modus ook weer $\hat{x} = 1$, de mediaan is $\tilde{x} \approx 0.39$ en het gemiddelde is $\overline{x} = 0.4$.

Soms kan bij een multimodale verdeling de modus juist wel interessant zijn, maar vaak is het in dit geval nodig de verdeling als combinatie van een aantal unimodale verdelingen te beschrijven en door de typische waarden van deze verdelingen te karakteriseren.

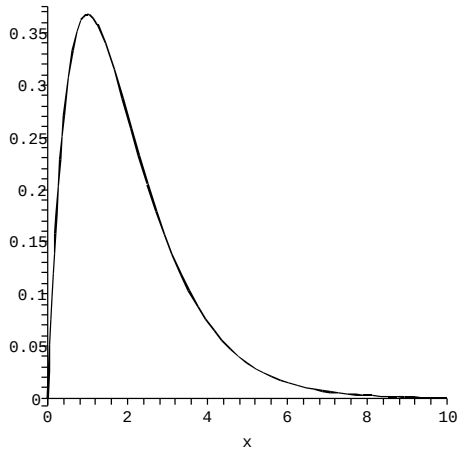
Samenhang tussen gemiddelde, mediaan en modus

Als een verzameling van gegevens een symmetrische unimodale verdeling heeft, vallen de waarden van het gemiddelde, de mediaan en de modus redelijk goed samen. Als de verdeling niet symmetrisch is en een langere staart naar rechts heeft, noemt men de verdeling *naar rechts scheef*. In dit geval is $\hat{x} < \tilde{x} < \overline{x}$. Omgekeerd heet een verdeling *naar links scheef* als hij een langere staart naar links heeft. In dit geval geldt $\overline{x} < \tilde{x} < \hat{x}$.

Een typische naar rechts scheve verdeling is

$$f(x) = \lambda^2 x e^{-\lambda x} \text{ met } \overline{x} = \frac{2}{\lambda}, \quad \tilde{x} \approx 1.678 \cdot \frac{1}{\lambda}, \quad \hat{x} = \frac{1}{\lambda}.$$

Deze verdeling is in Figuur 7 voor de parameter $\lambda = 1$ te zien. In het plaatje ligt dus de modus bij $\hat{x} = 1$, de mediaan bij $\tilde{x} \approx 1.678$ en het gemiddelde bij $\overline{x} = 2$.



Figuur 7: Naar rechts scheve verdeling $f(x) = xe^{-x}$.

Omdat de modus vaak niet eenvoudig te berekenen valt, wordt er voor unimodale verdelingen soms een heuristische formule voor de samenhang tussen modus, mediaan en gemiddelde toegepast, namelijk

$$\overline{x} - \hat{x} = 3(\overline{x} - \tilde{x}).$$

Voor de boven aangegeven verdeling $f(x) = \lambda^2 x e^{-\lambda x}$ zien we dat dit een uitstekende vuistregel is, maar let wel dat dit bij multimodale verdelingen meestal vreselijk mis gaat (zie de voorbeelden in Figuur 6).

Merk op: Het gemiddelde is veel gevoeliger voor *uitschieters* dan de mediaan. Op de modus heeft een uitschieter helemaal geen invloed. Als het erom gaat een robuuste schatting voor de typische waarde te hebben en er gevaar op uitschieters bestaat, is de mediaan soms een betere keuze dan het gemiddelde.

In ons voorbeeld van de tentamen resultaten kunnen we het gemiddelde en de mediaan makkelijk bepalen, we hebben $\overline{x} = 48.56$ en $\tilde{x} = 47.5$. Voor

de modus moeten we naar klassen kijken, als we bijvoorbeeld als klassen de intervallen van breedte 10 nemen, ligt de modus in het interval $[40, 50]$ en men neemt hiervoor de middelste waarde van het interval, dus $\hat{x} = 45$. Als we nu de uitslag van 80 punten als uitschieter beschouwen en weglaten, verandert dit het gemiddelde behoorlijk, we krijgen dan als nieuwe gemiddelde $\overline{x} = 46.47$, terwijl de mediaan veel minder verandert en nu $\tilde{x} = 47$ wordt. De modus blijft onveranderd.

> We kunnen zelfs algemeen aangeven hoe veel het weglaten van een waarde het gemiddelde verandert. Stel we hebben bij n waarden en gemiddelde $\overline{x}$ en willen de waarde x weglaten. Het nieuwe gemiddelde wordt dan $\frac{n\cdot\overline{x}-x}{n-1}$ en voor het verschil van het oude en het nieuwe gemiddelde krijgen we:
>
> $$\overline{x} - \frac{n\cdot\overline{x}-x}{n-1} = \frac{(n-1)\cdot\overline{x}-n\cdot\overline{x}+x}{n-1} = \frac{x-\overline{x}}{n-1}.$$
>
> Het gemiddelde verandert dus om de afstand van de uitschieter van het gemiddelde, gedeeld door $n-1$.

Andere gemiddelden

Soms is het rekenkundig gemiddelde niet geschikt om een gelijkmatige herverdeling te beschrijven. Dit is bijvoorbeeld het geval als de gegevens x_i een variabel beschrijven die niet opgeteld maar vermenigvuldigd wordt, zoals bij groeiprocessen. Stel een populatie groeit in n jaren met factoren $x_1, x_2, \ldots, x_n$, dan is de totale groei het product $\prod_{i=1}^{n} x_i$ van de x_i. Om nu een gemiddelde groei te berekenen, waarmee in n jaren dezelfde totale groei bereikt wordt, moeten we een waarde x_0 vinden zo dat $x_0^n = \prod_{i=1}^{n} x_i$. We moeten dus uit het product de n-de wortel trekken, dit geeft

$$x_0 = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

en x_0 heet het *meetkundig gemiddelde* van de x_i.

Een andere vorm van gemiddelde bestaat bij gegevens waarvoor eigenlijk x_i^{-1} opgeteld moet worden. Een beroemd voorbeeld hiervoor is het probleem van de piloot die op de heenweg wind tegen heeft maar de vertraging op de terugweg door de wind mee weer in te halen denkt. We noemen de afstand van de twee vliegvelden s, de tijd voor de heenweg t_1 en de tijd voor de terugweg t_2. Als de piloot zonder wind met een snelheid van v_0 vliegt, zou hij zonder wind de tijd $t = 2\frac{s}{v}$ nodig hebben. Bij wind met snelheid w is de snelheid op de heenweg $v_1 = v_0 - w$ en op de terugweg $v_2 = v_0 + w$. De tijden voor heen- en terugweg zijn $t_1 = \frac{s}{v_1}$ en $t_2 = \frac{s}{v_2}$. De vraag is nu, of $t_1 + t_2$ gelijk aan t is. Voor de gemiddelde snelheid $v = \frac{2s}{t_1+t_2}$ geldt:

$$v = \frac{2s}{t_1+t_2} = \frac{2s}{\frac{s}{v_1}+\frac{s}{v_2}} = \frac{2}{\frac{1}{v_1}+\frac{1}{v_2}} = \frac{2v_1v_2}{v_1+v_2} \text{ en dus } \frac{1}{v} = \frac{\frac{1}{v_1}+\frac{1}{v_2}}{2}.$$

Men noemt $v = \frac{2v_1v_2}{v_1+v_2}$ het *harmonisch gemiddelde* van v_1 en v_2 en dit is gewoon het inverse van het rekenkundig gemiddelde van de inversen van v_1 en v_2. In het geval met $v_1 = v_0 - w$ en $v_2 = v_0 + w$ hebben we

$$v = \frac{2(v_0 - w)(v_0 + w)}{(v_0 - w) + (v_0 + w)} = \frac{2(v_0^2 - w^2)}{2v_0} = \frac{v_0^2 - w^2}{v_0} < v_0.$$

De vliegreis duurt dus inderdaad langer.

Tussen de verschillende gemiddelden bestaat altijd de volgende keten van ongelijkheden:

$$\text{minimum} \leq \text{harmonisch} \leq \text{meetkundig} \leq \text{rekenkundig} \leq \text{maximum}\,.$$

1.4 Spreiding

Het is duidelijk dat een verzameling gegevens met een gemiddelde waarde (of zelfs de verschillende soorten van gemiddelden) nog niet goed beschreven is, want de verdelingen kunnen er nog erg verschillend uit zien. Bijvoorbeeld kan het zijn dat bij een tentamen met een gemiddelde van 7 iedereen het gehaald heeft, omdat er even veel 6en als 8en en geen 9en en 10en waren. Maar het kan ook zijn, dat slechts 40% het gehaald hebben, omdat 40% een 10 en 60% en 5 gehaald hebben (dit is een typisch voorbeeld van een bimodale verdeling). Men wil daarom ook een uitspraak over de afwijking van de waarden van het gemiddelde hebben. Ook hiervoor zijn er verschillende mogelijkheden.

Standaardafwijking

We hebben al gezien dat het gemiddelde $\overline{x}$ de waarde is waarvoor de kwadratische afstanden van de gegevens minimaal is. De wortel uit dit minimum heet de *standaardafwijking* s, we hebben dus

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Voor veel (en belangrijke) verdelingen ligt een 'groot deel' van de waarden binnen een afstand van s van het gemiddelde. Voor de normale verdeling zijn dit bijvoorbeeld 68% (en 95% liggen binnen een afstand van $2s$). Met behulp van het gemiddelde en de standaardafwijking laten zich gegevens normaliseren: De verschuiving $x_i' := x_i - \overline{x}$ geeft een verzameling gegevens met gemiddelde 0 en $z_i := \frac{x_i - \overline{x}}{s}$ geeft een verzameling gegevens met standaardafwijking 1. Men noemt de waarde

$$z := \frac{x_i - \overline{x}}{s}$$

ook de z-waarde van x_i. De z-waarde geeft de afwijking van een waarde van het gemiddelde in veelvouden van de standaardafwijking aan. Men zegt daarom ook soms dat een waarde *een afstand van 3 standaardafwijkingen* heeft, als de z-waarde 3 is.

Als we de standaardafwijking weer voor waarden bekijken die volgens een kansverdeling voor een stochast X geproduceerd zijn, zien we dat s^2 een benadering van de variantie $Var(X) = E[(X - E[X])^2]$ is. Voor een discrete kansverdeling is deze gegeven door $Var(X) = \sum_x (x - E[X])^2 \cdot p_x$, en voor een continue kansverdeling met dichtheidsfunctie $f(x)$ door $Var(X) = \int_{-\infty}^{\infty} (x - E[X])^2 \cdot f(x)\, dx$.

> In de kansrekening hebben we de wortel uit de variantie ook de *standaardafwijking* genoemd en toen met σ genoteerd. Het is inderdaad gebruikelijk, grootheden van kansverdelingen zo als verwachtingswaarde en standaardafwijking met *griekse letters* (μ, σ) te noteren, terwijl grootheden bij verdelingen van gegevens met *latijnse letters* genoteerd worden. Let wel dat niet iedere auteur dit soort conventies behartigt.

Kwartielen

Net als de mediaan voor de helft van de waarden worden ook *kwartielen* gedefinieerd waar een kwart van de waarden beneden of boven ligt. Het *onderste kwartiel* of *eerste kwartiel* is de waarde waar een kwart van de waarden onder en drie kwart boven liggen en is dus de mediaan van de onderste helft van de waarden. Net zo is het *bovenste kwartiel* of *derde kwartiel* de waarde waar drie kwart onder en een kwart boven ligt, dus de mediaan van de bovenste helft van de waarden. De mediaan zelfs heet soms ook het *tweede kwartiel*.

Algemeen noemt men de waarde waar p procent van de waarden onder en $100 - p$ procent boven liggen het *p-percentielpunt* en noteert dit met P_p. De mediaan is dus het 50-percentielpunt P_{50}, het onderste kwartiel het 25-percentielpunt P_{25} en het bovenste kwartiel het 75-percentielpunt P_{75}. Meestal zal een p-percentielpunt niet precies op een waarde vallen, en ook niet op het middelpunt tussen twee waarden. Bij n (geordende) waarden heeft het p-percentielpunt in de lijst de index $t = 1 + \frac{p}{100}(n-1)$. Als we t schrijven als $i + r$ met i een natuurlijk getal en $0 \leq r < 1$, dan berekenen we de waarde voor het p-percentielpunt als gewogen gemiddelde van x_i en x_{i+1} met gewichten $(1 - r)$ en r, dus als

$$P_p = (1 - r) \cdot x_i + r \cdot x_{i+1}.$$

Als we in ons voorbeeld van 16 waarden het 15-percentielpunt zouden willen vinden, hebben we $t = 1 + \frac{15}{100} \cdot 15 = 1 + \frac{225}{100} = 3 + \frac{1}{4}$. Het 15-percentielpunt ligt dus tussen x_3 en x_4, maar op een vierde van de afstand van x_3 naar x_4. We zouden dus in dit geval het 15-percentielpunt berekenen door $0.75 \cdot x_3 + 0.25 \cdot x_4$.

> Percentielpunten worden ook gebruikt om parameters van systemen vast te leggen. Bijvoorbeeld geeft een spraakherkenningssysteem voor elke herkenning een *score* die aangeeft hoe goed de kwaliteit van de herkenning was. Dit geeft in het algemeen niet de kans op een correcte herkenning weer, maar slechts een heuristische waarde die met toenemende kwaliteit stijgt. Als men met het automatische systeem nu 90% van de aanvragen wil behandelen en de rest naar een menselijke operator doorstuurt, dan moet men op een testset van aanvragen

het 90-percentielpunt van de scores bepalen en dit als grens vastleggen waaronder aanvragen naar de operator doorgestuurd worden.

De afstand tussen de kwartielen geeft informatie over de spreiding van de waarden. Het interval tussen de kwartielen P_{25} en P_{75} heet het *interkwartielbereik*, hun verschil de *interkwartielafstand* IQR (voor *inter quartile range*). Vaak wordt ook de helft van de interkwartielafstand gebruikt, de *semi-interkwartielafstand* $\frac{1}{2}IQR := \frac{P_{75}-P_{25}}{2}$.

Er is geen zuivere definitie mogelijk wanneer een waarde die uit het algemene patroon valt een uitschieter is, over dit probleem zijn veel boeken geschreven. Een veel gehanteerde vuistregel is, waarden als uitschieters te beschouwen die meer dan $1.5 \cdot IQR$ buiten het interkwartielbereik liggen, dus:

$$x < P_{25} - 1.5 \cdot IQR \text{ of } x > P_{75} + 1.5 \cdot IQR \Rightarrow x \text{ is een uitschieter.}$$

Voor waarden die volgens dit criterium uitschieters zijn moet men *met de hand* beslissen of het gewoon extreme maar geldige waarden zijn of ongeldige waarden die uit het bestand verwijderd moeten worden (bijvoorbeeld omdat er bij een meeting iets is mis gegaan).

Voor verdelingen die niet erg scheef zijn, bestaat er een verband tussen de standaardafwijking s en die semi-interkwartielafstand $\frac{1}{2}IQR$, namelijk

$$\frac{1}{2}IQR \approx \frac{2}{3}s.$$

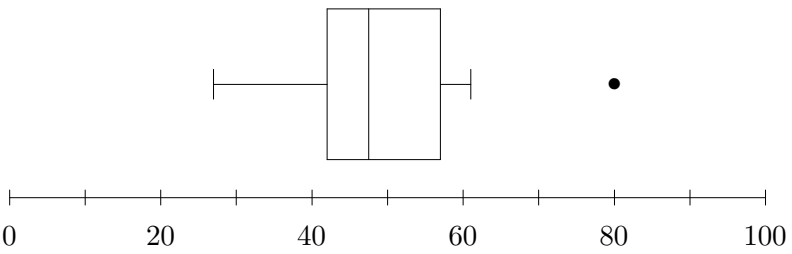
Dit is afgeleid van de normale verdeling, waarvoor $\frac{1}{2}IQR \approx 0.6745$ geldt.

Natuurlijk leveren naast de kwartielen ook de minimale en de maximale waarde informatie over de spreiding van een verdeling. Dit soort informatie wordt vaak in een *doos-en-snorren* figuur (*box-and-whiskers plot* of kort *box-plot*) samengevat. Dit is een doos tussen de kwartielen met de mediaan gemarkeerd. Voor de einden van de snorren zijn er verschillende conventies:

- minimale en maximale waarden;
- minimale en maximale waarden die binnen een afstand van $1.5 \cdot IQR$ van de kwartielen liggen, de andere waarden worden als uitschieters beschouwd (en soms wel als punten weergegeven);
- 5-percentielpunt en 95-percentielpunt.

In ons voorbeeld van de tentamenresultaten hebben we $P_{50} = 47.5$, $P_{25} = 42$ en $P_{75} = 57$. Hieruit volgt $IQR = 15$. Omdat $42 - 1.5 \cdot 15 = 19.5$ kleiner is dan alle waarden, hebben volgens het genoemde criterium geen uitschieters naar beneden. Aan de andere kant is $57 + 1.5 \cdot 15 = 79.5$, dus is de waarde 80 net een uitschieter.

De doos-en-snorren figuur voor het voorbeeld ziet er dus als volgt uit:



De doos-en-snorren figuur wordt soms horizontaal (zo als hier) en soms verticaal getekend. De verticale versie heeft het voordeel dat de figuren voor verschillende verdelingen makkelijk naast elkaar geplaatst kunnen worden.

1.5 Momenten

We hebben al een paar keer iets over de scheefheid van een verdeling gezegd. Natuurlijk laat zich dit aan de hand van een grafiek meestal goed aflezen, maar het is handig hiervoor ook een kwantitatief begrip te hebben. Hiervoor zijn de *momenten* van een verdeling handig. Het k-de *moment* van een verzameling gegevens is

$$m'_k := \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

en het k-de *centrale moment* rond het gemiddelde is

$$m_k := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k.$$

De eerste en tweede momenten zijn oude bekenden, we hebben $\overline{x} = m'_1$, $m_1 = 0$ en $s = \sqrt{m_2}$.

Om momenten voor verschillende verdelingen goed te kunnen vergelijken, is het handig om ze te normaliseren. Dit gebeurt net als bij de z-waarde door delen door de standaardafwijking en men krijgt

$$a_k := \frac{m_k}{s^k} = \frac{m_k}{\sqrt{m_2}^k}.$$

Momenten worden op een analoge manier ook voor kansverdelingen gedefinieerd. Voor een stochast X met een discrete kansverdeling met kansen p_x zijn de k-de momenten μ'_k en de k-de centrale momenten μ_k gedefinieerd door

$$\mu'_k := \sum_x x^k \cdot p_x \text{ en } \mu_k := \sum_x (x - E[X])^k \cdot p_x.$$

Voor een stochast X met een continue kansverdeling met dichtheidsfunctie $f(x)$ geldt

$$\mu'_k := \int_{-\infty}^{\infty} x^k \cdot f(x)\, dx \text{ en } \mu_k := \int_{-\infty}^{\infty} (x - E[X])^k \cdot f(x)\, dx.$$

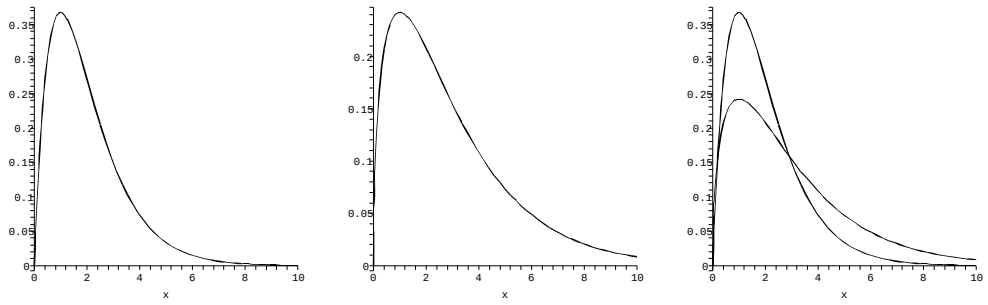
In het bijzonder is $\mu'_1 = E[X]$ en $\mu_2 = Var(X)$.

Merk op dat hogere momenten niet voor alle verdelingsfuncties van continue kansverdelingen hoeven te bestaan. Zo heeft bijvoorbeeld de integraal $\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx$ de waarde π, maar de integralen $\int_{-\infty}^{\infty} x^2 \cdot \frac{1}{1+x^2}\,dx$ en $\int_{-\infty}^{\infty} x^4 \cdot \frac{1}{1+x^2}\,dx$ hebben geen eindige waarde.

Scheefheid

Omdat voor een scheve verdeling de waarden in de langere staart een hoger gewicht krijgen, is het derde centrale moment een maat voor de scheefheid (*skewness*) van de verdeling. Bij positieve waarden van m_3 of a_3 is de verdeling scheef naar rechts, bij negatieve waarden scheef naar links. Men noemt a_3 ook de *coëfficiënt van scheefheid*. Verdelingen die symmetrisch ten opzichte van hun gemiddelde zijn (zo als de normale verdeling), hebben natuurlijk scheefheid 0.

In Figuur 8 zijn de grafieken van twee naar rechts scheve verdelingen te zien. De functie in het linkerplaatje is $f(x) := \lambda^2 x \cdot e^{-\lambda x}$ (voor $\lambda = 1$), de functie in het middelste plaatje is $g(x) := \frac{1}{\sqrt{2\pi}}\sqrt{x} \cdot e^{-\frac{x}{2}}$.



Figuur 8: Verdelingsfuncties van twee haar rechts scheve verdelingen.

De momenten voor $f(x)$ zijn $\overline{x} = m_1' = \frac{2}{\lambda}$, $s^2 = m_2 = \frac{2}{\lambda^2}$ en $m_3 = \frac{4}{\lambda^3}$. Hieruit volgt dat de coëfficiënt van scheefheid $a_3 = \frac{m_3}{\sqrt{m_2}^3} = \sqrt{2} \approx 1.414$ is. Merk op dat a_3 onafhankelijk van de parameter λ is.

De momenten voor $g(x)$ zijn $\overline{x} = m_1' = 3$, $s^2 = m_2 = 6$ en $m_3 = 24$. Hieruit volgt dat $g(x)$ de coëfficiënt van scheefheid $a_3 = \frac{m_3}{\sqrt{m_2}^3} = \frac{2}{3}\sqrt{6} \approx 1.633$ heeft. Zo als ook uit de plaatjes blijkt, heeft $g(x)$ een grotere scheefheid dan $f(x)$.

Een andere mogelijkheid om de scheefheid aan te geven gebruiken het verschil van gemiddelde en modus, bijvoorbeeld $\frac{\overline{x}-\hat{x}}{s}$. Als we hierin de heuristische benadering $\overline{x} - \hat{x} = (\overline{x} - \tilde{x})$ voor de modus toepassen, krijgen we $\frac{3(\overline{x}-\tilde{x})}{s}$ als uitdrukking voor de scheefheid.

Ook met behulp van de kwartielen of percentielen laat zich de scheefheid uitdrukken, bijvoorbeeld door

$$\frac{(P_{75} - \tilde{x}) - (\tilde{x} - P_{25})}{P_{75} - P_{25}} = \frac{P_{75} - 2\tilde{x} + P_{25}}{P_{75} - P_{25}} \quad \text{of}$$

$$\frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}.$$

Hierbij wordt gekeken hoe ver de p-percentielpunten P_{50-x} en P_{50+x} die bij een symmetrische verdeling even grote afstanden van de mediaan hebben van een symmetrische positie afwijken.

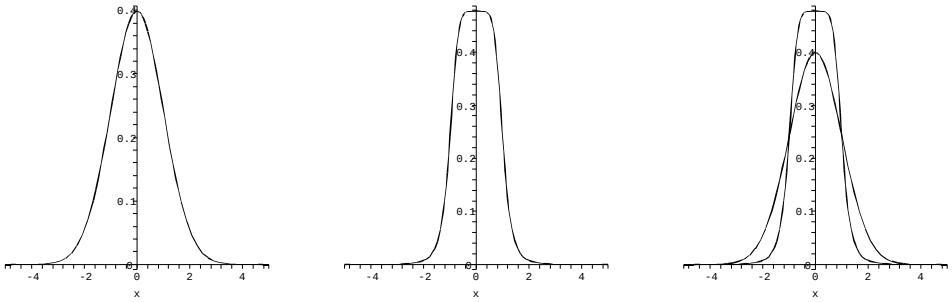
Scherptoppigheid

Het vierde moment zegt iets erover of een verdeling spits of plat is, dus over de *scherptoppigheid* of *gepiekdheid* (*kurtosis*) van de verdeling. Hiervoor vergelijkt men het genormaliseerde vierde moment a_4 met het vierde moment van de standaard-normale verdeling dat de waarde 3 heeft en noemt a_4 ook de *coëfficiënt van scherptoppigheid*. Voor $a_4 > 3$ noemt men een verdeling *gepiekd* (*leptokurtic*, van het griekse *lepto-* = smal) omdat de verdeling dan een scherpere top heeft dan de normale verdeling en de staarten dunner zijn. Voor $a_4 < 3$ noemt men de verdeling *afgeplat* (*platykurtic*, van *platy-* = plat) omdat ze een plattere top heeft dan de normale verdeling. Een verdeling met $a_4 \approx 3$ heet *mesokurtic* (van *meso-* = gemiddeld).

Merk op: In de literatuur wordt vaak ook $a_4 - 3$ als coëfficiënt van scherptoppigheid gehanteerd, een positieve waarde hiervan staat dan voor een gepiekde verdeling, een negatieve waarde voor een afgeplatte verdeling.

Als eenvoudig voorbeeld bekijken we de symmetrische uniforme verdeling op het interval $[-c, c]$, deze heeft de dichtheidsfunctie $f(x) = \frac{1}{2c}$. Er geldt $m_2 = \int_{-c}^{c} x^2 \cdot \frac{1}{2c}\, dx = \frac{1}{2c} \cdot \frac{x^3}{3} \,|_{-c}^{c} = \frac{1}{3}c^2$ en $m_4 = \int_{-c}^{c} x^4 \cdot \frac{1}{2c}\, dx = \frac{1}{2c} \cdot \frac{x^5}{5} \,|_{-c}^{c} = \frac{1}{5}c^4$. Hieruit volgt $a_4 = \frac{m_4}{m_2^2} = \frac{9}{5} < 3$, dus is de uniforme verdeling afgeplat. Merk op dat de schalingsfactor c geen invloed op de scherptoppigheid van de verdeling heeft.

Een interessanter voorbeeld is de verdeling met dichtheidsfunctie $f(x) = \frac{3}{2\pi} \cdot \frac{1}{1+x^6}$ die in het middelste plaatje van Figuur 9 te zien is. Hier hebben we $m_2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx = \frac{1}{2}$ en $m_4 = \int_{-\infty}^{\infty} x^4 \cdot f(x)\, dx = 1$, dus is $a_4 = \frac{m_4}{m_2^2} = 4$ en $f(x)$ is een gepiekde verdeling. Dit wordt ook in het vergelijk met de normale verdeling in Figuur 9 duidelijk.



Figuur 9: Verdelingsfuncties voor de normale verdeling en een gepiekde verdeling.

Merk op dat de scherptoppigheid vooral bij (redelijk) symmetrische verdelingen een rol speelt. Bij scheve verdelingen heeft de scheefheid een groot invloed

op de coëfficiënt van scherptoppigheid en is het vergelijken met symmetrische verdelingen meestal niet bijzonder verklarend.

Belangrijke begrippen in deze les

- stengel-en-blad diagram
- klassen, frequentieverdeling
- histogram, taart-diagram
- gemiddelde, mediaan, modus
- uni-, bi-, multimodale verdelingen
- kwartielen, p-percentielpunten
- standaardafwijking, interkwartielafstand
- doos-en-snorren figuur
- momenten, scheefheid, scherptoppigheid

Opgaven

1. Gegeven is de rij waarnemingen

$$15.813, 15.705, 15.748, 15.801, 15.720, 15.743.$$

   Bereken het gemiddelde en de standaardafwijking van deze gegevens

   (i) zonder af te ronden;

   (ii) met op twee decimalen achter de komma afgeronde waarden;

   (ii) met op een decimaal achter de komma afgeronde waarden.

2. Dit is een standaardafwijkings-wedstrijd: Kies als gegevens 4 getallen uit de getallen $0, 1, \ldots, 10$, waarbij herhalingen toegestaan zijn.

   (i) Vind getallen zo dat hun standaardafwijking minimaal is. Is het antwoord eenduidig?

   (ii) Vind getallen zo dat hun standaardafwijking maximaal is. Is het antwoord eenduidig?

   (iii) Behandel (i) en (ii) met 3 in plaats van 4 getallen.

3. Zij X het aantal ogen dat geworpen wordt met twee witte en één zwarte dobbelsteen, waarbij het aantal ogen van de zwarte dobbelsteen dubbel wordt geteld. In een experiment met 50 werpen zijn de volgende resultaten verkregen:

12	10	23	10	10	14	15	20	5	18
14	8	6	20	21	12	16	11	13	21
13	10	9	16	19	7	9	7	20	22
17	14	15	15	12	9	13	14	18	8
17	18	15	12	14	20	18	11	19	7

(i) Bereken de verwachtingswaarde $E[X]$ en de variantie $Var(X)$ van de stochast X (dit hangt niet van de verkregen resultaten af).

(ii) Bereken het gemiddelde $\overline{x}$ en de standaardafwijking s van de 50 waarnemingen.

(iii) Maak een histogram voor een zinvolle indeling van de waarnemingen in klassen.

4. De aantallen van stemmen voor de kandidaat presidenten in de VS in de verkiezingen sinds 1960 (dus sinds Kennedy) waren:

jaar	Republicans	Democrats	anderen
1960	34,108,157	34,226,731	0
1964	27,178,188	43,129,484	0
1968	31,785,480	31,275,166	9,906,473
1972	47,169,911	29,170,383	1,099,482
1976	39,147,973	40,830,763	756,631
1980	43,899,248	36,481,435	5,719,437
1984	54,455,075	37,577,185	0
1988	48,886,097	41,809,074	0
1992	39,104,545	44,909,889	19,742,267
1996	39,198,755	47,402,357	8,085,402
2000	50,456,002	50,999,897	2,882,955
2004	59,668,261	56,172,264	0

Met uitzondering van de verkiezingen in 2000 is steeds de kandidaat met de meeste stemmen president geworden.

(i) Maak frequentiepolygonen voor de relatieve aantallen stemmen voor de verschillende partijen.

(ii) Bepaal de verdeling van de stemaandelen die de gekozen president in de verschillende verkiezingen heeft behaald. Maak een doos-en-snorren figuur voor deze verdeling. Zijn er uitschieters? Kun je dit verklaren?

(iii) We beperken ons nu tot de stemmen voor de republikanen en de demokraten. In het jaar 2000 heeft dan bijvoorbeeld de kandidaat van de republikanen $50,456,002$ van $50,456,002+50,999,897 = 101,455,899$ stemmen, dus 49.73% van deze stemmen gehaald, en de kandidaat van de demokraten 50.27%. De *afstand* tussen republikanen en demokraten definiëren we als het verschil van deze aandelen, dus -0.54% voor het jaar 2000 (let op het teken).

Bepaal de verdeling van deze afstanden, hun gemiddelde, standaardafwijking, mediaan, kwartielen en interkwartielafstand.

Men zegt dat er een *aardverschuiving* heeft plaatsgevonden als de afstand bij een verkiezing sterk verschilt van de afstand bij de vorige verkiezing. Definieer een criterium, wanneer er sprake van een aardverschuiving is en geef aan bij welke verkiezingen een aardverschuiving heeft plaatsgevonden.

5. Zij $x_1, \ldots, x_n$ een verzameling gegevens waarbij de x_i alleen maar de waarden 0 of 1 kunnen hebben. Stel er zijn $p \cdot n$ gegevens met de waarde 0 en $(1-p) \cdot n$ gegevens met de waarde 1.

(i) Bereken het gemiddelde $\overline{x}$ en de centrale momenten m_k voor $k = 1, 2, 3, 4$.

(ii) Geef de scheefheid en scherptoppigheid van deze verzameling gegevens aan.

(iii) Laat zien dat de scheefheid 0 is dan en slechts dan als $p = 0.5$, dus als de verdeling over de twee mogelijke waarden symmetrisch is.

Les 2 Steekproeven en schatters

We zullen in deze les bekijken, hoe we gegevens van een populatie zo als het gemiddelde en de spreiding kunnen schatten, zonder naar elk individu van de populatie te kijken. Het idee hierbij is, alleen maar een deel van de populatie te pakken (dit noemen we een steekproef), dit als representatief te beschouwen en de gegevens hierop te bepalen. Een belangrijke vraag is dan hoe dicht de schatting bij de ware waarde zou liggen en wat voor een afwijking we moeten verwachten.

Voor dat we ons hiermee gaan bemoien moeten we een aantal feiten over de normale verdeling verzamelen (herhalen), omdat deze verdeling de basis voor de analyse van steekproeven vormt.

2.1 De normale verdeling

De belangrijkste verdeling in de statistiek is de *normale verdeling*. Deze wordt volledig bepaald door de verwachtingswaarde μ en de variantie σ^2 (of de standaardafwijking σ) en heeft de dichtheidsfunctie

$$f_{\mu,\sigma}(x) := \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Een stochast X die een kansverdeling met deze dichtheidsfunctie heeft, heet *normaal verdeeld* en wordt vaak met $X \in \mathcal{N}(\mu, \sigma^2)$ genoteerd. De verdelingsfunctie voor een normaal verdeelde stochast kan niet zonder integraal geschreven worden, er geldt

$$F(x) := P(X \leq x) = \int_{-\infty}^{x} f_{\mu,\sigma}(t)\, dt.$$

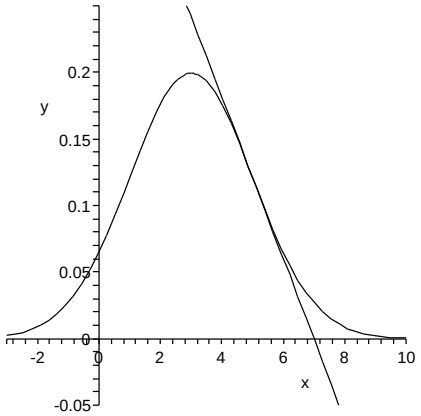
Voor een normaal verdeelde stochast X met verwachtingswaarde μ en variantie σ^2 heeft de *genormaliseerde stochast* $Z := \frac{X-\mu}{\sigma}$ de verwachtingswaarde 0 en variantie 1 en men noemt de stochast Z een *standaard-normaal verdeelde stochast*. De standaard-normale verdeling heeft de eenvoudigere dichtheidsfunctie

$$f(x) := f_{0,1}(x) := \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}.$$

De parameters μ en σ van een normale verdeling kunnen aan de grafiek van de dichtheidsfunctie $f(x)$ afgelezen worden zo als in Figuur 10 te zien.

De verwachtingswaarde μ is gewoon het punt waar $f(x)$ zijn maximum heeft en omdat de normale verdeling symmetrisch is, is dit ook de mediaan en de modus van de kansverdeling. De standaardafwijking σ vinden we op basis van het feit dat de punten $x = \mu - \sigma$ en $x = \mu + \sigma$ juist de punten zijn waar de grafiek van kromming verandert. Op de punten waar en grafiek van kromming verandert is de stijging van de grafiek maximaal of minimaal en heeft de afgeleide van de functie dus een maximum of minimum (en dus de tweede afgeleide een nulpunt).

Omdat de verdelingsfunctie $F(x)$ van de normale verdeling niet makkelijk te berekenen is, worden de waarden vaak in tabellen aangegeven. Hierbij is het



Figuur 10: Normale verdeling met $\mu = 3$ en $\sigma = 2$ en raaklijn aan de grafiek in $x = \mu + \sigma$.

voldoende, de waarden voor de standaard-normale verdeling aan te geven, voor een willekeurige normale verdeling worden de waarden op de z-waarden van de standaard-normale verdeling genormaliseerd. Voor $z = \frac{x-\mu}{\sigma}$ en $Z = \frac{X-\mu}{\sigma}$ geldt immers:

$$P(X \leq x) = P(Z \leq z) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t}\, dt.$$

De tabellen voor de standaard-normale verdeling worden op twee manieren aangelegd:

(i) De waarden $P(Z \leq z)$ voor waarden van z in regelmatige afstanden, bijvoorbeeld afstanden van 0.05 tussen $z = -3$ en $z = 3$.

(ii) Kritieke waarden van z zo dat $P(Z \leq z) = p$ voor zekere kansen p, bijvoorbeeld kansen in afstanden van 0.01 tussen 0 en 1.

Voorbeeld: Voor een normaal verdeelde stochast X met verwachtingswaarde 3 en standaardafwijking 2 willen we de kans $P(1 \leq X \leq 4)$ weten, dat een waarde tussen $x_1 = 1$ en $x_2 = 4$ ligt. De genormaliseerde z-waarden zijn $z_1 = \frac{x_1-3}{2} = -1$ en $z_2 = \frac{x_2-3}{2} = 0.5$. De gezochte kans is dus $P(Z \leq 0.5) - P(Z \leq -1)$ voor de standaard-normaal verdeelde stochast Z. Voor deze twee kansen vinden we in een tabel de waarden $P(Z \leq 0.5) \approx 0.6915$ en $P(Z \leq -1) \approx 0.1587$. De gezochte kans is dus $0.6915 - 0.1587 = 0.5328$.

Als we omgekeerd willen weten voor welke waarde van x de kans $P(X \leq x) = 0.8$ is, vinden we in een tabel dat dit voor de z-waarde 0.8416 het geval is, dus voor $x = \sigma \cdot z + \mu = 2 \cdot 0.8416 + 3 = 4.6832$.

De redenen voor de centrale stelling van de normale verdeling in de statistiek zijn veelvoudig, de volgende opmerkingen geven hier een idee van:

(1) Voor zekere parameters worden andere kansverdelingen zo als de binomiale verdeling of de Poisson-verdeling door de normale verdeling goed benadert.

(2) De combinatie van een groot aantal resultaten met bijna willekeurige kansverdelingen wordt goed benaderd door een normale verdeling.

(3) De frequentieverdelingen van de uitkomsten van veel experimenten worden goed benaderd door een normale verdeling, bijvoorbeeld merkmalen van populaties (grootte, gewicht), herhaald meten van gegevens, resultaten van een grote groep mensen bij een test, enz. Dit is ten dele een consequentie uit het punt (2), want vaak is een grootheid bepaald door een aantal enigszins onafhankelijke factoren en de combinatie daarvan geeft een normale verdeling.

Normale benadering van andere kansverdelingen

Stel een toevalsexperiment levert met kans p een succes op, dan heeft de stochast X die het aantal successen in n pogingen telt een *binomiale verdeling* en er geldt

$$P(X = k) = b(n, p; k) = \binom{n}{k} p^k (1-p)^k.$$

Een binomiaal verdeelde stochast X heeft de verwachtingswaarde $E[X] = np$ en de variantie $Var(X) = np(1-p)$. We transformeren X met behulp van $E[X]$ en $Var(X)$ op een stochast Z die verwachtingswaarde 0 en variantie (of standaardafwijking) 1 heeft door

$$Z := \frac{X - np}{\sqrt{np(1-p)}}$$

te definiëren. Als we n laten groeien, maakt de *stelling van De Moivre en Laplace* een belangrijke uitspraak over de stochast Z:

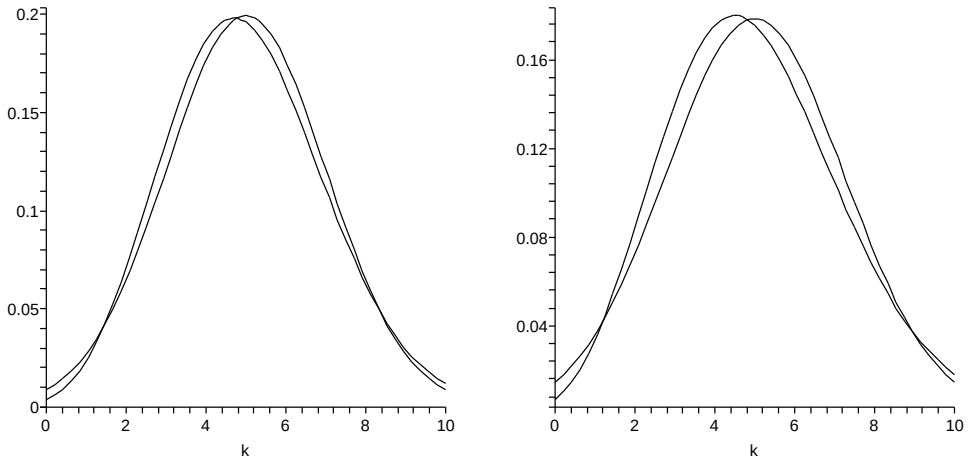
Stelling: De limiet $\lim_{n\to\infty} \dfrac{X - np}{\sqrt{np(1-p)}}$ is een standaard-normaal verdeelde stochast.

Omgekeerd betekent dit, dat voor niet te kleine waarden van n de binomiale verdeling met parameters n en p door de normale verdeling met parameters $\mu = np$ en $\sigma^2 = np(1-p)$ benaderd kan worden. We noemen dit de *normale benadering* van de binomiale verdeling.

De benadering is beter als p in de buurt van $\frac{1}{2}$ ligt en slechter als p dicht bij 0 of 1 ligt. Als vuistregel wordt vaak gehanteerd, dat de normale benadering van de binomiale verdeling toegestaan is als $np \geq 5$ en $n(1-p) \geq 5$ (soms wordt ook $np \geq 10$ en $n(1-p) \geq 10$ geëist).

We weten dat we voor een stochast X van zeldzame gebeurtenissen (dus met kleine p) de binomiale verdeling door de Poisson-verdeling met parameter $\lambda = np$ kunnen benaderen. Voor de kansen bij de Poisson-verdeling geldt

$$P(X = k) = po_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$



Figuur 11: Normale benadering van de binomiale verdeling met parameters $n = 25$ en $p = 0.2$ (links) en van de Poisson-verdeling met parameter $\lambda = 5$ (rechts).

en de stochast X heeft verwachtingswaarde $E[X] = \lambda$ en variantie $Var(X) = \lambda$.

Nadat we de binomiale verdeling behandeld hebben, is het nu geen verrassing, dat ook de Poisson-verdeling door de normale verdeling benaderd kan worden, als de parameter λ niet te klein is. Men noemt de normale verdeling met $\mu = \lambda$ en $\sigma^2 = \lambda$ de *normale benadering* van de Poisson-verdeling met parameter λ. Analoog met de binomiale verdeling wordt ook hier meestal $\lambda \geq 5$ als vuistregel gehanteerd.

Dat de benaderingen voor de aangegeven grenzen inderdaad redelijk goed zijn, kunnen we aan de voorbeelden in Figuur 11 zien. Merk op dat de binomiale verdeling en de Poisson-verdeling scheef naar rechts zijn. Daarom ligt de modus van de twee in Figuur 11 aangegeven verdelingen links van 5 (bij 4.69 voor de binomiale verdeling en bij 4.49 voor de Poisson-verdeling) en is de normale verdeling dus telkens de verdeling met het maximum meer rechts.

Centrale limietstelling

Dat de combinatie van min of meer willekeurige kansverdelingen door een normale verdeling benadert wordt, is ruwweg de uitspraak van een van de meest belangrijke (en misschien ook meest verbazende) stellingen in de kansrekening en statistiek, de *Centrale limietstelling*. Deze luidt als volgt:

Stelling: Als $X_1, X_2, \ldots$ onafhankelijke stochasten zijn met verwachtingswaarde $E[X_i]$ en variantie $Var(X_i)$, dan is de limiet

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n}(X_i - E[X_i])}{\sqrt{\sum_{i=1}^{n} Var(X_i)}}$$

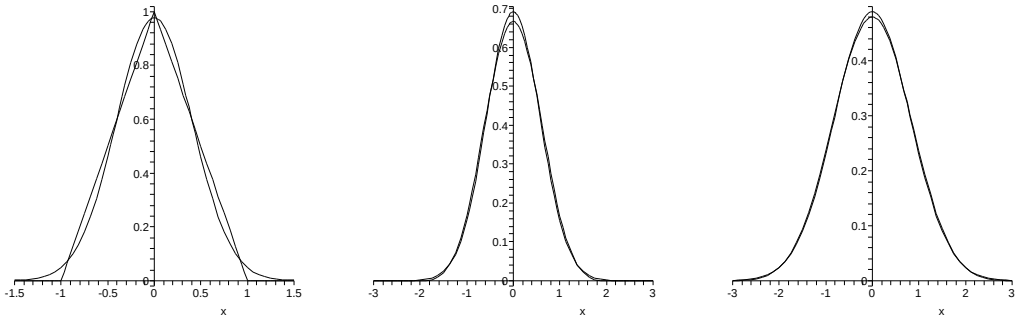
onder zwakke verdere voorwaarden aan de X_i een standaard-normaal verdeelde stochast. In het bijzonder wordt aan de voorwaarden voldaan als alle X_i

dezelfde standaardafwijking σ hebben, in dit geval convergeert

$$\frac{1}{\sqrt{n}\,\sigma}\left(\sum_{i=1}^{n} X_i - E[X_i]\right)$$

tegen de standaard-normale verdeling.

Uit deze stelling kunnen we omgekeerd concluderen dat de normale verdeling met verwachtingswaarde $\mu = \sum_{i=1}^{n} E[X_i]$ en variantie $\sigma^2 = \sum_{i=1}^{n} Var(X_i)$ een benadering geeft voor de kansverdeling van de stochast $X := \sum_{i=1}^{n} X_i$. Hoe goed deze benadering is, hangt van de verdelingen van de enkele stochasten X_i en natuurlijk van n af.



Figuur 12: Benadering van de som van n uniforme verdeling door een normale verdeling voor $n = 2$, $n = 4$ en $n = 8$.

Als voorbeeld kijken we naar de combinatie van n stochasten X_i met uniforme verdelingen op het interval $[-\frac{1}{2}, \frac{1}{2}]$. Omdat de verdelingen symmetrisch rond 0 liggen, is $E[X_i] = 0$ en voor de variantie geldt $Var(X_i) = \frac{1}{12}$. De som $X_1 + \ldots + X_n$ wordt dus benaderd door de normale verdeling met $\mu = 0$ en $\sigma^2 = \frac{n}{12}$. In Figuur 12 is de benadering voor $n = 2$, $n = 4$ en $n = 8$ te zien. Het is duidelijk, dat al voor $n = 4$ de normale verdeling een heel goede benadering geeft.

2.2 Steekproeven

We hebben gezien hoe we uit een verzameling gegevens uitspraken kunnen afleiden over typische waarden, spreiding, scheefheid, enz. van de gegevens. Hierbij hebben we altijd gebruik gemaakt van de kennis van *alle* gegevens. In de praktijk is dit vaak ondoenlijk of onwenselijk, omdat we uitspraken willen maken over een verzameling gegevens waarvan we niet ieder individu te pakken krijgen. In zo'n geval nemen we een deel van de gegevens - een *steekproef* - en proberen uit de resultaten op de steekproef conclusies over de volledige verzameling gegevens te trekken. Voorbeelden van deze situatie zijn:

- Verkiezingen: Om de percentages van de verschillende opties (verschillende partijen, ja/nee bij een referendum) bij een toekomstige verkiezing te schatten, wordt in een enquête een steekproef van typisch 1000 of 2000 mensen ondervraagd.

- Kwaliteitstoetsen: Om de percentage defecte stukken in een productie te schatten, nemen we een steekproef en testen de gekozen stukken. Het relatieve aantal defecte stukken in de steekproef nemen we als gok voor de percentage in de volledige productie.

- Gemiddelde waarden: Om de gemiddelde *intelligentiequotiënt* of *body-mass-index* in de bevolking te schatten, bepalen we deze voor een geselecteerde groep mensen.

Het idee achter het nemen van een steekproef zit in de veronderstelling, dat de steekproef *representatief* voor de volledige verzameling is. De manier hoe een steekproef wordt genomen, heeft natuurlijk een grote invloed erop of dit inderdaad klopt. Het is bijvoorbeeld bekend dat zekere groepen in de bevolking duidelijk verschillend resultaten bij verkiezingen opleveren, afhankelijk van inkom, leeftijd of burgerlijke staat. Men moet daarom ervoor zorgen, dat deze factoren in de steekproef met de juiste relatieve frequenties gerepresenteerd zijn.

Een voorbeeld van een slechte steekproef is, bij een enquête gewoon de eerste 100 mensen te vragen die je tegenkomt. Dit zou bijna nooit representatief zijn, omdat je op zekere plekken vooral mensen met gemeenschappelijke eigenschappen tegenkomt, op het station bijvoorbeeld mensen die naar hun werkplek reizen en op de campus van de universiteit studenten. Ook als je in de telefoongids willekeurig nummers kiest, is dit meestal niet representatief, omdat je mensen zonder telefoon buiten beschouwing laat en afhankelijk van de tijd verschillende bewoners van een woning bereikt.

Het juiste kiezen van een steekproef is een moeilijke taak waarmee zich een belangrijk speciaal gebied van de statistiek bezig houdt.

We zullen ons echter in dit college niet verder met de vraag van het juiste opzetten van steekproeven bemoeien, we gaan er van nu af van uit dat we het altijd goed hebben gedaan en het met een *aselecte steekproef* te maken hebben. Hiermee bedoelen we dat de steekproef aan de volgende twee eisen voldoet:

(1) De steekproef is *onbevooroordeeld* (unbiased): Elk individu heeft dezelfde kans om gekozen te worden.

(2) De steekproef is *onafhankelijk*: De keuze van één individu voor de steekproef heeft geen invloed op de kansen van de andere individuen om in de steekproef te komen.

Het gemiddelde van een steekproef

Vaak berekenen we het gemiddelde van een steekproef en gebruiken dit als schatting voor het gemiddelde (of de verwachtingswaarde) van de volledige populatie. Als we bijvoorbeeld bij een kwaliteitstoets de kans op een foutief stuk in een productieproces willen bepalen, nemen we hiervoor als *schatting* de relatieve frequentie van foutieve stukken in een (aselecte) steekproef. De vraag is nu, hoe goed de schatting vanuit de steekproef voor de echte kans is, dus hoe sterk het gemiddelde van de steekproef van het gemiddelde van de populatie afwijkt.

Het cruciale idee, om bij deze vraag verder te komen, is dat we ons voorstellen, het nemen van de steekproef vaak te herhalen en de uitslagen van de enkele steekproeven als toevalsexperiment, dus als stochast te beschouwen.

Stel we hebben een steekproef $x_1, \ldots, x_n$. Dan kunnen we ieder element x_i in de steekproef als resultaat van een stochast X_i beschouwen en als we veronderstellen dat de elementen in de steekproef op grond van hetzelfde proces geproduceerd worden, hebben de stochasten X_i alle dezelfde kansverdeling. Merk op dat we bij deze aanpak iets over het onderliggende proces veronderstellen, bijvoorbeeld dat bij de productie van de gecontroleerde stukken inderdaad elk stuk met kans p defect is en dat dit bij de verschillende stukken onafhankelijk gebeurt.

Als we nu naar *alle* mogelijke steekproeven $x_1, \ldots, x_n$ willen kijken, kunnen we dit met behulp van de stochasten $X_1, \ldots, X_n$ beschrijven, want X_i geeft juist de kans aan waarmee het resultaat x_i voorkomt. Op deze manier krijgen we bijvoorbeeld voor het *steekproefgemiddelde* $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$ de stochast $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ die de verdeling van de steekproefgemiddelden over alle mogelijke steekproeven aangeeft.

Merk op: Het is in de literatuur gebruikelijk, een concrete steekproef met kleine letters (zo als x_1, x_2, y) aan te geven, terwijl hoofdletters (zo als X_1, X_2, Y) de stochasten voor de verdeling over alle steekproeven aangeven.

Voorbeeld: Zij X de stochast van een Bernoulli-experiment met parameter p, d.w.z. er geldt $P(X = 1) = p$ en $P(X = 0) = 1 - p$. De verwachtingswaarde $E[X]$ is dan $E[X] = p \cdot 1 + (1 - p) \cdot 0 = p$ en de variantie $Var(X) = p \cdot (1 - p)^2 + (1 - p) \cdot p^2 = p(1 - p)$.

Als we een steekproef van grootte n nemen, herhalen we het Bernoulli-experiment n keer onafhankelijk en hebben hierbij n stochasten $X_1, \ldots, X_n$ met dezelfde verdeling als X. Voor de stochast $\overline{X} := \frac{1}{n}(X_1 + \ldots + X_n)$ die de relatieve frequentie van 1en bij n pogingen aangeeft, hebben we

$$E[\overline{X}] = \frac{1}{n}(p + \ldots + p) = \frac{1}{n}\, np = p$$

dus is de verwachtingswaarde van de steekproefgemiddelden inderdaad de juiste parameter p. Als we dus meerdere steekproeven nemen, kunnen we ervan uitgaan dat de *ware* waarde van p ongeveer het gemiddelde van de steekproefgemiddelden is.

Maar natuurlijk zullen we niet meerdere steekproeven apart nemen, dan zouden we ook meteen een grotere steekproef kunnen nemen. Interessanter is de vraag hoe ver het steekproefgemiddelde van de juiste waarden van p afwijkt. Maar hierover maakt juist de variantie $Var(\overline{X})$ van de stochast $\overline{X}$ een uitspraak, we kunnen verwachten dat het steekproefgemiddelde meestal binnen één standaardafwijking $\sqrt{Var(\overline{X})}$ van p ligt. De variantie van $\overline{X}$ berekenen we als

$$Var(\overline{X}) = \frac{1}{n^2}(p(1 - p) + \ldots + p(1 - p)) = \frac{1}{n^2}\, np(1 - p) = \frac{1}{n}\, p(1 - p).$$

Dit betekent dat het steekproefgemiddelde een standaardafwijking van $\sqrt{\frac{p(1-p)}{n}}$

heeft. In het bijzonder neemt de onzekerheid van de schatting met de wortel uit de grootte van de steekproef af.

Omdat we steeds van een aselecte steekproef uitgaan, is voor het n keer herhalen van een Bernoulli-experiment de Centrale limietstelling van toepassing en we krijgen voor niet te kleine n als verdeling voor de waarde van $\overline{X}$ (bij benadering) een normale verdeling. Dit betekent dat het steekproefgemiddelde met een kans van ongeveer 68% in het interval $\left[p - \sqrt{\frac{p(1-p)}{n}}, p + \sqrt{\frac{p(1-p)}{n}}\right]$ ligt.

Merk op dat we in het voorbeeld een alternatieve verdeling met parameter p verondersteld hebben, en hiermee iets over de verdeling van $\overline{X}$ konden zeggen. Dit is de situatie van een *hypothese* die we over de onderliggende kansverdeling hebben en die we met de realisaties $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ van $\overline{X}$ op concrete steekproeven kunnen toetsen. Het probleem van het toetsen van hypothesen zullen we later in dit college behandelen.

Het resultaat van het voorbeeld met het Bernoulli-experiment geldt inderdaad algemeen voor het bepalen van het gemiddelde van gegevens. Stel we willen het gemiddelde van een zekere grootheid bepalen, dan zien we elke meting als het resultaat van een kansexperiment met een stochast X die een zekere kansverdeling heeft. We *veronderstellen* dus een stochast X met verwachtingswaarde $E[X]$ en standaardafwijking $\sigma = \sigma_X = \sqrt{Var(X)}$.

Bij een steekproef van n metingen beschouwen we het *steekproefgemiddelde* $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$ als uitkomst voor de nieuwe stochast $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$, waarbij de stochasten X_i dezelfde kansverdeling als de veronderstelde stochast X hebben. Voor de stochast $\overline{X}$ van het steekproefgemiddelde geldt nu:

$$E[\overline{X}] = \frac{1}{n}(E[X_1] + \ldots + E[X_n]) = \frac{1}{n}\, n \cdot E[X] = E[X]$$

en

$$Var(\overline{X}) = \frac{1}{n^2}(Var(X_1) + \ldots + Var(X_n)) = \frac{1}{n^2}\, n \cdot Var(X) = \frac{1}{n}\, \sigma_X^2$$

dus geldt voor de variantie $\sigma^2_{\overline{X}}$ en de standaardafwijking $\sigma_{\overline{X}}$ van $\overline{X}$:

$$\sigma^2_{\overline{X}} = \frac{1}{n}\, \sigma_X^2 \quad \text{en} \quad \sigma_{\overline{X}} = \frac{1}{\sqrt{n}}\, \sigma_X.$$

De verdeling van het steekproefgemiddelde heeft dus dezelfde verwachtingswaarde als de onderliggende kansverdeling en de standaardafwijking neemt met de wortel uit de grootte van de steekproef af. Merk op dat we bij het berekenen van de variantie van $\overline{X}$ weer gebruik ervan hebben gemaakt dat de X_i onafhankelijk zijn, dus dat we het met een aselecte steekproef te maken hebben.

> Strikt genomen geldt $\sigma^2_{\overline{X}} = \frac{1}{n}\, \sigma_X^2$ voor de variantie van $\overline{X}$ alleen maar als we een steekproef uit een oneindige populatie nemen of als we de steekproef door trekken met terugleggen verkrijgen. Dit is bijvoorbeeld bij herhaalde metingen van een waarde van toepassing, want in principe kunnen we oneindig lang doorgaan met de metingen en de populatie is dus oneindig.

Als een steekproef van grootte n uit een eindige populatie met N elementen door trekken *zonder terugleggen* genomen wordt, geldt voor de variantie van het steekproefgemiddelde

$$\sigma^2_{\overline{X}} = \frac{1}{n}\,\sigma^2_X\left(\frac{N-n}{N-1}\right).$$

Maar deze correctie kunnen we in de praktijk bijn altijd verwaarlozen, omdat N veel groter is dan n (anders zouden we geen steekproef nemen, maar de hele populatie bekijken) en dus $\frac{N-n}{N-1}$ heel dicht bij 1 ligt.

Het probleem is nu, dat we over de kwaliteit van onze schatting voor het gemiddelde $E[X]$ alleen iets kunnen zeggen als we de standaardafwijking σ_X van X kennen.

De standaardafwijking van een steekproef

Net zo als we het steekproefgemiddelde als het rekenkundig gemiddelde $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$ van de waarden in een steekproef hebben gedefinieerd, kunnen we ook een *steekproefvariantie* en een *steekproefstandaardafwijking* definiëren. De voor de hand liggende gedachte zou zijn, de steekproefvariantie door $\frac{1}{n}((x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2)$ te definiëren. Maar met het steekproefgemiddelde is al een afhankelijkheid tussen de x_i gegeven, als we namelijk $x_1, \ldots, x_{n-1}$ en $\overline{x}$ kennen, ligt x_n vast. Men zegt daarom, dat we slechts nog $n-1$ *vrijheidsgraden* hebben, omdat we met $\overline{x}$ een afhankelijkheid tussen de x_i ingevoerd hebben. In plaats van de som van de kwadratische afstanden door n te delen, delen we door het aantal $n-1$ van onafhankelijke waarden in de steekproef en krijgen

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{en} \quad s := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

voor de steekproefvariantie en de steekproefstandaardafwijking.

Er is ook een minder heuristische verklaring voor het gebruiken van $n-1$ in plaats van n in de noemer. Dit hangt samen met de theorie van *schatters* die we straks gaan bediscussiëren. Het cruciale punt is, dat we graag willen dat de verwachtingswaarde van de steekproefvariantie de ware variantie σ^2 van de onderliggende verdeling geeft, net zo als de verwachtingswaarde voor het steekproefgemiddelde de ware verwachtingswaarde $E[X]$ geeft.

Om dit te analyseren, definiëren we weer een stochast X met de onderliggende kansverdeling en nemen aan dat alle mogelijke steekproeven door onafhankelijke stochasten $X_1, \ldots, X_n$ met dezelfde kansverdeling als X beschreven worden. De verwachtingswaarde en variantie van X noteren we met $\mu := E[X]$ en $\sigma^2 := Var(X)$. We weten dat $\sigma^2 = E[X^2] - E[X]^2$, dus is $E[X^2] = \sigma^2 + \mu^2$.

De stochast $\overline{X}$ voor het steekproefgemiddelde is weer gedefinieerd door $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + \ldots + X_n)$. Er geldt

$$(X_i - \overline{X})^2 = (X_i - \frac{1}{n}(\sum_j X_j))^2 = X_i^2 - \frac{2}{n}X_i(\sum_j X_j) + \frac{1}{n^2}\sum_{j,k} X_j X_k.$$

Als we dit over alle indices i optellen krijgen we

$$\begin{aligned}\sum_i (X_i - \overline{X})^2 &= \sum_i X_i^2 - \frac{2}{n}\sum_{i,j} X_i X_j + n\frac{1}{n^2}\sum_{j,k} X_j X_k \\ &= \sum_i X_i^2 - \frac{1}{n}\sum_{j,k} X_j X_k = \sum_i X_i^2 - \frac{1}{n}(\sum_i X_i)^2.\end{aligned}$$

Er geldt $E[X_i^2] = \sigma^2 + \mu^2$, $E[\sum_i X_i] = n\mu$ en $Var(\sum_i X_i) = n\sigma^2$. Hieruit volgt $E[(\sum_i X_i)^2] = Var(\sum_i X_i) + E[\sum_i X_i]^2 = n\sigma^2 + n^2\mu^2$ en hiermee krijgen we

$$\begin{aligned}E[\sum_i (X_i - \overline{X})^2] &= E[\sum_i X_i^2] - \frac{1}{n}E[(\sum_i X_i)^2] \\ &= n(\sigma^2 + \mu^2) - \frac{1}{n}(n\sigma^2 + n^2\mu^2) = n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\ &= (n-1)\sigma^2.\end{aligned}$$

We moeten dus de steekproefvariantie als $s^2 := \frac{1}{n-1}(\sum_i (x_i - \overline{x})^2)$ definiëren, om als verwachtingswaarde van de steekproefvariantie over alle steekproeven de variantie σ^2 te krijgen. De stochast die de verdeling van de steekproefvarianties beschrijft noemen we S^2 en definiëren deze door

$$S^2 := \frac{1}{n-1}(\sum_i (X_i - \overline{X})^2).$$

2.3 Student t-verdeling en χ^2-verdeling

Bij een stochast X krijgen we de verdeling van de z-waarden door $Z := \frac{X-\mu}{\sigma}$ en analoog krijgen we bij een steekproef van n waarden de z-waarde van het steekproefgemiddelde als

$$z := \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\overline{x} - \mu}{s}\sqrt{n}$$

waarbij we de onbekende standaardafwijking σ door de steekproefstandaardafwijking s vervangen.

Om de verdeling van de z-waarden van het steekproefgemiddelde te beschrijven, interpreteren we de elementen x_i van een steekproef weer als realisaties van stochasten X_i, dan wordt de verdeling van de z-waarden beschreven door de stochast

$$T := \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\overline{X} - \mu}{S}\sqrt{n} \text{ met } \overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \text{ en } S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Voor een normaal verdeelde stochast X heet de kansverdeling van T de *Student t-verdeling* met $n-1$ vrijheidsgraden. De Student t-verdeling is platter dan de standaard-normale verdeling maar komt voor groeiende n steeds dichter bij de standaard-normale verdeling. De oorzaak hiervoor is de onzekerheid over de variantie die de steekproefgemiddelden sterker om de ware waarde van het gemiddelde verspreid.

De rare naam van deze verdeling gaat terug op William Sealey Gosset (1876-1937), die 1908 een artikel hierover gepubliceerd heeft. Omdat hij als medewerker van de *Guinness* brouwerij niet onder zijn eigen naam mocht publiceren, koos hij het pseudoniem *Student* voor zijn wetenschappelijke artikelen. Een beschrijving van hem zegt: *To many in the statistical world "Student"was regarded as a statistical advisor to Guinness's brewery, to others he appeared to be a brewer devoting his spare time to statistics.*

De dichtheidsfunctie van de Student t-verdeling met n vrijheidsgraden is

$$f_n(x) := C_n\,(1+\frac{x^2}{n})^{-\frac{n+1}{2}}$$

waarbij de normaliseringsconstante C_n gegeven is door

$$C_n := \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\cdot\frac{1}{\sqrt{\pi n}}.$$

De hierbij optredende *Gamma-functie* $\Gamma(t)$ is gedefinieerd door

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,dt.$$

Ook dit is (net als de verdelingsfunctie van de normale verdeling) een functie die niet zonder integraal te schrijven is. Uit de eigenschappen $\Gamma(t+1) = t\Gamma(t)$ en $\Gamma(1) = 1$ volgt dat $\Gamma(n+1) = n!$ voor natuurlijke getallen n. De Gamma-functie is dus een soort interpolatie van de faculteit en speelt daarom in veel gebieden van de wiskunde een belangrijke rol.

Omdat de Student t-verdeling symmetrisch is, heeft een stochast T met deze verdeling de verwachtingswaarde $E[T] = 0$. Heeft T een verdeling met $n \geq 3$ vrijheidsgraden, dan geldt

$$Var(T) = \frac{n}{n-2},$$

de variantie is dus inderdaad groter dan bij de standaard-normale verdeling.

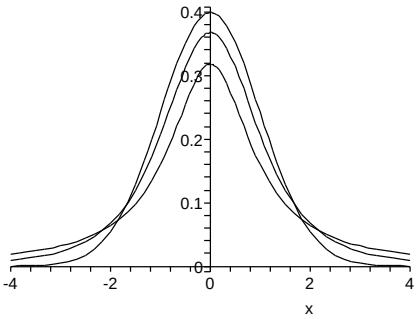
Met de Student t-verdeling hebben we iets over de verdeling van de steekproefgemiddelden kunnen zeggen. Een andere klasse van functies is geschikt om de verdeling van de steekproefvarianties te beschrijven.

Voor n standaard-normaal verdeelde stochasten $X_1, \ldots, X_n$ heet de verdeling van de stochast $Y = X_1^2 + \ldots + X_n^2$ een χ^2-*verdeling* met n vrijheidsgraden.

Voor de stochast S^2 van de steekproefvarianties geldt

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2 = \frac{\sigma^2}{n-1}\sum_i (\frac{X_i - \overline{X}}{\sigma})^2$$

maar $\frac{X_i-\overline{X}}{\sigma}$ is zelf niet standaard-normaal verdeeld. Dit geldt echter wel voor $\frac{X_i-\mu}{\sigma}$ dus is $\sum_i(\frac{X_i-\mu}{\sigma})^2$ een χ^2-verdeling met n vrijheidsgraden. Met behulp



Figuur 13: Student t-verdeling voor $n = 1$ en $n = 3$ in relatie tot standaard-normale verdeling.

van de relatie

$$\sum_i (X_i - \overline{X})^2 = \sum_i (X_i - \mu)^2 - n(\overline{X} - \mu)^2$$

laat zich aantonen dat $\sum_i (\frac{X_i - \overline{X}}{\sigma})^2$ inderdaad wel een χ^2-verdeling met $n-1$ vrijheidsgraden is, dus geldt samengevat:

$$\frac{n-1}{\sigma^2} S^2 = \sum_i (\frac{X_i - \overline{X}}{\sigma})^2 \text{ heeft een } \chi^2\text{-verdeling met } n-1 \text{ vrijheidsgraden.}$$

Ook de χ^2-verdelingen kunnen we expliciet aangeven, de χ^2-verdeling met n vrijheidsgraden heeft de dichtheidsfunctie

$$f_n(x) = \begin{cases} C_n\, x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{voor } x > 0 \\ 0 & \text{voor } x \leq 0, \end{cases} \quad \text{waarbij } C_n = (2^{\frac{n}{2}} \cdot \Gamma(\frac{n}{2}))^{-1}.$$

Voor een stochast Y met χ^2-verdeling met n vrijheidsgraden geldt

$$E[Y] = n \text{ en } Var(X) = 2n$$

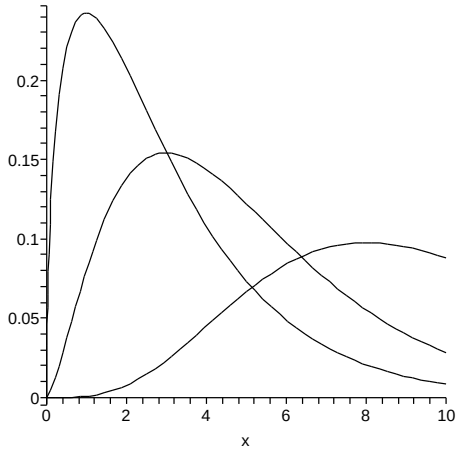
en voor $n \to \infty$ wordt de χ^2-verdeling steeds beter benaderd door een normale verdeling met $\mu = n$ en $\sigma^2 = 2n$.

We zullen de χ^2-verdeling in het kader van betrouwbaarheidsintervallen en het toetsen van hypothesen nog vaker tegen komen.

2.4 Schatters

We hebben vaak gezegd dat het steekproefgemiddelde $\overline{x} := \frac{1}{n}\sum_{i=1}^n x_i$ een *schatting* voor het gemiddelde van de populatie is. We zullen nu kort het algemene begrip van een schatting toelichten.

De meeste kansverdelingen die in de statistiek een rol spelen, hangen van een of meerdere parameters af, de normale verdeling bijvoorbeeld van de verwachtingswaarde μ en de variantie σ^2 en de exponentiële verdeling met dichtheidsfunctie $f(x) = \lambda\, e^{-\lambda x}$ van de intensiteit λ.



Figuur 14: χ^2-verdelingen voor $n = 3$, $n = 5$ en $n = 10$.

Een *schatting* is een functie (of procedure) die uit een steekproef $x_1, \ldots, x_n$ een parameter $\theta = t(x_1, \ldots, x_n)$ van een kansverdeling probeert te bepalen. Hierbij neemt men aan, dat de gegevens door een stochast X met deze kansverdeling met parameter θ voortgebracht zijn. Als we de elementen x_i in de steekproef nu weer als realisaties van stochasten X_i zien die alle dezelfde kansverdeling hebben als X, dan noemen we de stochast $T = t(X_1, \ldots, X_n)$ die de verdeling van de schattingen over alle steekproeven aangeeft een *schatter* voor θ. We hebben al voorbeelden van schatters gezien:

- $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ is een schatter voor de verwachtingswaarde $\mu = E[X]$ van X.
- $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is een schatter voor de variantie $\sigma^2 = Var(X)$ van X.

We zeggen dat een schatter T *zuiver* (unbiased) is, als voor elke waarde van de parameter θ voor de kansverdeling van de stochast X geldt, dat de verwachtingswaarde $E[T]$ juist θ oplevert.

We hebben gezien dat $\overline{X}$ en S^2 zuivere schatters zijn. Deze eigenschap van S^2 was juist de reden om bij de steekproefvariantie door $n-1$ en niet door n te delen. Voor de schatter $T := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ hadden we namelijk gezien dat $E[T] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$ voor een stochast X met variantie σ^2. Omdat $\lim_{n\to\infty} E[T] = \sigma^2$ noemt men T een *asymptotisch zuivere* schatter. Dit betekent, dat de schatter voor grote steekproeven wel een goede schatting geeft.

> Alhoewel $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ een zuivere schatter voor de variantie σ^2 is, is $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ *geen* zuivere schatter voor de standaardafwijking σ, d.w.z. in het algemeen is $E[S] \neq \sigma$. Dit ligt simpelweg eraan dat $\sqrt{a+b} \neq \sqrt{a} + \sqrt{b}$.

Er zijn verschillende algemene principes hoe men schatters voor de parameters van kansverdelingen bepaald. We zullen twee van de meest gebruikte van deze principes nu kort bekijken.

Momentenschatters

Meestal is een kansverdeling die van een aantal parameters $\theta_1, \ldots, \theta_s$ afhangt door een dichtheidsfunctie $f(x)$ gegeven, die van deze parameters afhangt. Als we voor zo'n verdeling de momenten μ'_k (of centrale momenten μ_k) berekenen, hangen deze natuurlijk ook van de parameters $\theta_1, \ldots, \theta_s$ af. Bij de normale verdeling met parameters μ en σ^2 hebben we bijvoorbeeld $\mu'_1 = \mu$ en $\mu'_2 = \sigma^2 + \mu^2$. Vaak is het mogelijk, deze vergelijkingen naar de parameters op te lossen, waarbij men steeds net zo veel momenten als parameters bekijkt. Voor de normale verdeling geeft dit de relaties

$$\mu = \mu'_1 \quad \text{en} \quad \sigma^2 = \mu'_2 - \mu'^2_1.$$

Het idee van een *momentenschatter* is nu, als schatting voor de momenten μ'_k de steekproefmomenten $m'_k := \frac{1}{n}\sum_{i=1}^n x_i^k$ te bepalen en $M'_k := \frac{1}{n}\sum_{i=1}^n X_i^k$ als schatter voor het k-de moment μ'_k te definiëren.

Door de schatters voor de momenten in de relaties tussen parameters en momenten in te vullen, krijgen we zo schatters voor de parameters.

Bij de normale verdeling levert dit de oude bekende

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$$

als schatter voor μ en

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n X_i^2 - (\frac{1}{n}\sum_{i=1}^n X_i)^2 = \frac{1}{n}(\sum_{i=1}^n X_i^2 - \frac{1}{n}(\sum_{i=1}^n X_i)^2) = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \\ = \frac{n-1}{n}S^2\end{aligned}$$

als schatter voor σ^2.

De momentenschatter is dus niet noodzakelijk een zuivere schatter.

Maximum likelihood schatters

Als een dichtheidsfunctie $f(x)$ van parameters $\theta_1, \ldots, \theta_s$ afhangt, kunnen we dit ook expliciet uitdrukken door $f(x) = f(x; \theta_1, \ldots, \theta_s)$ te schrijven. Voor een steekproef $x_1, \ldots, x_n$ is dan het product

$$L(\theta_1, \ldots, \theta_s) := \prod_{i=1}^n f(x_i; \theta_1, \ldots, \theta_s)$$

een maat voor de *aannemelijkheid* waarmee een stochast X met parameters $\theta_1, \ldots, \theta_s$ de elementen van de steekproef geproduceerd heeft. Hoe groter deze

aannemelijkheid, hoe beter past de verdeling van de stochast bij de gevonden steekproef.

De *maximum likelihood schatter* (meest aannemelijke schatter) bepaalt daarom de waarden $\theta_1, \ldots, \theta_s$ zo, dat de aannemelijkheid maximaal wordt. Bij een aantal van kansverdelingen is het mogelijk dit expliciet met behulp van afgeleiden uit te rekenen.

Als voorbeeld kijken we naar een exponentiële verdeling met parameter λ. De dichtheidsfunctie is $f(x;\lambda) = \lambda\, e^{-\lambda x}$ en als aannemelijkheid voor een steekproef $x_1, \ldots, x_n$ krijgen we

$$L(\lambda) = \prod_{i=1}^{n} \lambda\, e^{-\lambda x_i} = \lambda^n e^{-\lambda(\sum_i x_i)}.$$

De aannemelijkheid is maximaal als $L'(\lambda) = 0$ en voor de afgeleide krijgen we

$$L'(\lambda) = n\lambda^{n-1} e^{-\lambda(\sum_i x_i)} - \lambda^n e^{-\lambda(\sum_i x_i)}(\sum_i x_i) = \lambda^{n-1} e^{-\lambda(\sum_i x_i)}(n - \lambda(\sum_i x_i))$$

en er geldt $L'(\lambda) = 0$ als $n - \lambda(\sum_i x_i) = 0$, dus voor

$$\lambda = \frac{n}{\sum_i x_i} = \frac{1}{\overline{x}}.$$

Dit is natuurlijk precies het verwachte resultaat en dit is inderdaad voor de meeste van de gebruikelijke verdelingen het geval.

> Omdat de aannemelijkheid $L(\theta) = f(x_1;\theta) \cdot \ldots \cdot f(x_n;\theta)$ een product van n uitdrukkingen in θ is, is het vaak onhandig de afgeleide van deze functie te bepalen. Wegens de productregel krijgt men hierbij namelijk heel veel termen. Het is daarom vaak handig, in plaats van de functie $L(\theta)$ zelfs de logaritme $\log(L(\theta))$ te bekijken, omdat
>
> $$\log(L(\theta)) = \log(f(x_1;\theta)) + \ldots + \log(f(x_n;\theta)).$$
>
> Omdat de logaritme een monotoon stijgende functie is, neemt $\log(L(\theta))$ precies voor dezelfde waarde van θ zijn maximum aan als $L(\theta)$, daarom kan men in plaats van de nulpunten van $L'(\theta)$ ook de nulpunten van $\log(L(\theta))'$ bepalen.

Voor de normale verdeling levert de maximum likelihood schatter hetzelfde resultaat als de momentenschatter, dus krijgt men ook hier niet in elk geval een zuivere schatter. Er laat zich wel aantonen dat de maximum likelihood schatters altijd asymptotisch zuiver zijn.

Belangrijke begrippen in deze les

- normale verdeling
- normale benadering

- Centrale limietstelling
- steekproef, aselecte steekproef
- steekproefgemiddelde, -variantie, -standaardafwijking
- Student t-verdeling
- χ^2-verdeling
- momentenschatter
- maximum likelihood schatter

Opgaven

6. Een populatie bestaat uit de vier waarden 3, 7, 11 en 13. Een mogelijke schatter voor het gemiddelde van de populatie is, steekproeven van 2 elementen *met* terugleggen te nemen en hiervan het gemiddelde te bepalen.

   (i) Bereken het gemiddelde van de schattingen over alle mogelijke steekproeven (dus de verwachtingswaarde van de schatter). Vergelijk dit met het echte gemiddelde van de populatie.

   (ii) Bepaal de standaardafwijking voor deze schatter van het gemiddelde van de populatie.

   (iii) Bij een alternatieve schatter neem je steekproeven van 2 elementen *zonder* terugleggen. Bepaal weer de verwachtingswaarde en de standaardafwijking van deze schatter, dus het gemiddelde van de steekproefgemiddelden over alle mogelijke steekproeven en de standaardafwijking van alle steekproefgemiddelden.

7. Bij een steekproef van n stukken worden s defecte stukken gevonden, de schatting voor de kans p op een defect stuk is dus $\overline{p} = \frac{s}{n}$. Voor een *gegeven* waarde van p laat zich de kwaliteit van de schatting makkelijk toetsen, omdat in dit geval de standaardafwijking van de verdeling van schattingen (dus de standaardafwijking van de schatter) gegeven is door $\sqrt{\frac{p(1-p)}{n}}$. Maar in veel gevallen is de ware waarde van p onbekend en we moeten onze conclusies alleen uit de steekproef trekken.

   (i) Bij een steekproef van 100 stukken werden 20 defecte stukken gevonden. Bepaal de minimale en de maximale waarde van p zo dat de schatting $\overline{p} = 0.2$ binnen één standaardafwijking (van de schatter) van p ligt.

   (ii) We noteren de grootste waarde van p waarvoor de schatting $\overline{p}$ nog net binnen één standaardafwijking van p ligt met p_{max}. Geef een formule afhankelijk van p_{max}, $\overline{p}$ en n aan, waar p_{max} aan voldoet.
   (Hint: Bepaal een functie van p die p_{max} als nulpunt heeft. Het nulpunt van deze functie kan niet expliciet bepaald worden, maar moet numeriek benaderd worden.)
   Geef ook een formule voor de kleinste waarde p_{min} van p aan, waarvoor $\overline{p}$ nog binnen één standaardafwijking van p ligt.

   (iii) Stel iemand beweert dat zijn schatting van $\overline{p} = 0.2$ binnen één standaardafwijking van 0.01 van de ware waarde van p ligt. Hoe groot moet zijn steekproef voor deze bewering minstens zijn?

8. Zij X een stochast met de drie mogelijke uitkomsten -1, 0 en 1 en met de kansverdeling $P(X=-1) = P(X=1) = \frac{1}{2}p$ en $P(X=0) = 1-p$ die van een parameter $0 \leq p \leq 1$ afhangt. Zij T_0 de stochast die het aantal 0en in een steekproef van grote n aangeeft, en T_1 de stochast die het aantal 1en aangeeft.

   Laat zien dat $\frac{1}{n}(n - T_0)$ en $\frac{2}{n}T_1$ zuivere schatters voor p zijn.

9. Zij X een stochast met uniforme verdeling op het interval $[0, \theta]$, dan is $P(X \leq x) = \frac{x}{\theta}$ voor $0 \leq x \leq \theta$. We willen uit een steekproef $x_1, \ldots, x_n$ een schatting voor θ maken.

   (i) Laat zien dat de schatting $t := \frac{2}{n}(x_1 + \ldots + x_n)$ een zuivere schatter $T := \frac{2}{n}(X_1 + \ldots + X_n) = 2\,\overline{X}$ voor θ geeft.

   (ii) Een andere mogelijke schatting voor θ is het maximum van de gevonden waarden, dus $t_{max} := \max(x_1, \ldots, x_n)$. Laat zien dat voor de schatter $T_{max} := \max(X_1, \ldots, X_n)$ geldt dat $P(T \leq x) = (\frac{x}{\theta})^n$ en concludeer dat T de verdelingsfunctie $f(x) = n\,\frac{x^{n-1}}{\theta^n}$ heeft.

   Ga na dat T_{max} geen zuivere schatter, maar wel een asymptotisch zuivere schatter voor θ is, door te laten zien dat $E[T] = \frac{n}{n+1}\,\theta$.

   (Hint: Er geldt $\int_0^\theta x^n\,dx = \frac{1}{n+1}\theta^{n+1}$.)

   (iii) Laat zien dat $\frac{n+1}{n}\,T_{max}$ een zuivere schatter voor θ is.

10. Laat zien dat voor een stochast X met uniforme verdeling op het interval $[0, \theta]$ de schatter $T_{max} := \max(X_1, \ldots, X_n)$ de maximum likelihood schatter is.

    (Hint: Ga na dat de aannemelijkheid $L(\theta)$ voor een steekproef $x_1, \ldots, x_n$ gegeven is door $L(\theta) = 0$ als $\theta < \max(x_1, \ldots, x_n)$ en $L(\theta) = \frac{1}{\theta^n}$ als $\theta \geq \max(x_1, \ldots, x_n)$.)

11. Voor een stochast X met uniforme verdeling op het interval $[0, \theta]$ wordt van een steekproef x_1, x_2 van twee waarden de schatting $t := 3|x_1 - x_2|$ voor θ gemaakt. Laat zien dat $T := 3|X_1 - X_2|$ een zuivere schatter voor θ is.

12. Bij een zeker chemisch proces wordt de afgegeven energie (warmte) gemeten en er wordt verondersteld dat de afgegeven energie door een stochast X met verwachtingswaarde μ en variantie σ^2 wordt beschreven. Bij 10 metingen zijn de volgende resultaten verkregen:

$$\begin{array}{llllll} x_1 = 1244, & x_2 = 1198, & x_3 = 1212, & x_4 = 1235, & x_5 = 1245, \\ x_6 = 1190, & x_7 = 1202, & x_8 = 1220, & x_9 = 1233, & x_{10} = 1208. \end{array}$$

    (i) Bepaal het steekproefgemiddelde $\overline{x}$ en de steekproefvariantie s^2 van de metingen.

    (ii) In plaats van over alle steekproefwaarden te middelen, zou men ook het gemiddelde van de eerste en de laatste waarde, of het gemiddelde van de waarden x_3 t/m x_8 kunnen nemen. Dit geeft aanleiding tot de schatters $Y := \frac{1}{2}(X_1 + X_{10})$ en $Z := \frac{1}{6}(X_3 + X_4 + X_5 + X_6 + X_7 + X_8)$. Bepaal de schattingen voor het gemiddelde van de aangegeven steekproef met deze twee schatters.

    (iii) Laat zien dat de schatters Y en Z uit (ii) zuiver zijn, d.w.z. dat $E[Y] = E[Z] = E[X]$. Bepaal ook de varianties van deze schatters.

    (iv) De schatter Y voor de verwachtingswaarde μ van X kunnen we ook voor een algemeen steekproef van grote n definiëren door $Y := \frac{1}{2}(X_1 + X_n)$. Laat zien dat deze schatter Y verwachtingswaarde $E[Y] = \mu$ en variantie $Var(Y) = \frac{\sigma^2}{2}$ heeft.

13. We hebben gezien dat $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ een zuivere schatter voor de verwachtingswaarde $\mu = E[X]$ is. Laat zien dat $\overline{X}^2$ *geen* zuivere schatter voor μ^2 is.

Les 3 Betrouwbaarheidsintervallen

In de vorige les hebben we erna gekeken hoe we grootheden van een populatie met behulp van steekproeven kunnen schatten. We hebben daarbij gezien dat de nauwkeurigheid van een schatting met de grootte van de steekproef toeneemt, want de steekproefstandaardafwijking neemt met $\frac{1}{\sqrt{n}}$ af. We zullen in deze les bekijken hoe we uitspraken erover kunnen maken dat een schatting met een foutmarge de juiste waarde met een gegeven kans bevat. Hierbij moeten we in het bijzonder precies formuleren, wat de uitspraak dat *een waarde met een betrouwbaarheid van* 95% *in een zeker interval ligt* eigenlijk betekent.

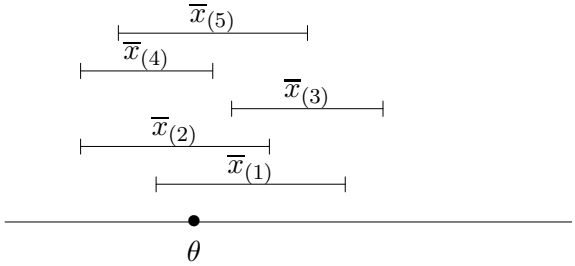
3.1 Intervalschatters

De schatters die we in de vorige les hebben bekeken, noemt men vaak *puntschatters* omdat ze voor een gegeven steekproef een precieze waarde voor een parameter opleveren. Bijvoorbeeld levert de schatter $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ voor het gemiddelde van een populatie op een gegeven steekproef $x_1, \ldots, x_n$ de schatting $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

In tegenstelling hiertoe geeft een *intervalschatter* voor een gegeven steekproef een interval aan waarin de juiste waarde θ van de parameter moet liggen. Hierbij wordt altijd een level γ van *betrouwbaarheid* geëist, waarmee het interval de juiste waarde bevat. De betrouwbaarheid γ wordt als volgt geïnterpreteerd: Voor een gegeven waarde van θ is γ de kans dat een steekproef een interval oplevert dat θ bevat. We kijken dus weer naar alle mogelijke steekproeven en analyseren de verdeling van de schattingen.

Merk op: Een betrouwbaarheid van 95% voor een interval betekent *niet* dat de juiste waarde θ met kans 95% in het interval ligt, maar dat onze methode om het interval te schatten voor 95% van de mogelijke steekproeven een interval oplevert, dat θ bevat.

Bij een betrouwbaarheid van $\gamma = 0.8$ zouden we dus bij vijf steekproeven verwachten, dat de juiste parameter vier keer in het geschatte interval ligt, bijvoorbeeld zo als in het volgende plaatje met de intervallen rond de schattingen $\overline{x}_{(i)}$ aangegeven.



In de taal van stochasten en schatters levert dit idee van betrouwbaarheid het volgende concept op. Zij X een stochast met dichtheidsfunctie $f(x) := f(x;\theta)$ en verdelingsfunctie $F(x) := F(x;\theta)$ die van een parameter θ afhangen, dan berekenen we de kansen voor X door

$$P(X \leq x) = P_\theta(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

We noemen een paar (T_1, T_2) van stochasten een *intervalschatter van betrouwbaarheid γ voor θ* als

$$P(T_1 \leq \theta \leq T_2) = \gamma \text{ voor elke mogelijke waarde van de parameter } \theta.$$

Een realisatie van een intervalschatter op een concrete steekproef $x_1, \ldots, x_n$ heet een *betrouwbaarheidsinterval* van betrouwbaarheid γ voor θ.

Omdat we de waarde van θ van twee zijden ingeschakeld hebben, noemen we het paar (T_1, T_2) ook een *tweezijdige intervalschatter*.

> Als we in de praktijk een betrouwbaarheidsinterval voor de verwachtingswaarde $\mu := E[X]$ schatten, zal het interval bijna altijd symmetrisch rond het steekproefgemiddelde $\overline{x}$ liggen. Dit is geen noodzakelijke voorwaarde maar wel heel gebruikelijk. Er laat zich aantonen dat voor een normaal verdeelde stochast X het symmetrische interval rond $\overline{x}$ de kleinste lengte van alle intervallen met betrouwbaarheid γ heeft.

Soms is het interessant om alleen maar een boven- of een benedengrens voor een parameter te schatten. Dit levert *éénzijdige intervalschatters*. We noemen een stochast T_1 een *rechtséénzijdige intervalschatter* van betrouwbaarheid γ als

$$P(T_1 \leq \theta) = \gamma \text{ voor elke mogelijke waarde van de parameter } \theta$$

en we noemen een stochast T_2 een *linkséénzijdige intervalschatter* van betrouwbaarheid γ als

$$P(\theta \leq T_2) = \gamma \text{ voor elke mogelijke waarde van de parameter } \theta.$$

De reden waarom de stochast T_1 met $P(T_1 \leq \theta) = \gamma$ *rechts*éénzijdig heet, hangt met de éénzijdige toetsen samen die we in de volgende les gaan behandelen.

3.2 Betrouwbaarheidsintervallen bij gegeven variantie

Als belangrijk voorbeeld zullen we naar een intervalschatter kijken die voor een normaal verdeelde stochast X met bekende variantie σ^2 een betrouwbaarheidsinterval voor de verwachtingswaarde μ van X geeft.

Hetzelfde principe werkt bij benadering voor de verwachtingswaarde van niet normaal verdeelde stochasten, in het bijzonder voor de verwachte kans op succes bij een binomiale verdeling.

> De centrale limietstelling zegt dat de som van onafhankelijke stochasten goed benaderd wordt door een normale verdeling. Hieruit volgt dat de vorm van de onderzochte stochast X geen grote rol speelt als de steekproefgrootte n niet te klein is. Maar er zijn wel andere problemen, waardoor de verdeling van schattingen van de normale verdeling afwijkt. Deze hebben vooral met de veronderstelling te maken dat we een *aselecte* steekproef hebben genomen. Dit is in de praktijk vaak

lastig, omdat mensen bijvoorbeeld een enquête weigeren, maar dit niet representatief over de populatie gebeurt. Ook is het vaak niet realistisch, dat de verschillende steekproefelementen onafhankelijk van elkaar genomen worden. Het is de kunst van de instituten voor opinieonderzoek deze factoren zo ver mogelijk te onderdrukken of de resultaten navenant te corrigeren.

Stel we hebben een normaal verdeelde stochast $X \in \mathcal{N}(\mu, \sigma^2)$ dan weten we dat $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ een zuivere schatter voor μ is. Omdat X normaal verdeeld is, geldt dit ook voor $\overline{X}$ (de som van onafhankelijke normaal verdeelde stochasten is weer normaal verdeeld) en we weten dat $Var(\overline{X}) = \frac{\sigma^2}{n}$. Hieruit volgt dat de stochast

$$Z := \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma}$$

standaard-normaal verdeeld is.

Als X een niet-normaal verdeelde stochast met verwachtingswaarde μ en variantie σ^2 is, geldt voor $\overline{X}$ nog steeds dat $E[\overline{X}] = \mu$ en $Var(\overline{X}) = \frac{\sigma^2}{n}$, maar $\overline{X}$ is niet meer normaal verdeeld. Uit de Centrale limietstelling volgt echter dat voor een niet te kleine n de verdeling van $\overline{X}$ sterk op een normale verdeling lijkt en hierdoor goed benaderd kan worden.

Voor een stochast $Z \in \mathcal{N}(0, 1)$ met standaard-normale verdeling definiëren we nu de z-waarde z_α van level $\alpha := 1 - \gamma$ door

$$P(Z > z_\alpha) = \alpha.$$

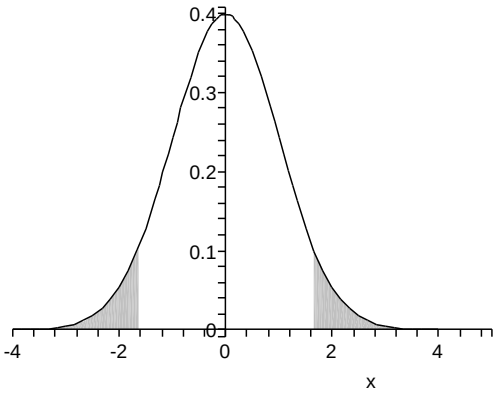
Voor een betrouwbaarheid van 95% is dus $\alpha = 0.05 = 1 - 0.95$ en geeft z_α de waarde aan, waarvoor slechts 5% van de waarden van Z boven z_α liggen en de waarden van Z dus met betrouwbaarheid 95% hoogstens z_α zijn. De level $\alpha = 1 - \gamma$ wordt ook wel de *onbetrouwbaarheid* genoemd.

Omdat de normale verdeling symmetrisch rond 0 is, geldt $P(Z < -z_\alpha) = \alpha$ en dus $P(|Z| > z_\alpha) = 2\alpha$. Hieruit volgt in het bijzonder:

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha = \gamma.$$

De waarden van de standaard-normale verdeling liggen dus met kans $\gamma = 1 - \alpha$ tussen $-z_{\frac{\alpha}{2}}$ en $z_{\frac{\alpha}{2}}$. In Figuur 15 is dit voor $\gamma = 0.9$ aangeduid. Het witte stuk onder de grafiek bevat 90% van de totale oppervlakte onder de grafiek, de resterende 10% liggen in de grijze staarten, dus telkens 5% in de linker- en rechterstaart. De z-waarde $z_{0.05}$ is dus juist het punt waar de rechterstaart begint.

Als we de relatie $P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = \gamma$ nu op de standaard-normaal verdeelde stochast $Z = \frac{(\overline{X}-\mu)\sqrt{n}}{\sigma}$ toepassen, krijgen we voor de betrouwbaarheid



Figuur 15: Standaard-normale verdeling met betrouwbaarheidsinterval voor $\gamma = 0.9$.

γ en onbetrouwbaarheid $\alpha := 1 - \gamma$:

$$\begin{aligned}
&P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = \gamma \Leftrightarrow P(-z_{\frac{\alpha}{2}} \leq \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \leq z_{\frac{\alpha}{2}}) = \gamma \\
\Leftrightarrow\ &P(-z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = \gamma \\
\Leftrightarrow\ &P(\mu - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = \gamma \\
\Leftrightarrow\ &P(\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = \gamma.
\end{aligned}$$

We weten dus dat het steekproefgemiddelde met kans γ niet meer dan $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ van de juiste waarde μ afwijkt. Als intervalschatter voor het gemiddelde nemen we dus (T_1, T_2) met

$$T_1 := \overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{en} \quad T_2 := \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

en een betrouwbaarheidsinterval is een realisatie van de intervalschatter voor een concrete steekproef, dus het interval

$$\left[\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right].$$

Omdat $P(\mu - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = P(\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$ is dit precies het interval van de waarden van μ waarvoor $\overline{x}$ binnen het symmetrische interval rond μ met kansmassa γ valt. Merk op dat de lengte van het betrouwbaarheidsinterval alleen maar van de gekozen betrouwbaarheid γ, de grootte n van de steekproef en de variantie σ^2 van de stochast X afhangt.

Voor éénzijdige betrouwbaarheidsintervallen kunnen we op dezelfde manier als bij de tweezijdige intervallen argumenteren. Voor een rechtséénzijdig interval

met betrouwbaarheid γ en $\alpha := 1 - \gamma$ krijgen we:

$$P(Z \leq z_\alpha) = \gamma \Leftrightarrow P(\frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \leq z_\alpha) = \gamma \Leftrightarrow P(\overline{X} - \mu \leq z_\alpha \frac{\sigma}{\sqrt{n}}) = \gamma$$
$$\Leftrightarrow P(\overline{X} \leq \mu + z_\alpha \frac{\sigma}{\sqrt{n}}) = \gamma \Leftrightarrow P(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu) = \gamma$$

dus is

$$T_1 := \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

een rechtséénzijdige intervalschatter en een concrete steekproef geeft het rechtséénzijdige betrouwbaarheidsinterval

$$\left[\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right].$$

Dit is precies het interval van de waarden van μ waarvoor $\overline{x}$ binnen het naar *rechts* begrensde en naar links open interval rond μ met kansmassa γ valt.

> We zien hier dus de reden waarom de stochast T_1 met $P(T_1 \leq \mu) = \gamma$ een *rechts*éénzijdig betrouwbaarheidsinterval geeft. De waarden van μ die in dit éénzijdige betrouwbaarheidsinterval liggen, zijn namelijk juist de waarden waarvoor $\overline{x}$ een plausibele schatting aangeeft, als we met plausibel bedoelen, dat de schatting $\overline{x}$ niet de ver *rechts* van de ware waarde ligt.

Analoog krijgen we voor het linkséénzijdige betrouwbaarheidsinterval met betrouwbaarheid γ de schatter

$$T_2 := \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}} \quad \text{met} \quad P(\mu \leq T_2) = P(\mu \leq \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}) = \gamma$$

en het linkséénzijdige betrouwbaarheidsinterval

$$\left[-\infty, \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right].$$

Typische waarden voor de betrouwbaarheid γ zijn 90%, 95% en 99%, Tabel 1 geeft de z_α- en $z_{\frac{\alpha}{2}}$-waarden voor een paar gebruikelijke betrouwbaarheden:

γ	α	z_α	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$
0.80	0.20	0.8416	0.10	1.2816
0.90	0.10	1.2816	0.05	1.6449
0.95	0.05	1.6449	0.025	1.9600
0.98	0.02	2.0537	0.01	2.3263
0.99	0.01	2.3263	0.005	2.5758
0.999	0.001	3.0902	0.0005	3.2905

Tabel 1: Kritieke waarden voor de standaard-normale verdeling.

We hebben gezien dat betrouwbaarheidsintervallen door drie parameters beschreven worden:

(i) De grote n van de steekproef.

(ii) De gewenste betrouwbaarheid γ.

(iii) De lengte van het betrouwbaarheidsinterval.

Als we de betrouwbaarheid willen verhogen, moeten we of de steekproef vergroten of een groter interval accepteren. Omgekeerd kunnen we het betrouwbaarheidsinterval alleen maar kleiner maken door of de steekproef te vergroten of een lagere level van betrouwbaarheid te kiezen. Bij een gegeven grootte van de steekproef zijn dus de lengte van het betrouwbaarheidsinterval en de betrouwbaarheid parameters, die elkaar tegenstrijdig beïnvloeden.

Bij het opzetten van een experiment (bijvoorbeeld een enquête) heeft men vaak andere voorwaarden: Voor een gegeven level γ van betrouwbaarheid is er een maximale lengte $2l$ van het betrouwbaarheidsinterval dat als acceptabel beschouwd wordt. Hierdoor wordt de noodzakelijke grootte van de steekproef bepaald, namelijk door:

$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq l \Rightarrow n \geq \left(z_{\frac{\alpha}{2}} \frac{\sigma}{l} \right)^2 = z_{\frac{\alpha}{2}}^2 \frac{\sigma^2}{l^2}.$$

Betrouwbaarheidsinterval voor relatieve frequenties

Als we de kans p schatten waarmee een Bernoulli-experiment een succes oplevert, tellen we het aantal k van successen bij n pogingen een nemen $\overline{p} := \frac{k}{n}$ als schatting voor p. De stochast X die de verdeling van de aantallen van successen beschrijft, is binomiaal verdeeld met parameter p en er geldt $E[X] = np$ en $Var(X) = np(1-p)$. Voor de stochast $\overline{P} := \frac{X}{n}$ die de verdeling van de relatieve aantallen beschrijft, geldt dus $E[\overline{P}] = p$ en $Var(\overline{P}) = \frac{p(1-p)}{n}$. Als n niet te klein en p niet te dicht bij 0 of 1 is, kunnen we met de normale benadering van de binomiale verdeling werken, d.w.z. we kunnen aannemen dat de stochast

$$Z := \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{(\overline{P} - p)\sqrt{n}}{\sqrt{p(1-p)}}$$

goed door de standaard-normale verdeling benaderd wordt. In dit geval kunnen we de redenering van de normale verdeling weer toepassen en we krijgen

$$P\left(\overline{P} - z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \leq p \leq \overline{P} + z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right) = \gamma.$$

Dit geeft het betrouwbaarheidsinterval

$$\left[\overline{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, \overline{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right]$$

voor de schatting van de parameter p.

Het probleem bij de binomiale verdeling is, dat de variantie $\frac{p(1-p)}{n}$ en dus ook de lengte van het betrouwbaarheidsinterval van de gezochte parameter p

afhangt. In de praktijk wordt dit meestal opgelost door p gewoon door $\overline{p}$ te vervangen, men gebruikt hiervoor de *standaard fout* (standard error)

$$SE(\overline{p}) := \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

van $\overline{p}$. De standaard fout is dus een schatting voor de standaardafwijking $\sqrt{Var(\overline{P})}$ van de schatter $\overline{P}$. Met behulp van de standaard fout krijgt men het betrouwbaarheidsinterval

$$\left[\overline{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}, \overline{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right] = \left[\overline{p} - z_{\frac{\alpha}{2}}SE(\overline{p}), \overline{p} + z_{\frac{\alpha}{2}}SE(\overline{p})\right].$$

> Bij een precieze analyse komt men erachter dat de zuivere grenzen voor het betrouwbaarheidsinterval
>
> $$\frac{\overline{p} + \frac{z^2_{\frac{\alpha}{2}}}{2n} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}(1-\overline{p})}{n} + \frac{z^2_{\frac{\alpha}{2}}}{4n^2}}}{1 + \frac{z^2_{\frac{\alpha}{2}}}{n}}$$
>
> zijn, maar voor $n\overline{p} \geq 50$ en $n(1-\overline{p}) \geq 50$ kunnen de correctie termen veilig verwaarloosd worden.

Ook in het geval van de relatieve frequenties kan men de benodigde grootte van de steekproef afschatten om een betrouwbaarheid γ en een maximale lengte van $2l$ voor het betrouwbaarheidsinterval te bereiken. Er geldt dezelfde relatie als bij de normale verdeling, met σ^2 vervangen door $p(1-p)$, dus

$$n \geq z^2_{\frac{\alpha}{2}}\frac{p(1-p)}{l^2}.$$

Merk op dat we hierbij ook weer de gezochte relatieve frequentie p nodig hebben. Omdat we juist willen bepalen, hoe groot we de steekproef moeten kiezen, kunnen we hier niet eens de schatting $\overline{p}$ voor p invullen, maar we kunnen natuurlijk wel een gok doen wat voor een waarde van p we verwachten.

Voorbeeld: Bij een enquête onder 1000 mensen hebben 52% aangegeven *voor* de Europese grondwet te stemmen. Een betrouwbaarheidsinterval op de level 99% geeft een nauwkeurigheid van $z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = 2.5758 \cdot \sqrt{\frac{0.2496}{1000}} \approx 0.041$ voor de schatting $\overline{p} = 0.52$ van de echte proportie van toestemming. Het betrouwbaarheidsinterval is dus $[47.9\%, 56.1\%]$.

Natuurlijk is de interessante vraag, of de toestemming boven de 50% ligt. Om hierover een uitspraak met betrouwbaarheid 99% te kunnen doen, moet de lengte van het betrouwbaarheidsinterval tot 4% worden beperkt. De benodigde grootte van de steekproef hiervoor is $n \geq z^2_{\frac{\alpha}{2}}\frac{p(1-p)}{l^2} = 2.5758^2 \cdot \frac{0.25}{0.02^2} \approx 4147$. Hierbij hebben we voor p de schatting $p = 0.5$ ingevuld, voor $p = 0.52$ zouden we $n \geq 4140$ krijgen, dus bijna hetzelfde.

3.3 Betrouwbaarheidsintervallen bij onbekende variantie

We zijn er tot nu toe van uitgegaan dat we het met een normaal verdeelde stochast X met *bekende* variantie te maken hebben. Omdat dit in de praktijk niet realistisch is, kijken we nu naar het geval van een stochast met onbekende variantie. In dit geval hebben we helaas niets meer aan de stochast $Z := \frac{(\overline{X}-\mu)\sqrt{n}}{\sigma}$, omdat we de variantie σ^2 gewoon niet kennen. Maar we weten wel, dat $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ een zuivere schatter voor σ^2 is, dus kunnen we proberen de onbekende variantie σ^2 door de schatter S^2 te vervangen. Dit geeft de stochast

$$T := \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{(\overline{X}-\mu)\sqrt{n}}{S}$$

die we al in de laatste les zijn tegengekomen: Voor een normaal verdeelde stochast X heeft T de *Student-t verdeling met $n-1$ vrijheidsgraden*. We weten dat deze verdeling voor kleine n meer uitgespreid is dan de standaard-normale verdeling en voor grote n steeds meer op de standaard-normale verdeling lijkt.

Met dezelfde argumenten als in het geval van bekende variantie komen we nu weer naar betrouwbaarheidsintervallen, als we de standaard-normale verdeling altijd door de Student-t verdeling met $n-1$ vrijheidsgraden vervangen.

Analoog met de standaard-normale verdeling definiëren we de t-waarde $t_\alpha := t_{n-1,\alpha}$ van level $\alpha = 1-\gamma$ door

$$P(T > t_\alpha) = \alpha$$

waarbij het aantal $n-1$ van vrijheidsgraden meestal niet aangeven wordt, omdat het uit de samenhang duidelijk is.

Een soortgelijke berekening als boven geeft:

$$\begin{aligned}
&P(-t_{\frac{\alpha}{2}} \le T \le t_{\frac{\alpha}{2}}) = \gamma \Leftrightarrow P(-t_{\frac{\alpha}{2}} \le \frac{(\overline{X}-\mu)\sqrt{n}}{S} \le t_{\frac{\alpha}{2}}) = \gamma \\
\Leftrightarrow\ &P(\mu - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \le \overline{X} \le \mu + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}) = \gamma \\
\Leftrightarrow\ &P(\overline{X} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}) = \gamma.
\end{aligned}$$

Voor een steekproef $x_1, \ldots, x_n$ met steekproefgemiddelde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ en steekproefstandaardafwijking $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ noemen we (net als bij de binomiale verdeling) de schatting $\frac{s}{\sqrt{n}}$ voor de standaardafwijking $\sqrt{Var(\overline{X})}$ van de schatter $\overline{X}$ de *standaard fout* van $\overline{x}$ en noteren dit met $SE(\overline{x})$. Hiermee krijgen we het betrouwbaarheidsinterval

$$\left[\overline{x} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{x} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right] = \left[\overline{x} - t_{\frac{\alpha}{2}}SE(\overline{x}), \overline{x} + t_{\frac{\alpha}{2}}SE(\overline{x})\right]$$

van betrouwbaarheid γ voor μ.

Net zo als bij de standaard-normale verdeling worden de t-waarden voor de meest gebruikelijke levels van betrouwbaarheid en voor de verschillende vrijheidsgraden in tabellen opgeslagen. Inmiddels worden in plaats van tabellen meestal software pakketten gebruikt, die de t-waarden voor een gewenste betrouwbaarheid γ en een gegeven aantal van vrijheidsgraden uitrekenen. Typische waarden van $t_{n,\alpha}$ zijn in Tabel 2 te zien (waarbij we met $n = \infty$ de waarden voor de standaard-normale verdeling aangeven):

$n\backslash\alpha$	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
5	1.476	2.015	2.571	3.365	4.032
10	1.372	1.812	2.228	2.764	3.169
30	1.310	1.697	2.042	2.457	2.750
∞	1.282	1.645	1.960	2.326	2.576

Tabel 2: Kritieke waarden $t_{n,\alpha}$ voor de Student-t verdelingen met n vrijheidsgraden.

Voorbeeld: Men neemt aan dat het aantal lijnen die in een grote telefooncentrale tijdens het spitsuur in gebruik zijn normaal verdeeld is. Uit een steekproef over 11 dagen blijkt een steekproefgemiddelde van $\overline{x} = 120$ voor het aantal lijnen, met een steekproefstandaardafwijking van $s = 10$. Als we een betrouwbaarheidsinterval op level 99% voor het gemiddelde aantal μ van lijnen in gebruik willen bepalen, hebben we de t-waarde $t_{10,0.005}$ nodig, want $n = 11$ en $\alpha = 0.01$. In de tabel vinden we $t_{10,0.005} = 3.169$, dus is de afwijking $t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 3.169 \cdot \frac{10}{\sqrt{11}} \approx 9.6$ en we krijgen het betrouwbaarheidsinterval $[110.4, 129.6]$ voor μ.

3.4 Betrouwbaarheidsintervallen voor de variantie

We hebben in de vorige les aangegeven dat voor een standaard-normaal verdeelde stochast X de stochast

$$Y := \frac{n-1}{\sigma^2} S^2 = \sum_{i=1}^{n} (\frac{X_i - \overline{X}}{\sigma})^2$$

een χ^2-verdeling met $n-1$ vrijheidsgraden heeft. Deze stochast Y is nu geschikt om een betrouwbaarheidsinterval voor de variante aan te geven.

Analoog met de z-waarde voor de standaard-normale verdeling en de t-waarde voor de Student-t verdeling definiëren we de χ^2-waarde $\chi^2_\alpha := \chi^2_{n-1,\alpha}$ door

$$P(Y > \chi^2_\alpha) = \alpha$$

waarbij de index voor het aantal vrijheidsgraden weer weggelaten is.

Omdat de χ^2-verdeling niet symmetrisch is, kunnen we niet meer zo makkelijk uit χ^2_α een waarde χ^2_β afleiden zo dat $P(Y < \chi^2_\beta) = P(Y > \chi^2_\alpha) = \alpha$ is.

Maar uit $P(Y > \chi^2_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$ volgt dat tussen $\chi^2_{1-\frac{\alpha}{2}}$ en $\chi^2_{\frac{\alpha}{2}}$ de kansmassa $(1 - \frac{\alpha}{2}) - \frac{\alpha}{2} = 1 - \alpha = \gamma$ ligt.

> Bij symmetrische verdelingen zo als de normale verdeling laat zich aantonen dat de symmetrische betrouwbaarheidsintervallen de intervallen van minimale lengte voor een gegeven betrouwbaarheid zijn. De χ^2-verdeling is niet symmetrisch, en men kan voor het interval rond Y dat de kansmassa γ bevat ook een willekeurig interval van de vorm $[\chi^2_{\gamma+c}, \chi^2_c]$ kiezen. Zo'n interval heeft inderdaad niet voor $c = \frac{\alpha}{2}$ de minimale lengte, maar de waarde c waarvoor de lengte minimaal is ligt in de praktijk meestal zo dicht bij $\frac{\alpha}{2}$ dat men dit verwaarloost.

Met een analoge redenering als eerder krijgen we voor de stochast Y:

$$P(\chi^2_{1-\frac{\alpha}{2}} \leq Y \leq \chi^2_{\frac{\alpha}{2}}) = 1 - \alpha = \gamma \Leftrightarrow P(\chi^2_{1-\frac{\alpha}{2}} \leq \frac{n-1}{\sigma^2} S^2 \leq \chi^2_{\frac{\alpha}{2}}) = \gamma$$
$$\Leftrightarrow P(\chi^2_{1-\frac{\alpha}{2}} \frac{\sigma^2}{n-1} \leq S^2 \leq \mu + \chi^2_{\frac{\alpha}{2}} \frac{\sigma^2}{n-1}) = \gamma$$
$$\Leftrightarrow P(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}) = \gamma.$$

Voor een concrete steekproef $x_1, \ldots, x_n$ met steekproefvariantie s^2 krijgen we hieruit als betrouwbaarheidsinterval van betrouwbaarheid γ voor σ^2 het interval

$$\left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} \right].$$

We kunnen ook een betrouwbaarheidsinterval voor de standaardafwijking σ aangeven, want worteltrekken geeft

$$P\left(\sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2}}}}\, S \leq \sigma \leq \sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}}}\, S \right) = P\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} \right) = \gamma$$

en hieruit krijgen we het betrouwbaarheidsinterval

$$\left[\sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2}}}}\, s, \sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}}}\, s \right]$$

van betrouwbaarheid γ voor de standaardafwijking σ.

Belangrijke begrippen in deze les

- betrouwbaarheid
- tweezijdige / éénzijdige intervalschatter
- betrouwbaarheidsintervallen

- z-waarde, t-waarde, χ^2-waarde
- standaard fout

Opgaven

14. Zij X een uniform verdeelde stochast op het interval $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ en zij $x_1, \ldots, x_n$ een steekproef voor deze stochast. Laat zien dat $[\min(x_1, \ldots, x_n), \max(x_1, \ldots, x_n)]$ een betrouwbaarheidsinterval voor θ is (dus de realisatie van een intervalschatter) en bepaal de level γ van betrouwbaarheid van dit interval.

15. Bij het bedrijf *Bonanza Banana* heeft een steekproef van 225 aanvragen een gemiddelde verwerkingstijd van $\overline{x} = 7$ `jerks` opgeleverd. Uit langdurige ervaring is bekend dat de standaardafwijking voor de verwerkingstijd $\sigma = 3$ `jerks` bedraagt.

    (i) Bepaal een betrouwbaarheidsinterval voor de level 95% voor de gemiddelde verwerkingstijd.

    (ii) Hoe groot moet de steekproef minstens zijn om op level 95% een betrouwbaarheidsinterval van lengte hoogstens 0.5 `jerks` te hebben?

16. In een aselecte steekproef van 100 studenten geven 18 studenten aan dat ze bekend met de binomiale verdeling zijn.

    (i) Bepaal betrouwbaarheidsintervallen op de levels 90%, 95% en 99% voor het relatieve aantal $\overline{p}$ van studenten die de binomiale verdeling kennen.

    (ii) Hoe groot moet voor ieder van de drie levels uit (i) de steekproef zijn om de lengte van het betrouwbaarheidsinterval op hoogstens 0.05 te beperken?

17. Gegeven is een aselecte steekproef (12.05, 12.71, 12.25, 12.40, 12.15, 12.94, 12.00, 12.40, 12.49, 12.33, 12.37) van 11 waarnemingen van een normaal verdeelde stochast met onbekende verwachtingswaarde μ en (bekende) standaardafwijking $\sigma = 0.3$.

    (i) Bereken een betrouwbaarheidsinterval op level 95% voor μ.

    (ii) Bereken een linkséénzijdig betrouwbaarheidsinterval op level 90% voor μ.

    (iii) Vergelijk het betrouwbaarheidsinterval uit (i) met het betrouwbaarheidsinterval op level 95% bij onbekende standaardafwijking σ.

18. Een onderzoek naar het atoomgewicht van thallium leverde de volgende waarden op: 203.628, 203.636, 203.639, 203.644, 203.650, 203.666.

    (i) Bereken een betrouwbaarheidsinterval van level 95% voor het atoomgewicht.

    (ii) Hoeveel waarnemingen moeten er extra worden gedaan om op level 95% het atoomgewicht met een nauwkeurigheid van 0.002 te kunnen bepalen?

19. Iemand werpt 600 keer met een dobbelsteen en vindt 70 keer een `6`. Geef een betrouwbaarheidsinterval op level 95% voor de kans op een `6` bij deze dobbelsteen. Doe hetzelfde voor de levels 99% en 99.9%. Lijkt je dit een eerlijke dobbelsteen?

Les 4 Toetsen van hypothesen

We hebben tot nu toe enigszins algemeen naar grootheden van populaties gekeken en bediscussieerd hoe we deze grootheden uit steekproeven kunnen schatten. Vaak hebben we echter redelijk concrete voorstellingen over de verwachte waarde van een zeker parameter. In dit geval kan het resultaat van een steekproef onze verwachting over de parameter bevestigen of aanduiden dat onze verwachting onjuist was. Vaak wordt deze situatie door het opstellen van een *hypothese* gerealiseerd en een steekproef kan wel of niet evidentie voor het verwerpen van de hypothese geven. We zullen zien dat het toetsen van een hypothese min of meer een herformulering van de ideeën achter intervalschatters en betrouwbaarheidsintervallen zijn.

4.1 Hypothesen

In een hypothese maken we een uitspraak over een eigenschap van een stochast, bijvoorbeeld over de verwachtingswaarde. Hiervoor geven we aan dat een parameter θ waarvan de kansverdeling van de stochast afhangt een zekere waarde heeft. Vervolgens proberen we aan de hand van een steekproef voor de stochast evidentie voor of tegen de hypothese te vinden. Als we bijvoorbeeld de hypothese hebben dat de gemiddelde Nederlander $180cm$ groot is, dan geeft een (aselecte) steekproef van 1000 Nederlanders met een steekproefgemiddelde van $190cm$ hier sterke evidentie tegen, terwijl een steekproefgemiddelde van $181cm$ dit niet doet.

Hypothesen worden altijd in paren bekeken:

(i) De *nulhypothese* H_0 zegt dat een parameter θ een zekere waarde θ_0 heeft.

(ii) De *alternatieve hypothese* H_1 of H_a zegt dat de parameter θ van θ_0 afwijkt.

In het eenvoudigste geval zien de hypothesen er dus als volgt uit:

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0.$$

In het voorbeeld van de gemiddelde grootte houdt de alternatieve hypothese de mogelijkheden in, dat de gemiddelde Nederlander (duidelijk) groter of kleiner is dan $180cm$. Dit geval leidt tot een tweezijdige toets.

Vaak is men echter alleen maar geïnteresseerd of een parameter in een zekere richting van de nulhypothese afwijkt. Bijvoorbeeld wil een sporter weten of hij door een nieuwe training methode (of door een nieuw dopingmiddel) harder kan lopen dan eerder. In dit geval zijn de hypothesen

$$H_0 : \theta \leq \theta_0 \qquad H_1 : \theta > \theta_0$$

en dit geeft aanleiding tot een *rechtséénzijdige* toets. Analoog zal men met een *linkséénzijdige* toets de hypothesen

$$H_0 : \theta \geq \theta_0 \qquad H_1 : \theta < \theta_0$$

testen.

Een *toets* is nu een procedure die op grond van een steekproef de beslissing neemt om de nulhypothese te verwerpen of niet. Hierbij kunnen er twee fouten gemaakt worden:

I: De nulhypothese wordt verworpen terwijl hij juist is. Dit heet een *type I fout* of een *fout van de eerste soort*. De kans α op een type I fout heet de *onbetrouwbaarheid* (of *onbetrouwbaarheidsdrempel*) van de toets.

II: De nulhypothese wordt niet verworpen terwijl hij onjuist is. Dit heet een *type II fout* of een *fout van de tweede soort*. De kans β op een type II fout levert het *onderscheidingsvermogen* (power) $1-\beta$ van de toets.

We kunnen deze terminologie in het volgende schema weergeven:

	H_0 is juist	H_0 is onjuist
H_0 niet verwerpen	juiste beslissing kans $1-\alpha$	type II fout kans β
H_0 verwerpen	type I fout kans α	juiste beslissing kans $1-\beta$

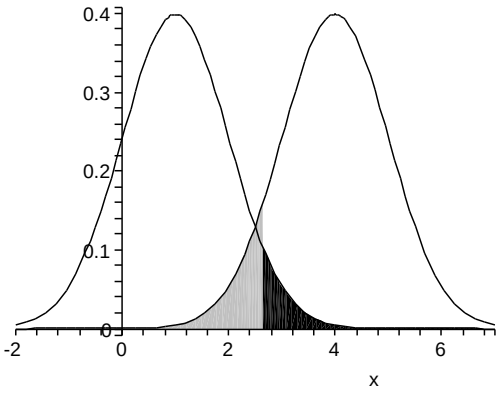
Het is natuurlijk heel eenvoudig, de kans op een type I fout te minimaliseren door de nulhypothese bijna nooit te verwerpen. Maar dit betekent dat veel resultaten van steekproeven als niet strijdig met H_0 geaccepteerd worden die eigenlijk evidentie voor de alternatieve hypothese geven. In dit geval is dus de kans op een type II fout hoog en het onderscheidingsvermogen van de toets slecht.

Merk op dat het onderscheidingsvermogen $1-\beta$ van een toets alleen bepaald kan worden als de alternatieve hypothese $H_1 : \theta \neq \theta_0$ vervangen wordt door een concrete alternatieve hypothese

$$H_1 : \theta = \theta_1.$$

Vaak worden toetsen vergeleken, door bij een vaste onbetrouwbaarheid α naar het onderscheidingsvermogen te kijken. De betere toets heeft dan het hogere onderscheidingsvermogen. Men kan ook het onderscheidingsvermogen $1-\beta$ als functie van de onbetrouwbaarheid opvatten, dit geeft de zogeheten *operating characteristic*. (Let wel: Er zijn ongeveer zo veel definities van operating characteristic als er auteurs zijn, maar de achterliggende gedachten zijn hetzelfde.) Een ideale toets zou al voor zeer kleine waarden van α naar een onderscheidingsvermogen $1-\beta$ dicht bij 1 stijgen.

In Figuur 16 is het concept van type I en type II fouten geïllustreerd. We kijken hierbij naar de nulhypothese $H_0 : \theta = 1$ en kiezen een onbetrouwbaarheid α van $\alpha = 0.05$. Het zwarte gebied onder de linker normale verdeling heeft juist de oppervlakte 0.05, dus leiden steekproefwaarden $\overline{\theta}$ die in dit gebied vallen tot verwerpen van de nulhypothese. Als we als alternatieve hypothese $H_1 : \theta = 4$ nemen, dan is de kans op een type II fout de oppervlakte onder de rechter normale verdeling, waar we de nulhypothese niet verwerpen, dus het grijze gebied. In het voorbeeld is deze oppervlakte ongeveer 0.0877, dus is het onderscheidingsvermogen van deze toets ongeveer 92.3%.



Figuur 16: Gebieden voor type I (zwart) en type II fouten (grijs).

4.2 Toetsen en betrouwbaarheidsintervallen

Aan de hand van het begrip van een type I fout kunnen we nu een verband leggen tussen toetsen en betrouwbaarheidsintervallen. We hadden een betrouwbaarheidsinterval op level γ rond een schatting $\overline{\theta}$ van een parameter zo gekozen, dat het interval precies de waarden van θ bevat zo dat $\overline{\theta}$ in een interval rond θ met kansmassa γ ligt.

Deze aanpak kunnen we nu omdraaien om een toets met onbetrouwbaarheid $\alpha = 1 - \gamma$ te krijgen: Voor de nulhypothese $H_0 : \theta = \theta_0$ kiezen we een interval $[\theta_-, \theta_+]$ rond θ_0 zo dat onder de aanname dat H_0 juist is de kans op een steekproefwaarde $\overline{\theta}$ buiten dit interval hoogstens α is, dus

$$P(\theta_- \leq \overline{\theta} \leq \theta_+) = 1 - \alpha = \gamma.$$

Als de schatting $\overline{\theta}$ buiten het interval $[\theta_-, \theta_+]$ ligt, wordt dit als evidentie *tegen* de nulhypothese H_0 beschouwd omdat dit slechts met de (kleine) kans α gebeurt en de nulhypothese wordt daarom verworpen.

Bij éénzijdige toetsen is het interval $[\theta_-, \theta_+]$ aan een kant open, omdat we de nulhypothese alleen maar bij afwijking in één richting verwerpen. Bij een rechtséénzijdige toets wordt H_0 verworpen, als $\overline{\theta}$ buiten het interval $[-\infty, \theta_+]$ ligt, en bij een linkséénzijdige toets, als $\overline{\theta}$ buiten het interval $[\theta_-, \infty]$ ligt.

Merk op: Het lijkt op het eerste gezicht verwarrend, dat bij een rechtséénzijdige toets het interval $[-\infty, \theta_+]$ waarvoor we de nulhypothese niet verwerpen naar *links* open is, terwijl het rechtséénzijdige betrouwbaarheidsinterval voor een schatting naar *rechts* open is. Maar dit schijnbare paradox maakt juist het verband tussen toetsen en betrouwbaarheidsintervallen duidelijk:

Het betrouwbaarheidsinterval op level $\gamma = 1 - \alpha$ rond een schatting $\overline{\theta}$ bevat precies de waarden θ_0 waarvoor $\overline{\theta}$ bij een toets met onbetrouwbaarheid α geen aanleiding geeft tot verwerpen van de nulhypothese $\theta = \theta_0$.

Andersom: Een toets met onbetrouwbaarheid α verwerpt de nulhypothese $H_0 : \theta = \theta_0$ op grond van de schatting $\overline{\theta}$ dan en slechts dan als θ_0 buiten het betrouwbaarheidsinterval van level $1 - \alpha$ rond $\overline{\theta}$ valt.

Toetsen voor gemiddelden

In de meeste situaties zal onder de voorwaarde dat de nulhypothese juist is de schatter T voor de schattingen $\overline{\theta}$ een normale verdeling met gemiddelde θ_0 en variantie $\frac{\sigma^2}{n}$ hebben. Dit is in het bijzonder het geval als T de schatter voor het gemiddelde van een normale verdeling is, maar bij benadering ook voor de schatter van het gemiddelde van niet-normale verdelingen (als n niet te klein is). In dit geval weten we dat de stochast

$$Z := \frac{T - \theta_0}{\frac{\sigma}{\sqrt{n}}} = \frac{(T - \theta_0)\sqrt{n}}{\sigma}$$

standaard-normaal verdeeld is en we kunnen daarom net zo als bij de betrouwbaarheidsintervallen met behulp van de z-waarden makkelijk een interval aangeven, dat een *tweezijdige toets* met onbetrouwbaarheid α oplevert, want er geldt

$$P\left(\theta_0 - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq T \leq \theta_0 + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

We zullen bij deze toets de nulhypothese dus verwerpen als de schatting $\overline{\theta}$ meer dan $z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ van θ_0 afwijkt, dus als

$$|\overline{\theta} - \theta_0| > z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

Dit zou namelijk onder de aanname van H_0 slechts met kans α gebeuren en omdat de kans α laag is, geeft dit evidentie tegen H_0. De kans α dat de beslissing om H_0 te verwerpen onjuist is, is juist de kans op een type I fout.

Merk op: De foutmarge rond θ_0 die we toelaten zonder H_0 te verwerpen is precies hetzelfde als de foutmarge die we voor het betrouwbaarheidsinterval rond $\overline{\theta}$ hebben gekozen. Dit is geen toeval, omdat de definitie van een toets met onbetrouwbaarheid α in principe alleen maar een herformulering van de definitie van een betrouwbaarheidsinterval van level $1 - \alpha$ is.

Als we een *rechtséénzijdige toets* met onbetrouwbaarheid α willen hebben, moeten we een interval $[-\infty, \theta_+]$ vinden zo dat $P(T > \theta_+) = \alpha$. Maar omdat

$$P\left(T \leq \theta_0 + z_\alpha\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

is $[-\infty, \theta_0 + z_\alpha\frac{\sigma}{\sqrt{n}}]$ zo'n interval en we verwerpen $H_0 : \theta \leq \theta_0$ als

$$\overline{\theta} > \theta_0 + z_\alpha\frac{\sigma}{\sqrt{n}}.$$

Analoog krijgen we een *linkséénzijdige toets* met onbetrouwbaarheid α door H_0 te verwerpen als

$$\overline{\theta} < \theta_0 - z_\alpha\frac{\sigma}{\sqrt{n}},$$

want $P(T < \theta_0 - z_\alpha \frac{\sigma}{\sqrt{n}}) = \alpha$, of te wel

$$P\left(T \geq \theta_0 - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Voorbeeld: Een eierhandelaar koopt een grote partij eieren van een kippenfokker. We mogen aannemen dat het gewicht X van de eieren in een homogene partij normaal verdeeld is en dat de standaardafwijking van de gewichten $6g$ is. De fokker garandeert dat het gemiddelde van de eieren in deze partij boven de $60g$ ligt. De handelaar neemt nu een steekproef van 5 eieren en constateert dat deze samen $275g$ wegen. Hij wil de levering alleen maar reclameren als hij de nulhypothese $H_0 : \mu = 60$ op een onbetrouwbaarheidslevel van $\alpha = 0.05$ kan verwerpen. Omdat hij natuurlijk alleen maar bij te lichte eieren gaat reclameren, past hij een linkséénzijdige toets toe. Er geldt $z_{0.05} = 1.6449$ en dus zal hij de nulhypothese verwerpen, als zijn schatting $\overline{\mu}$ voldoet aan $\overline{\mu} < 60 - z_{0.05}\frac{6}{\sqrt{5}} \approx 55.6$. Maar zijn steekproef geeft $\overline{\mu} = \frac{275}{5} = 55$, dus zal hij inderdaad reclameren.

Aanpassingen bij kleine steekproeven

We zijn er tot nu toe van uit gegaan dat de schatter T voor de schattingen $\overline{\theta}$ de variantie $\frac{\sigma^2}{n}$ heeft. Vaak is de hiervoor benodigde variantie σ^2 van de onderliggende kansverdeling echter onbekend, in dit geval wordt de variantie $\frac{\sigma^2}{n}$ vervangen door de schatting $\frac{s^2}{n}$, waarbij s^2 de steekproefvariantie is. Maar het vervangen van σ^2 door de schatting s^2 leidt ertoe dat de getransformeerde stochast

$$\frac{(T - \theta_0)\sqrt{n}}{s}$$

geen normale verdeling maar een Student-t verdeling met $n-1$ vrijheidsgraden heeft. We moeten dus de z-waarden in de boven aangegeven intervallen voor de verschillende toetsen vervangen door de t-waarden van de Student-t verdeling, net zo als bij de betrouwbaarheidsintervallen. We krijgen dus een tweezijdige toets met onbetrouwbaarheid α door de nulhypothese H_0 te verwerpen als

$$|\overline{\theta} - \theta_0| > t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

Bij de rechts- en linkséénzijdige toetsen zijn de criteria voor het verwerpen van de nulhypothese analoog

$$\overline{\theta} > \theta_0 + t_{n-1,\alpha}\frac{s}{\sqrt{n}} \quad \text{en} \quad \overline{\theta} < \theta_0 - t_{n-1,\alpha}\frac{s}{\sqrt{n}}.$$

Als n groot is (meestal wordt hier $n \geq 50$ als vuistregel gehanteerd), ligt de Student-t verdeling met $n-1$ vrijheidsgraden zo dicht bij de standaard-normale verdeling, dat deze correctie verwaarloosd kan worden omdat dan $z_\alpha \approx t_{n-1,\alpha}$ is. Maar bij onbekende variantie σ^2 *en* kleine steekproeven moeten de toetsen inderdaad zo als aangegeven aangepast worden.

Toetsen voor relatieve frequenties

Stel we willen de hypothese toetsen dat defecte stukken bij een productie met kans p_0 optreden, dus dat de parameter p van een binomiale verdeling gelijk is aan p_0. Hiervoor tellen we met de stochast X het aantal k van successen bij n pogingen en krijgen hiermee de schatting $\overline{p} = \frac{k}{n}$ voor p. We weten dat bij een niet te kleine steekproef ($np_0 \geq 5$, $n(1-p_0) \geq 5$) de stochast

$$Z := \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

bij benadering standaard-normaal verdeeld is. Omdat net als boven $P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1-\alpha$, geldt

$$P\left(np_0 - z_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)} \leq Z \leq np_0 + z_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)}\right) = 1-\alpha,$$

dus zullen we bij een tweezijdige toets met onbetrouwbaarheid α de nulhypothese $H_0 : p = p_0$ verwerpen als bij een steekproef met k successen in n pogingen geldt dat

$$|k - np_0| > z_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)}.$$

Als we beide zijden door n delen, kunnen we dit ook rechtstreeks als criterium voor de relatieve frequenties formuleren, we verwerpen de nulhypothese als

$$|\overline{p} - p_0| > z_{\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0)}{n}}.$$

De rechts- en linkséénzijdige toetsen kunnen we inmiddels zonder na te denken afleiden, we verwerpen bij de relatieve frequenties de nulhypothese H_0 als

$$\overline{p} > p_0 + z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} \text{ (rechts) of } \overline{p} < p_0 - z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} \text{ (links)}.$$

Voorbeeld: Een handelaar verkoopt een grote partij goederen en deelt de koper mee dat er hoogstens 5% ondeugdelijke exemplaren in zitten. Om dit te verifiëren neemt de koper een steekproef van 150 stuks. Hij zal reclameren als hij op een onbetrouwbaarheidslevel van $\alpha = 0.05$ de bewering van de handelaar kan verwerpen. Omdat $0.05 \cdot 150 = 7.5 > 5$, kunnen we de normale benadering van de binomiale verdeling toepassen. Te koper zal natuurlijk alleen maar bij een te hoog aantal ondeugdelijke exemplaren reclameren, daarom moeten we een rechtséénzijdige toets toepassen. Er geldt $z_{0.05} = 1.6449$, $n = 150$ en $p_0 = 0.05$, dus is $z_\alpha\sqrt{np_0(1-p_0)} \approx 4.39$, de koper zal dus vanaf $7.5 + 4.39$, dus vanaf 12 ondeugdelijke stukken reclameren.

Als een steekproef te klein is om de normale benadering toe te passen, is het meestal mogelijk de kans op een steekproef met k of meer successen expliciet met de binomiale verdeling te berekenen, namelijk door

$$P(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}.$$

Bij een rechtséénzijdige toets wordt H_0 verworpen als $P(X \geq k) < \alpha$. Analoog berekent men met

$$P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

de kans op een steekproef met hoogstens k successen en verwerpt bij een links-éénzijdige toets de nulhypothese als $P(X \leq k) < \alpha$. Bij een tweezijdige toets hangt het criterium ervan af of $k > np_0$ of $k < np_0$. Als kans dat een steekproef zo sterk van p_0 afwijkt als $\overline{p}$ krijgt men in dit geval $2 \cdot \min(P(X \geq k), P(X \leq k))$ omdat ook met de afwijking in de andere richting rekening gehouden moet worden. Als criterium voor het verwerpen van de nulhypothese krijgt men zo

$$\min(P(X \geq k), P(X \leq k)) < \frac{\alpha}{2}.$$

Significantie en P-waarden

Als we een toets zo opzetten dat we de nulhypothese verwerpen als de schatting voor een parameter θ buiten het betrouwbaarheidsinterval van level $\gamma = 1 - \alpha$ rond de nulhypothese θ_0 ligt, dan noemen we α ook de *significantie level* van de toets. De significantie is dus gelijk aan de kans op een type I fout onder de aanname dat de nulhypothese juist is. We noemen een resultaat dus significant op level α als de kans dat de nulhypothese geldt, hoogstens α is. Het woord *significant* (van het Latijnse *signum* = teken) is gekozen om aan te duiden, dat het gevonden resultaat iets *betekent* en niet meer als toevallig beschouwd kan worden.

Soms ligt een schatting $\overline{\theta}$ veel verder af van de nulhypothese dan het betrouwbaarheidsinterval op de gekozen level α aangeeft. De schatting geeft dus zelfs op een hogere level nog evidentie tegen de nulhypothese. In dit geval kijkt men vaak naar de hoogste mogelijke waarde van α, zo dat de schatting nog net tot verwerpen van de nulhypothese zou leiden en noemt dit de *P-waarde* van de schatting:

De P-waarde p van een schatting $\overline{\theta}$ zegt dat steekproeven die zo ver of verder van de nulhypothese θ_0 afwijken als $\overline{\theta}$ onder de aanname van de nulhypothese slechts met kans p voorkomen.

De P-waarde van een schatting maakt dus een kwantitatieve uitspraak over de evidentie tegen de nulhypothese, terwijl een gewone toets met significantie niveau α alleen maar aangeeft of de evidentie sterker dan een gekozen level is of niet.

> Soms wordt de mate van significantie met zekere intervallen van P-waarden verbonden, men leest bijvoorbeeld aanduidingen zo als
>
> | | |
> |---|---|
> | zeer sterk significant: | $P < 0.001$ |
> | sterk significant: | $0.001 < P < 0.01$ |
> | zwak significant: | $0.01 < P < 0.05$ |

maar er bestaan geen conventies die enigszins uniform gehandhaafd worden.

4.3 Toetsen op verschillen tussen twee verdelingen

We hebben tot nu toe naar de situatie gekeken dat we een hypothese over een parameter van een kansverdeling hebben en deze hypothese met een steekproef willen toetsen. In de praktijk is echter vaak een iets andere vraag van belang, namelijk of een parameter bij twee verdelingen dezelfde waarde heeft, dus bijvoorbeeld of twee verdelingen hetzelfde gemiddelde hebben. In dit geval is het niet zo interessant wat de waarden van de gemiddelden zijn, maar alleen maar of hun verschil 0 is of niet.

In plaats van een enkele steekproef moeten we hier voor ieder van de twee verdelingen een aparte steekproef nemen, en de verdelingen van deze steekproeven worden door twee onafhankelijke schatters T_1 en T_2 beschreven. We gaan ervan uit dat T_1 een zuivere schatter voor de parameter θ_1 van de eerste verdeling is, en T_2 een zuivere schatter voor de parameter θ_2 van de tweede verdeling is en veronderstellen verder dat de varianties σ_1^2 en σ_2^2 van de twee verdelingen bekend zijn en we steekproeven van grootte n_1 en n_2 nemen. In dit geval geldt

$$E[T_1 - T_2] = \theta_1 - \theta_2 \quad \text{en} \quad Var(T_1 - T_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

De nulhypothese is dat de parameters θ_1 en θ_2 gelijk zijn, dus

$$H_0 : \theta_1 = \theta_2 \quad \text{of} \quad \theta_1 - \theta_2 = 0.$$

Als we weer veronderstellen dat T_1 en T_2 bij benadering normaal verdeeld zijn dan is

$$Z := \frac{(T_1 - T_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(bij benadering) een standaard-normale verdeling en we kunnen weer de z-waarden gebruiken om een toets te formuleren: Als de steekproef voor de eerste verdeling een schatting van $\overline{\theta_1}$ en de steekproef voor de tweede verdeling een schatting van $\overline{\theta_2}$ oplevert, dan wordt op een significantie niveau α de nulhypothese $\theta_1 = \theta_2$ verworpen als

$$|\overline{\theta_1} - \overline{\theta_2}| > z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Voorbeeld: Stel de normaal verdeelde stochast X heeft variantie $\sigma_X^2 = 0.09$ en de normaal verdeelde stochast Y heeft variantie $\sigma_Y^2 = 0.16$. Een steekproef van 9 stuk geeft een gemiddelde van $\overline{x} = 21.7$ voor X en een steekproef van 4 stuk geeft een gemiddelde van $\overline{y} = 21.2$ voor Y. Kunnen bij op een onbetrouwbaarheidslevel van $\alpha = 0.05$ de nulhypothese verwerpen dat X en Y hetzelfde gemiddelde hebben? Er geldt $z_{0.025} = 1.96$ en $\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}} = \sqrt{0.05}$, dus zullen we de nulhypothese inderdaad verwerpen omdat $|\overline{x} - \overline{y}| = 0.5 > 1.96 \cdot \sqrt{0.05} \approx 0.44$.

Ook éénzijdige toetsen spelen hier weer een belangrijke rol, bijvoorbeeld wil men aantonen dat een nieuwe medicijn beter is dan een oude. Als de parameter θ_1 de oude en de parameter θ_2 de nieuwe medicijn beschrijft, is de nulhypothese $H_0 : \theta_2 \leq \theta_1$ en men probeert met een rechtséénzijdige toets evidenties ervoor te vinden om deze hypothese te verwerpen, dus $\theta_2 > \theta_1$ te ondersteunen. Met dezelfde redeneringen die we eerder hebben toegepast, geeft dit op het significantie niveau α het criterium

$$\overline{\theta_2} - \overline{\theta_1} > z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

om de nulhypothese te verwerpen. De schatting voor het verschil tussen de nieuwe en oude medicijn moet dus een zekere marge overschrijden om een toevallig effect met hoge kans uit te sluiten.

Aanpassingen bij kleine steekproeven

We zijn weer ervan uitgegaan dat de varianties σ_1^2 en σ_2^2 van de twee onderliggende verdelingen bekend zijn. Als dit niet het geval is moeten we net als bij de toetsen voor een enkele verdeling de varianties door de geschatte steekproefvarianties s_1^2 en s_2^2 vervangen. Het probleem is, dat de verdeling van

$$T := \frac{(T_1 - T_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

geen Student-t verdeling meer is en we dus niet zonder meer met de t-waarden kunnen werken. Maar gelukkig laat zich de verdeling van T wel door een Student-t verdeling benaderen, alleen moet men hiervoor nog een geschikt aantal ν van vrijheidsgraden bepalen.

> Men kan inzien, dat het aantal vrijheidsgraden groter dan het minimum van $n_1 - 1$ en $n_2 - 1$ moet zijn, omdat dit de vrijheidsgraden voor de aparte stochasten T_1 en T_2 zijn. Aan de andere kant kan het aantal vrijheidsgraden ook niet groter dan $n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$ zijn, want dit zou men bij samenvoegen van de twee steekproeven krijgen. Als men aan de conservatieve kant zit en de nulhypothese niet te snel wil verwerpen, is $\nu := \min(n_1 - 1, n_2 - 1)$ een mogelijke keuze voor het aantal vrijheidsgraden. Maar meestal wordt het aantal vrijheidsgraden uit de grootten van de steekproeven en de steekproefvarianties berekend, bijvoorbeeld door
>
> $$\nu := \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1 - 1}\frac{s_1^2}{n_1} + \frac{1}{n_2 - 1}\frac{s_2^2}{n_2}}$$

De situatie is iets eenvoudiger en overzichtelijker als bekend is dat de twee verdelingen dezelfde (onbekende) variantie hebben. In dit geval noemt men het gewogen gemiddelde

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

van de steekproefvarianties de *gepoolde variantie* van de twee steekproeven.

> Het idee achter de gepoolde variantie is, de twee steekproeven samen te vatten en uit de verzamelde waarden een schatting voor de variantie te maken. Stel X en Y zijn stochasten met dezelfde variantie σ^2. Voor een steekproef van grootte n_1 is $S_1^2 := \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2$ een zuivere schatter voor σ^2 en net zo is $S_2^2 := \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_i-\overline{Y})^2$ een zuivere schatter voor σ^2. Hieruit volgt, dat $(n_1-1)S_1^2+(n_2-1)S_2^2$ een zuivere schatter voor $(n_1+n_2-2)\sigma^2$ is, en dus is
>
> $$S^2 := \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \text{ een zuivere schatter voor } \sigma^2.$$
>
> De gepoolde variantie is dus juist de realisering van deze zuivere schatter voor σ^2 op twee concrete steekproeven.

Het voordeel van de gepoolde variantie is, dat men hiermee weer naar een Student-t verdeling met een bekend aantal vrijheidsgraden komt, er geldt namelijk dat

$$T := \frac{(T_1-T_2)-(\theta_1-\theta_2)}{\sqrt{\frac{s^2}{n_1}+\frac{s^2}{n_2}}} = \frac{(T_1-T_2)-(\theta_1-\theta_2)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

een Student-t verdeling met n_1+n_2-2 vrijheidsgraden is.

Een tweezijdige toets zou in deze situatie de nulhypothese $H_0 : \theta_1 = \theta_2$ verwerpen als

$$|\overline{\theta_1}-\overline{\theta_2}| > t_{n_1+n_2-2,\frac{\alpha}{2}} \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

> De vraag of de aanname dat twee steekproeven uit verdelingen met dezelfde variantie $\sigma_1^2 = \sigma_2^2 = \sigma^2$ komen juist is, kan zijnerzijds ook weer met een toets onderzocht worden. Hiervoor kijkt men naar het quotiënt $\frac{\sigma_1^2}{\sigma_2^2}$, waarvoor $\frac{s_1^2}{s_2^2}$ een schatting is en de verdeling van deze schattingen heet de F-verdeling. De nulhypothese is $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ en de zogeheten F-toets geeft aan, wanneer H_0 op een zekere onbetrouwbaarheidslevel moet worden verworpen. In dit college gaan we de F-toets echter niet behandelen.

Verschillen tussen relatieve frequenties

De ideeën die we net hebben bediscussieerd, kunnen we ook toepassen op de vraag of twee relatieve frequenties significant verschillen. Als P_1 een zuivere schatter voor de relatieve frequentie p_1 is en P_2 een zuivere schatter voor de relatieve frequentie p_2, dan is $P_1 - P_2$ een schatter met verwachtingswaarde $E[P_1 - P_2] = p_1 - p_2$ en met variantie $Var(P_1 - P_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$, waarbij n_1 en n_2 de grootten van de steekproeven zijn.

Als we willen laten zien, dat de twee relatieve frequenties verschillend zijn, is de nulhypothese natuurlijk dat p_1 en p_2 gelijk zijn, dus

$$H_0 : p_1 = p_2.$$

Onder de aanname dat de nulhypothese juist is, is dus

$$Var(P_1 - P_2) = p_1(1-p_1)(\frac{1}{n_1} + \frac{1}{n_2}) = p_2(1-p_2)(\frac{1}{n_1} + \frac{1}{n_2}).$$

Omdat we niet ervan kunnen uitgaan dat p_1 of p_2 bekend is, moeten we hier weer een schatting invullen, en hiervoor nemen we de schatting p_0 die we uit de combinatie van de twee steekproeven krijgen, dus

$$p_0 := \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}.$$

Als de steekproeven niet te klein zijn (dus weer $n_1 p_1 \geq 5$ en $n_2 p_2 \geq 5$, d.w.z. in ieder steekproef hebben we minsten 5 successen) is de stochast

$$Z := \frac{P_1 - P_2}{\sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

bij benadering standaard-normaal verdeeld en we kunnen hiermee weer met behulp van de z-waarden tweezijdige en éénzijdige toetsen formuleren.

Als we de schattingen $\overline{p_1}$ en $\overline{p_2}$ voor de relatieve frequenties in de twee steekproeven vinden, zullen we bij een tweezijdige toets de nulhypothese $H_0 : p_1 = p_2$ verwerpen als

$$|\overline{p_1} - \overline{p_2}| > z_{\frac{\alpha}{2}} \sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}.$$

Bij een éénzijdige toets krijgen we analoog, dat we de nulhypothese moeten verwerpen als

$$\overline{p_2} - \overline{p_1} > z_\alpha \sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})} \quad \text{of} \quad \overline{p_1} - \overline{p_2} < z_\alpha \sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}$$

afhankelijk ervan of we willen aantonen dat p_2 groter of kleiner is dan p_1.

Belangrijke begrippen in deze les

- nulhypothese, alternatieve hypothese
- toets (tweezijdig, éénzijdig)
- onbetrouwbaarheid van een toets
- onderscheidingsvermogen van een toets
- type I fout, type II fout

- significantie
- P-waarde
- aanpassingen bij kleine steekproeven
- gepoolde variantie

Opgaven

20. Men past op elk van twee (aselecte, onafhankelijke) steekproeven een toets met onbetrouwbaarheid α toe. Hoe groot moet α worden gekozen zo dat de kans dat minstens één van de nulhypothesen ten onrechte wordt verworpen hoogstens 10% is?

21. Het gewicht van sinaasappels was tot nu toe normaal verdeeld met gemiddelde $\mu_0 = 50g$ en standaardafwijking $\sigma = 2g$. Van een nieuwe goedkopere behandeling van de sinaasappelbomen wordt beweerd dat ze minstens even zware vruchten oplevert. Een kweker wil deze bewering toetsen tegen het alternatief dat $\mu < 50g$ (waarbij de standaardafwijking onveranderd blijft). De sinaasappels in een steekproef van 100 stuks hebben een gemiddeld gewicht van $49.65g$. Heeft de kweker op een onbetrouwbaarheidslevel van $\alpha = 0.05$ reden om de nieuwe methode niet toe te passen?

22. Zij X een normaal verdeelde stochast met standaardafwijking $\sigma = 10$ en onbekende gemiddelde μ. Op grond van een steekproef willen we de hypothese $H_0 : \mu = 50$ rechtséénzijdig toetsen met onbetrouwbaarheid $\alpha = 0.05$. We eisen daarbij dat het onderscheidingsvermogen bij de alternatieve hypothese $H_1 : \mu = 52$ gelijk aan 90% moet zijn.

    (i) Hoe groot moet de steekproef minstens zijn?

    (ii) Hoe groot is bij de steekproefgrootte uit (i) het onderscheidingsvermogen bij de alternatieve hypothese $\mu = 51$?

23. In een fabriek staan 2 vulmachines, A en B, waarmee flessen worden gevuld. Bij een juiste instelling van de machines is de inhoud van de flessen normaal verdeeld met een gemiddelde van $250g$. De standaardafwijking is onafhankelijk van de instelling steeds $2.5g$. Om na te gaan of de machines goed zijn ingesteld wordt voor elke machine de inhoud van 4 flessen nauwkeurig bepaald. De gemiddelde inhoud voor flessen van machine A bedraagt $251.68g$ en $252.68g$ voor flessen van machine B.

    (i) Toets met onbetrouwbaarheid $\alpha = 0.05$ of de machines A en B op het juist vulgewicht van $250g$ ingesteld zijn.

    (ii) Toets met onbetrouwbaarheid $\alpha = 0.05$ of de instellingen van de machines A en B onderling verschillen.

24. Een examen bestaat uit 20 vragen met telkens 4 mogelijke antwoorden. De kandidaten zijn geslaagd als op minstens 10 vragen het juist antwoord is gekozen. Beschouw het tentamen als een statistische toets.

    (i) Formuleer een nulhypothese H_0 en een alternatieve hypothese H_1.

    (ii) Definieer de grootheid die voor de toets uit de steekproef bepaald wordt en bepaal de kansverdeling van deze grootheid onder de aanname van H_0.

(iii) Bereken de onbetrouwbaarheid van de toets.

(iv) Bereken het onderscheidingsvermogen van de toets als de kans op het geven van het juiste antwoord door een kandidaat per vraag $\frac{1}{2}$ is.

25. Uit een baal katoen werd een aselecte steekproef genomen van 4000 draden om de vezellengte te bepalen. De gemiddelde lengte was $2.33cm$ en de standaardafwijking $0.48cm$. Uit dezelfde baal werd een andere steekproef genomen van 200 draden volgens een andere methode dan de eerste. Van deze tweede steekproef was de gemiddelde vezellengte $2.54cm$. Aangenomen mag worden dat de vezellengte normaal verdeeld is. Toets met onbetrouwbaarheid $\alpha = 0.05$ of er verschil is tussen de twee steekproefmethoden.

26. Een fabrikant betrekt al jaren transistoren van A, die hem gemiddeld 8% kapotte levert. Van een vertegenwoordiger van B koopt hij 75 stuks die wat duurder zijn, maar waarvan beweerd wordt dat er minder kapot zijn. Bij controle blijken 5 van deze 75 transistoren ondeugdelijk te zijn. Zijn de percentages kapotte exemplaren in de producten van A en B op een significantie niveau van $\alpha = 0.05$ verschillend?

27. Een medicus beweert dat de kans op een jongensgeboorte groter is dan die op de geboorte van een meisje. Hij komt tot deze conclusie omdat 51% van de pasgeboren babies uit zijn praktijk jongens zijn. Hoeveel geboorten moeten dat zijn om deze conclusie of een onbetrouwbaarheidslevel van $\alpha = 0.05$ te rechtvaardigen?

Les 5 Vergelijken van verdelingen

In de vorige les hebben we naar toetsen voor hypothesen gekeken, waarbij de hypothese een uitspraak over een parameter van een kansverdeling was, bijvoorbeeld over het gemiddelde of een relatieve frequentie. Maar als we bijvoorbeeld willen toetsen, of een dobbelsteen eerlijk is, zullen we na 120 worpen niet alleen maar het gemiddelde en de variantie bepalen, maar kijken of de getallen 1 t/m 6 alle ongeveer 20 keer gevallen zijn. Op deze manier zouden we natuurlijk onmiddellijk zien, dat de stochast X met

$$P(X=1)=\frac{5}{24},\quad P(X=2)=\frac{1}{6},\quad P(X=3)=\frac{1}{12},$$
$$P(X=4)=0,\quad P(X=5)=\frac{13}{24},\quad P(X=6)=0$$

geen eerlijke dobbelsteen beschrijft, terwijl $E[X]=3\frac{1}{2}$ en $Var(X)=\frac{35}{12}$, net zo als bij een eerlijke dobbelsteen (ga dit na). We zouden dus met een toets op het gemiddelde of de variantie niet aan het licht kunnen brengen dat de dobbelsteen oneerlijk is, maar natuurlijk zouden we dit ook niet op zo'n stomme manier proberen.

De dobbelsteen is een voorbeeld van een verdeling, waar we niet alleen maar een parameter van de kansverdeling willen toetsen, maar de volledige verdeling willen bekijken. De nulhypothese, die we in dit geval zouden toetsen is

$$H_0 : P(X=1)=\frac{1}{6},\ldots,P(X=6)=\frac{1}{6}$$

en de alternatieve hypothese luidt, dat niet alle van deze kansen gelijk aan $\frac{1}{6}$ zijn. Natuurlijk kunnen we niet verwachten, dat we bij een steekproef precies de kansen van de nulhypothese vinden, maar naarmate de steekproef groter wordt, zouden we steeds kleinere afwijkingen verwachten. Het vergelijken van de onder de nulhypothese verwachte aantallen en de daadwerkelijk waargenomen aantallen geeft aanleiding tot een belangrijke klasse van toetsen voor hypothesen over kansverdelingen, namelijk de χ^2-toetsen.

5.1 De χ^2-aanpassingstoets

De situatie die we nu bekijken is als volgt: Gegeven is een stochast X met een zekere kansverdeling, bijvoorbeeld de uniforme verdeling voor een eerlijke dobbelsteen. De nulhypothese is, dat een steekproef door de stochast X is voortgebracht en we willen toetsen of deze hypothese plausibel is.

De algemene aanpak is, de mogelijke uitkomsten van de stochast X in een aantal klassen in te delen. Voor een stochast met een discrete kansverdeling zijn de klassen vaak de verschillende mogelijke uitkomsten, maar soms is het handig verschillende uitkomsten in één klasse samen te vatten.

Voor continue kansverdelingen kiest men als klassen meestal intervallen, deze zijn vaak van dezelfde breedte, maar dit is niet noodzakelijk zo.

Voorbeeld: Voor een stochast $X \in \mathcal{N}(\mu,\sigma^2)$ waarvoor men een normale verdeling met verwachtingswaarde μ en variantie σ^2 verondersteld, worden de

intervalgrenzen vaak op veelvouden van de standaardafwijking σ gelegd. Men krijgt zo bijvoorbeeld de klassen

$$\begin{array}{ll} K_1 : -\infty < X < \mu - 3\sigma, & K_2 : \mu - 3\sigma \leq X < \mu - 2\sigma, \\ K_3 : \mu - 2\sigma \leq X < \mu - \sigma, & K_4 : \mu - \sigma \leq X < \mu, \\ K_5 : \mu < X \leq \mu + \sigma, & K_6 : \mu + \sigma \leq X < \mu + 2\sigma, \\ K_7 : \mu + 2\sigma \leq X < \mu + 3\sigma, & K_8 : \mu + 3\sigma \leq X < \infty \end{array}$$

Als de mogelijke uitkomsten van X in k klassen ingedeeld zijn, wordt voor elke van de klassen de kans p_i bepaalt, dat X een uitkomst in de i-de klasse produceert. Bij een steekproef van n stuks zullen we dan np_i waarden in de i-de klasse verwachten.

In het voorbeeld van de normale verdeling met 8 klassen kunnen we uit de standaard-normale verdeling de volgende kansen afleiden:

i	1	2	3	4	5	6	7	8
p_i	0.0013	0.0214	0.1359	0.3413	0.3413	0.1359	0.0214	0.0013

We beschrijven nu met een stochast X_i het aantal uitkomsten in een steekproef van n stuks, die in de i-de klasse vallen. Uit de verschillen van X_i en np_i moeten we nu een toets afleiden, die aangeeft of het plausibel is dat de steekproef volgens de veronderstelde kansverdeling is voortgebracht.

Voor het speciaal geval van slechts 2 klassen hebben we dit probleem al eerder bekeken, in dit geval vallen de uitkomsten met kans p in de eerste klasse en met kans $q = 1-p$ in de tweede klasse. Maar dit betekent, dat X de stochast van een Bernoulli-experiment met kans p is en de stochast X_1 is binomiaal verdeeld met parameters n en p. De relatieve frequentie p van een binomiale verdeling hadden we in de vorige les getoetst, door X_1 op een (bij benadering) standaard-normale verdeling te transformeren, namelijk door

$$Z := \frac{X_1 - np}{\sqrt{np(1-p)}}.$$

Als Z standaard-normaal verdeeld is, heeft Z^2 een χ^2-verdeling met 1 vrijheidsgraad en we kunnen Z^2 als volgt herschrijven:

$$\begin{aligned} Z^2 &= \frac{(X_1 - np)^2}{np(1-p)} = (1-p)\,\frac{(X_1 - np)^2}{np(1-p)} + p\,\frac{(X_1 - np)^2}{np(1-p)} \\ &= \frac{(X_1 - np)^2}{np} + \frac{((n - X_1) - n(1-p))^2}{n(1-p)} \\ &= \frac{(X_1 - np)^2}{np} + \frac{(X_2 - nq)^2}{nq}. \end{aligned}$$

We zien dus dat we Z^2 kunnen beschrijven als som van de kwadratische afwijkingen tussen waargenomen aantallen en verwachte aantallen, genormeerd op de verwachte aantallen.

In plaats van de waarde van Z met de z-waarden van de standaard-normale verdeling te vergelijken, kunnen we de waarde van Z^2 tegen de waarden χ^2_α van een χ^2-verdeling met 1 vrijheidsgraad toetsen die gedefinieerd zijn door

$$P(Z^2 > \chi^2_\alpha) = \alpha$$

want er geldt $P(Z^2 > \chi^2_\alpha) = P(Z > z_\alpha) = \alpha$.

De veralgemening van 2 tot k klassen is nu enigszins voor de hand liggend: De gekwadrateerde afwijkingen van de waargenomen aantallen van de verwachte aantallen worden door de verwachte aantallen gedeeld en deze hoeveelheden worden voor de verschillende klassen bij elkaar opgeteld. Het idee achter de normering op het aantal verwachte uitkomsten in een klasse is dat bij een verwacht aantal van 100 uitkomsten een afwijking van 3 minder sterk weegt dan bij een verwacht aantal van 10 uitkomsten. Men definieert dus de stochast χ^2 door

$$\chi^2 := \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = \frac{(X_1 - np_1)^2}{np_1} + \ldots + \frac{(X_k - np_k)^2}{np_k}.$$

Er laat zich aantonen dat χ^2 voor $n \to \infty$ een χ^2-verdeling met $k-1$ vrijheidsgraden heeft. Voor het geval $k = 2$ hebben we dit boven ingezien, want we hebben aangetoond dat

$$\frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \left(\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} \right)^2$$

en het laatste heeft voor $n \to \infty$ inderdaad een χ^2-verdeling met 1 vrijheidsgraad. Het bewijs voor algemene k vergt behoorlijk meer moeite en wordt hier onderdrukt.

We geven wel een iets handigere manier aan om χ^2 uit te rekenen. Uit $(X_i - np_i)^2 = X_i^2 - 2X_i np_i + n^2 p_i^2$ volgt dat $\frac{(X_i - np_i)^2}{np_i} = \frac{X_i^2}{np_i} - 2X_i + np_i$. We hebben $\sum_{i=1}^k p_i = 1$ en omdat de som van de X_i het totaal aantal n van uitkomsten aangeeft, geldt $\sum_{i=1}^k X_i = n$. Hiermee krijgen we

$$\chi^2 := \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{X_i^2}{np_i} - \sum_{i=1}^{k} 2X_i + \sum_{i=1}^{k} np_i = \left(\sum_{i=1}^{k} \frac{X_i^2}{np_i} \right) - n.$$

De kansverdeling die de verdeling van n uitkomsten over k klassen beschrijft, waarbij een uitkomst met kans p_i in de i-de klasse valt, heet de *multinomiale verdeling* met parameters $p_1, \ldots, p_k$ (die aan $p_1 + \ldots + p_k = 1$ moeten voldoen). Er geldt

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1!\, n_2! \ldots n_k!}\, p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

waarbij $n_1 + \ldots + n_k = n$ is. De multinomiale verdeling voor het speciaal geval $k = 2$ is natuurlijk juist de binomiale verdeling.

Het idee van een toets, de zogeheten χ^2-*aanpassingstoets* of kort χ^2-*toets*, is nu hetzelfde als bij de toetsen die we in de vorige les hebben gezien. Voor de verschillende aantallen ν van vrijheidsgraden en de verschillende levels α van onbetrouwbaarheid worden waarden $\chi^2_{\nu,\alpha}$ bepaald zo dat

$$P(\chi^2 > \chi^2_{\nu,\alpha}) = \alpha.$$

Onder de aanname van de nulhypothese geeft een steekproef dus (slechts) met kans α een χ^2-waarde die zo groot of groter is dan χ^2 en de nulhypothese wordt verworpen als een waarde van χ^2 wordt gevonden die groter is dan $\chi^2_{\nu,\alpha}$ voor de gekozen level α. Vaak wordt ook hier de P-waarde van χ^2 bepaald, dus de kans waarmee de stochast X van de nulhypothese een steekproef oplevert die een χ^2-waarde oplevert die zo groot of groter is dan de gevonden χ^2.

Merk op: Een belangrijke voorwaarde voor de toepasbaarheid van de χ^2-toets is, dat voor iedere klasse de verwachte aantallen $np_i \geq 5$ zijn, want anders wordt de verdeling van χ^2 niet nauwkeurig genoeg door een χ^2-verdeling benaderd. Dit eist soms dat men klassen samenvoegt die anders te weinig waarnemingen laten verwachten.

In het voorbeeld van de normale verdeling heeft de klasse K_1 de verwachte relatieve frequentie $p_1 = 0.0013$: Om hier op $np_1 \geq 5$ te komen, moeten we een steekproef van grootte $n \geq 3847$ hebben. Als dit niet haalbaar is, kunnen we bijvoorbeeld de klassen K_1 en K_2 samenvoegen, de gecombineerde kans voor deze twee klassen is $p'_1 = 0.02275$ en om nu aan de voorwaarde $np'_1 \geq 5$ te voldoen is een steekproef van grootte $n \geq 220$ voldoende.

Voorbeeld: We nemen aan dat we voor onze oneerlijke dobbelsteen met kansen $(\frac{5}{24}, \frac{1}{6}, \frac{1}{12}, 0, \frac{13}{24}, 0)$ bij een steekproef met $n = 120$ worpen precies de juiste aantallen vinden, dus $(25, 20, 10, 0, 65, 0)$. Bij een eerlijke dobbelsteen is $p_1 = \ldots = p_6 = \frac{1}{6}$ en we zouden dus voor elke klasse 20 uitkomsten verwachten. De waarde voor χ^2 is in dit geval

$$\begin{aligned}\chi^2 &= \frac{(25-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(0-20)^2}{20} + \frac{(65-20)^2}{20} + \frac{(0-20)^2}{20} \\ &= \frac{1}{20}(25 + 0 + 100 + 400 + 2025 + 400) = 147.5\end{aligned}$$

en voor $\alpha = 0.01$ vind men in de tabellen voor een χ^2-verdeling met 5 vrijheidsgraden de waarde $\chi^2_{5,0.01} = 15.1$ en zelfs voor $\alpha = 0.001$ is $\chi^2_{5,0.001} = 20.5$ veel kleiner dan de gevonden waarde voor χ^2. De P-waarde voor $\chi^2 = 147.5$ is in feite $4.5 \cdot 10^{-30}$ dus is het nagenoeg uitgesloten dat een resultaat met zo'n grote waarde voor χ^2 toevallig door een eerlijke dobbelsteen opgeleverd zou worden.

Voorbeeld: Van een bepaalde plantensoort komen volgens de wetten van Mendel vier variaties voor in de verhouding $9 : 3 : 3 : 1$. De verwachte relatieve frequenties zijn dus $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$ en $p_4 = \frac{1}{16}$. In een steekproef van 160 exemplaren vindt men de volgende aantallen n_i, de met de verwachte aantallen np_i vergeleken worden:

	variatie				
	1	2	3	4	totaal
n_i	88	35	24	13	160
np_i	90	30	30	10	160

Omdat de verdeling 4 klassen bevat, hebben we de kritieke waarden van de χ^2-verdeling met 3 vrijheidsgraden nodig. Voor $\alpha = 0.1$ is $\chi^2_{3,0.1} = 6.25$ en voor $\alpha = 0.05$ is $\chi^2_{3,0.05} = 7.81$. Als waarde voor χ^2 krijgen we

$$\chi^2 = \frac{(88-90)^2}{90} + \frac{(35-30)^2}{30} + \frac{(24-30)^2}{30} + \frac{(13-10)^2}{10} \approx 2.98$$

dus geeft dit experiment niet eens op een onbetrouwbaarheidslevel van 10% evidentie tegen de wetten van Mendel. De P-waarde van $\chi^2 = 2.98$ is 0.395, dit betekent dat 39.5% van de steekproeven minstens een χ^2-waarde van 2.98 zou opleveren, dus is onze steekproef zeker geen atypisch resultaat.

Meestal wordt de χ^2-aanpassingstoets als rechtséénzijdige toets toegepast, die aangeeft wat de kans is dat een steekproef in het geval van de nulhypothese een zo grote χ^2-waarde geeft. Er zijn echter ook gevallen waarbij een tweezijdige χ^2-toets uitgevoerd wordt, omdat men steekproeven ook verdacht vindt, als ze *te goed* bij de nulhypothese passen.

Een voorbeeld hiervoor is het toetsen van *toevalsgetallen*. Voor toevalsgetallen tussen 0 en 1 kan men bijvoorbeeld als klassen bijvoorbeeld de deelintervallen van lengte 0.1 kiezen. Als een toevalsgenerator nu 10000 toevalsgetallen produceert, zou men ongeveer 1000 getallen in ieder deelinterval verwachten en men berekent hiervoor de waarde van χ^2. Natuurlijk mag χ^2 in dit geval niet te groot zijn, omdat dit evidentie tegen de nulhypothese geeft dat de toevalsgenerator onbevooroordeeld is. Maar omgekeerd geeft een te kleine χ^2-waarde aanleiding tot de aanname dat er te veel regelmaat in de toevalsgetallen zit en de rij toevalsgetallen voorspelbaar is. Dit is evidentie tegen de nulhypothese dat de toevalsgenerator de getallen onafhankelijk van elkaar produceert. Men zou in dit geval de toevalsgenerator als ongeschikt verwerpen als de χ^2-waarde niet tussen $\chi^2_{0.05}$ en $\chi^2_{0.95}$ ligt.

> Een van de grondleggers van de statistiek, R.A. Fisher, heeft de χ^2-toets op de experimenten van Gregor Mendel met erwten toegepast, waardoor deze tot de ontdekking van de genen werd geleid (zonder ze zo te noemen). Fisher kwam tot het resultaat dat χ^2 een P-waarde van 0.99996 had, dus slechts 4 in 100000 steekproeven zouden een zo kleine χ^2-waarde opleveren. Het lijkt erop dat Mendel's tuin assistent precies wist, welke uitslag Mendel bij zijn experimenten verwachte en hier een handje bij heeft geholpen.

De waarden $\chi^2_{\nu,\alpha}$

De $\chi^2_{\nu,\alpha}$-waarden zijn net zo als de z-waarden en t-waarden voor verschillende parameters ν en α in tabellen opgeslagen of worden door software pakketten berekend. Voor grotere aantallen van vrijheidsgraden zijn er zekere benaderingen die op het verband van de χ^2-verdeling met de normale verdeling berusten.

(1) Voor een stochast χ^2 met een χ^2-verdeling met ν vrijheidsgraden is

$$Z := \sqrt{2\chi^2} - \sqrt{2\nu - 1}$$

bij benadering standaard-normaal verdeeld, waarbij deze benadering zeker voor $\nu > 100$ toegepast mag worden. Door dit naar χ^2 op te lossen, volgt dat men $\chi^2_{\nu,\alpha}$ kan benaderen door

$$\chi^2_{\nu,\alpha} \approx \frac{1}{2}\left(z_\alpha + \sqrt{2\nu - 1}\right)^2.$$

(2) Een betere benadering krijgt men uit het feit dat ook

$$Z := \frac{\sqrt[3]{\frac{\chi^2}{\nu}} - (1 - \frac{2}{9\nu})}{\frac{2}{9\nu}}$$

bij benadering standaard-normaal verdeeld is. Oplossen hiervan naar χ^2 geeft de benadering

$$\chi^2_{\nu,\alpha} \approx \nu \cdot \left(1 - \frac{2}{9\nu} + z_\alpha\sqrt{\frac{2}{9\nu}}\right)^3.$$

Er wordt soms aangegeven de benadering (1) voor $\nu > 100$ en de betere benadering (2) voor $\nu > 30$ toe te passen, maar met deze voorwaarden zit men zeker aan de veilige kant.

Voor $\nu = 50$ en $\alpha = 0.05$ is bijvoorbeeld de juiste waarde $\chi_{50,0.05} = 67.5048$, benadering (1) geeft $\chi_{50,0.05} \approx 67.2189$ en benadering (2) $\chi_{50,0.05} \approx 67.5006$. Zelfs voor $\nu = 10$ en $\alpha = 0.05$ is de fout van de twee benaderingen nog klein, de juiste waarde is hier $\chi_{10,0.05} = 18.3070$, benadering (1) geeft $\chi_{10,0.05} \approx 18.0225$ en benadering (2) $\chi_{10,0.05} \approx 18.2918$.

Verschillende kritieke waarden $\chi^2_{\nu,\alpha}$ zijn in Tabel 3 te vinden. Voor aantallen van vrijheidsgraden die niet in de tabel genoteerd zijn, kan men (voor voldoende grote ν) de boven aangegeven benaderingen toepassen, of een waarde voor een hoger aantal vrijheidsgraden kiezen, die wel genoteerd is. Op deze manier wordt in ieder geval de kans op een type I fout niet vergroot.

Onbekende parameters

In veel gevallen wil men toetsen of een steekproef door een stochast met een zeker *type* van kansverdeling geproduceerd is, bijvoorbeeld met een binomiale verdeling of een normale verdeling. In dit geval hangt de verdeling voor de nulhypothese van onbekende parameters af die uit de steekproef geschat moeten worden. Bij een schatter voor het gemiddelde van een kansverdeling hebben we gezien dat door het vervangen van de variantie door een schatting de verdeling breder wordt, omdat er meer onzekerheid in de schatting zit. We moesten daarom de normale verdeling door de Student-t verdeling vervangen.

Iets soortgelijks gebeurt ook bij de χ^2-toetsen. Als we de parameters van de verdeling waarmee we de verwachte kansen p_i berekenen door schattingen

$\nu\backslash\alpha$	0.95	0.1	0.05	0.01	0.001
1	.0039	2.71	3.84	6.63	10.8
2	.103	4.61	5.99	9.21	13.8
3	.352	6.25	7.81	11.3	16.3
4	.711	7.78	9.49	13.3	18.5
5	1.15	9.24	11.1	15.1	20.5
6	1.64	10.6	12.6	16.8	22.5
7	2.17	12.0	14.1	18.5	24.3
8	2.73	13.4	15.5	20.1	26.1
9	3.33	14.7	16.9	21.7	27.9
10	3.94	16.0	18.3	23.2	29.6
12	5.23	18.5	21.0	26.2	32.9
15	7.26	22.3	25.0	30.6	37.7
20	10.9	28.4	31.4	37.6	45.3
25	14.6	34.4	37.7	44.3	52.6
30	18.5	40.3	43.8	50.9	59.7
40	26.5	51.8	55.8	63.7	73.4
50	34.8	63.2	67.5	76.2	86.7
70	51.7	85.5	90.5	100	112
100	77.9	118	124	136	149

Tabel 3: Kritieke waarden $\chi_{\nu,\alpha}$ voor de χ^2-verdelingen met ν vrijheidsgraden.

vervangen, passen we de waarden p_i al aan de steekproef aan, daarom wordt in dit geval de onzekerheid kleiner tegenover het geval van bekende parameters. Op een gegeven onbetrouwbaarheidslevel α moeten de kritieke waarden dus kleiner worden. Gelukkig laat zich bewijzen dat dit op een overzichtelijke manier gebeurt, er moet namelijk voor elke parameter die we uit de steekproef schatten één vrijheidsgraad afgetrokken worden. Er geldt:

Stelling: Als voor het berekenen van de verwachte kansen p_i voor een uitkomst in de i-de klasse r parameters voor de kansverdeling van X met een maximum likelihood schatting worden bepaald, dan heeft $\chi^2 := \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$ voor $n \to \infty$ een χ^2-verdeling met $k - 1 - r$ vrijheidsgraden.

Merk op: Voor het gemiddelde μ van een verdeling is de maximum likelihood schatting gewoon het steekproefgemiddelde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ en voor de parameter p van een binomiale verdeling is $\overline{p} = \frac{k}{n}$ de maximum likelihood schatting, waarbij k het aantal successen bij n pogingen is. Aan de andere kant geldt dat de maximum likelihood schatting voor de variantie niet de steekproefvariantie $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is, maar $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n-1}{n}\, s^2$. Maar omdat de verdeling van χ^2 toch alleen maar voor $n \to \infty$ een χ^2-verdeling geeft, maakt het niet zo veel uit of we de variantie σ^2 door de (zuivere) schatting s^2 of de asymptotisch zuivere maximum likelihood schatting $\frac{n-1}{n}\, s^2$ vervangen. Vaak wordt daarom in de literatuur ook alleen maar aangegeven, dat een parameter door *een* schatting wordt vervangen, maar niet of door de maximum likelihood schatting of door een andere schatting.

Voorbeeld: Om het uur worden uit een productieproces steekproeven genomen van 5 stuks en het aantal defecte stukken wordt genoteerd. In 200 zulke steekproeven zijn de volgende resultaten gevonden:

aantal defecte stukken	0	1	2	3	4	5
aantal steekproeven	104	58	26	8	4	0

We willen toetsen of het aantal defecte stukken een binomiale verdeling heeft omdat dit het geval zou zijn als de kans op defecte stukken over de tijd constant gebleven is. Omdat de parameter p van de binomiale verdeling niet bekend is, moeten we deze uit de steekproeven schatten. We krijgen hiervoor

$$\overline{p} = \frac{1}{1000}(104 \cdot 0 + 58 \cdot 1 + 26 \cdot 2 + 8 \cdot 3 + 4 \cdot 4 + 0 \cdot 5) = \frac{150}{1000} = 0.15.$$

Als indeling van de steekproeven in klassen kiezen we de aantallen defecte stukken in een steekproef (van 5 stuks). De verwachte relatieve frequentie p_i voor de i-de klasse (met i defecte stukken) is dan volgens de binomiale verdeling met parameters $m = 5$ en $p = \overline{p} = 0.15$ gegeven door

$$p_i = \binom{m}{i}\overline{p}^i \cdot (1-\overline{p})^{m-i} = \binom{5}{i}0.15^i \cdot 0.85^{5-i}$$

en voor de $n = 200$ steekproeven krijgen we als verwachte aantallen voor de klassen

defect	0	1	2	3	4	5
p_i	0.444	0.392	0.138	0.024	0.002	0.0001
np_i	88.74	78.30	27.64	4.88	0.43	0.02

Omdat de verwachte aantallen voor de klassen met 3, 4 en 5 defecte stukken te klein zijn, voegen we deze samen tot één klasse met ≥ 3 defecte stukken. We krijgen zo de volgende statistiek waarvoor we de χ^2-waarde moeten bepalen:

defect	0	1	2	≥ 3
n_i	104	58	26	12
np_i	88.74	78.30	27.64	5.32

Omdat we de parameter $\overline{p}$ van de binomiale verdeling uit de steekproeven hebben geschat, heeft de χ^2-verdeling $4 - 1 - 1 = 2$ vrijheidsgraden. Op de levels $\alpha = 0.05$ en $\alpha = 0.01$ hebben we de kritieke waarden $\chi^2_{2,0.05} = 5.99$ en $\chi^2_{2,0.01} = 9.21$. Er geldt nu

$$\chi^2 = \frac{(104-88.74)^2}{88.74} + \frac{(58-78.30)^2}{78.30} + \frac{(26-27.64)^2}{27.64} + \frac{(12-5.32)^2}{5.32} \approx 16.37$$

dus kunnen we de nulhypothese van een binomiale verdeling zelfs op de onbetrouwbaarheidslevel $\alpha = 0.01$ veilig verwerpen. De P-waarde van $\chi^2 = 16.37$ is in feite 0.0003, een veel te lage waarde voor de aanname dat de afwijking van de binomiale verdeling toevallig is. We zouden dus concluderen, dat de kans p op defecte stukken in het productieproces over de tijd niet constant was.

5.2 χ^2-toets voor contingentietabellen

We hebben met de χ^2-aanpassingstoets getoetst of een steekproef bij een zekere kansverdeling past. Vaak komt men echter een iets andere vraag tegen, namelijk of twee of meer steekproeven bij een gemeenschappelijke kansverdeling horen, waarbij het niet nodig is deze gemeenschappelijke verdeling nader te bepalen. Dit probleem wordt meestal met een variatie van de χ^2-toets uit de vorige sectie aangepakt, waarbij men de verwachte aantallen uit de steekproeven bepaald. Hierbij gebruikt men een *contingentietabel*.

Stel we hebben r steekproeven met omvangen $n_1, \ldots, n_r$. Ieder van de steekproeven wordt op k klassen verdeeld, dit geeft de aantallen n_{ij} van elementen in de i-de steekproef, die in de j-de klasse vallen. We krijgen zo een $r \times k$-matrix met als elementen de hoeveelheden van elementen in de doorsnede van een steekproef en een klasse en dit noemen we een *contingentietabel*.

Met $n := \sum_{i=1}^{r} n_i = n_1 + \ldots + n_r$ noteren we de gemeenschappelijke omvang van alle steekproeven. We definiëren nu

$$p_j := \frac{n_{1j} + \ldots + n_{rj}}{n}$$

als kans voor een uitkomst in de j-de klasse, dit is juist de relatieve frequentie van uitkomsten in de j-de klasse in alle steekproeven. Met de kansen p_j krijgen we als verwachte waarde op positie (i, j) in de contingentietabel de waarde $n_i \cdot p_j$, want dit is het aantal uitkomsten in de j-de klasse die we bij een steekproef van omvang n_i zouden verwachten. We vatten nu de cellen van de contingentietabel als nieuwe klassen op en berekenen voor deze klassen de χ^2-waarde, dus

$$\chi^2 := \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - n_i \cdot p_j)^2}{n_i \cdot p_j}.$$

Er laat zich ook in dit geval aantonen, dat χ^2 voor $n \to \infty$ een χ^2-verdeling heeft, en het aantal vrijheidsgraden is $\nu = (r-1)(k-1)$. Dit kunnen we als volgt inzien: Als de p_j bekend waren, hadden we voor iedere steekproef $k-1$ vrijheidsgraden, dus in het geheel $r(k-1)$ vrijheidsgraden. Maar omdat we de p_j uit de steekproeven schatten, moeten we hiervan $k-1$ aftrekken (niet k, want p_k laat zich door $p_k = 1 - p_1 - \ldots - p_{k-1}$ uit de andere schattingen berekenen). Dit geeft dus $\nu = r(k-1) - (k-1) = (r-1)(k-1)$ vrijheidsgraden.

Voorbeeld: Bij een enquête in drie steden A, B en C werd een contingentietabel met de volgende resultaten gevonden:

stad	voor	tegen	neutraal	geen antwoord	totaal
A	105	61	87	167	420
B	118	60	130	145	453
C	88	58	62	101	309
totaal	311	179	279	413	1182

We hebben dus

$$n_1 = 420, \quad n_2 = 453, \quad n_3 = 309, \quad n = 1182,$$

$p_1 = \frac{311}{1182} \approx 0.263$, $p_2 = \frac{179}{1182} \approx 0.151$, $p_3 = \frac{279}{1182} \approx 0.236$, $p_4 = \frac{413}{1182} \approx 0.349$ en dit geeft als tabel met de verwachte aantallen $n_i \cdot p_j$:

stad	voor	tegen	neutraal	geen antwoord
A	110.5	63.6	99.1	146.8
B	119.2	68.6	106.9	158.3
C	81.3	46.8	72.9	108.0

Als we nu de waarde van χ^2 berekenen, zijn de cellen van de tabellen de nieuwe klassen en we krijgen

$$\chi^2 = \frac{(105-110.5)^2}{110.5} + \frac{(61-63.6)^2}{63.6} + \ldots + \frac{(101-108.0)^2}{108.0} \approx 17.2.$$

Dit moeten we vergelijken met de kritieke waarden van de χ^2-verdeling met $(3-1)\cdot(4-1) = 6$ vrijheidsgraden. We hebben $\chi^2_{6,0.05} = 12.6$ en $\chi^2_{6,0.01} = 16.8$, dus zijn de resultaten van de drie steden op de level $\alpha = 0.01$ significant verschillend.

In het geval van $r = 2$ steekproeven hebben we natuurlijk al eerder toetsen op verschillen van de verdelingen gezien, bijvoorbeeld toetsen op hetzelfde gemiddelde. Het hangt vaak van de vraagstukken af, of een χ^2-toets hier beter geschikt zou zijn. In het algemeen is de χ^2-toets minder scherp dan een toets op verschillen van de gemiddelden, aan de andere kant kan deze ook nog verschillen detecteren als de gemiddelden wel overeenkomen. In het bijzonder is de χ^2-toets ook toepasbaar, als de veronderstelling van een normaal verdeelde schatter niet meer houdbaar is.

Voorbeeld: Bij een niet nader toegelicht experiment met mogelijke uitslagen $1, \ldots, 10$ worden met twee verschillende methoden I en II de volgende aantallen uitslagen bereikt:

methode	1	2	3	4	5	6	7	8	9	10	totaal
I	6	16	22	38	44	30	18	12	8	6	200
II	2	6	12	22	29	30	21	16	8	4	150
totaal	8	22	34	60	73	60	39	28	16	10	350

Als geschatte kansen p_j voor de uitkomsten krijgen we

j	1	2	3	4	5	6	7	8	9	10
p_j	0.023	0.063	0.097	0.171	0.209	0.171	0.111	0.080	0.046	0.029

en als we hiermee de χ^2-waarde berekenen, krijgen we $\chi^2 \approx 11.12$. Voor een χ^2-verdeling met $(2-1)\cdot(10-1) = 9$ vrijheidsgraden hebben we $\chi_{9,0.1} = 14.7$, dus geeft de χ^2-toets met onbetrouwbaarheid $\alpha = 0.1$ geen evidentie voor een verschil van de twee methoden. De P-waarde van $\chi^2 = 11.12$ is 0.268.

Maar we kunnen met onze kennis uit de vorige les natuurlijk ook toetsen, of de twee methoden hetzelfde gemiddelde hebben. Hiervoor kijken we naar de

steekproefgemiddelden $\overline{x_I}$ en $\overline{x_{II}}$ en de steekproefvarianties s_I^2 en s_{II}^2 voor de twee steekproeven met omvangen $n_I = 200$ en $n_{II} = 150$. We hebben

$$\overline{x_I} = \frac{1}{200}(6 \cdot 1 + \ldots + 6 \cdot 10) = 5.05, \quad \overline{x_{II}} = \frac{1}{150}(2 \cdot 1 + \ldots + 4 \cdot 10) = 5.67$$

$$s_I^2 = 4.29, \quad s_{II}^2 = 3.86$$

en hieruit krijgen we voor de gepoolde variantie s^2 en standaardafwijking s:

$$s^2 = \frac{(n_I - 1)s_I^2 + (n_{II} - 1)s_{II}^2}{n_I + n_{II} - 2} = \frac{199 \cdot s_I^2 + 149 \cdot s_{II}^2}{348} = 4.11, \quad s = 2.03.$$

Als t-waarde die we met de kritieke waarden van de Student-t verdeling met 348 vrijheidsgraden moeten toetsen, hebben we

$$t = \frac{\overline{x_{II}} - \overline{x_I}}{s\sqrt{\frac{1}{n_I} + \frac{1}{n_{II}}}} \approx 2.82.$$

De verdeling van t is nagenoeg een standaard-normale verdeling en als P-waarde voor $t = 2.82$ vinden we 0.0024, dus vinden we met deze toets een significant verschil voor de gemiddelden van de twee methoden.

Toets op onafhankelijkheid van kenmerken

Een variatie op het vergelijken van r steekproeven geeft een toets op onafhankelijkheid van twee kenmerken in een steekproef. Bijvoorbeeld wil men weten, of het interesse in verschillende studievakken onafhankelijk is van het geslacht van de student. Men interpreteert nu de studenten van de verschillende studievakken als verschillende steekproeven en de indeling vrouw/man als indeling in klassen. De nulhypothese is, dat de kenmerken onafhankelijk zijn, in dit geval zou de kansverdeling voor iedere steekproef hetzelfde zijn en we zijn terug bij de situatie van de vorige sectie.

Voor het gemak nemen we aan dat het eerste kenmerk de waarden $\{1, \ldots, r\}$ kan hebben en het tweede kenmerk de waarden $\{1, \ldots, k\}$. Als n elementen in de steekproef zitten, noteren we met n_{ij} het aantal elementen met waarde i voor het eerste kenmerk en waarde j voor het tweede kenmerk. Als schatting p_{i*} voor de relatieve frequentie van elementen met waarde i voor het eerste kenmerk krijgen we

$$p_{i*} := \frac{n_{i1} + \ldots + n_{ik}}{n}$$

en als schatting p_{*j} voor de relatieve frequentie van elementen met waarde j voor het tweede kenmerk krijgen we

$$p_{*j} := \frac{n_{1j} + \ldots + n_{rj}}{n}.$$

De kansen p_{i*} en p_{*j} heten ook *marginale kansen*, omdat ze met de totale aantallen corresponderen die we aan de rand van de contingentietabel schrijven.

Onder de aanname van de nulhypothese zijn de twee kenmerken onafhankelijk, dus is de kans op een uitkomst in de cel (i, j) van de contingentietabel

$p_{i*} \cdot p_{*j}$ en het verwachte aantal uitkomsten voor deze cel is dus $n \cdot p_{i*} \cdot p_{*j}$. We kijken dus in dit geval naar de χ^2-waarde

$$\chi^2 := \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - np_{i*}p_{*j})^2}{np_{i*}p_{*j}}$$

en er laat zich weer aantonen dat dit voor $n \to \infty$ een χ^2-verdeling heeft. Omdat het schatten van de p_{i*} uit de steekproef $r-1$ vrijheidsgraden wegneemt en het schatten van de p_{*j} het aantal vrijheidsgraden om $k-1$ reduceert, hebben we $\nu = rk - 1 - (r-1) - (k-1) = (r-1)(k-1)$ vrijheidsgraden.

Voorbeeld: In een onderzoek werd getoetst of de prestaties van leerlingen in de vakken Engels en Wiskunde onafhankelijk van elkaar zijn. Men deelt de resultaten in 3 klassen, cijfers 6 en lager, cijfers 7 en 8 en cijfer 9 en 10.

	Wiskunde			
Engels	≤ 6	$7,8$	$9,10$	totaal
≤ 6	85	42	14	141
$7,8$	38	163	47	248
$9,10$	12	71	56	139
totaal	135	276	117	528

Hieruit krijgen we voor de marginale kansen:

$$p_{1*} = \frac{141}{528} = 0.267, \quad p_{2*} = \frac{248}{528} = 0.470, \quad p_{3*} = \frac{139}{528} = 0.263$$

$$p_{*1} = \frac{135}{528} = 0.256, \quad p_{*2} = \frac{276}{528} = 0.523, \quad p_{*3} = \frac{117}{528} = 0.222$$

Onder de aanname van de nulhypothese dat de twee kenmerken onafhankelijk zijn, zouden we voor de combinatie (i,j) van de kenmerken $n \cdot p_{i*} \cdot p_{*j}$ leerlingen in de steekproef verwachten. Dit geeft de verwachte waarden in de volgende tabel:

	Wiskunde		
Engels	≤ 6	$7,8$	$9,10$
≤ 6	36.1	73.7	31.2
$7,8$	63.4	129.6	55.0
$9,10$	35.5	72.7	30.8

We zien al dat dit behoorlijk afwijkt van de gevonden waarden. Als we hiervoor de χ^2-waarde berekenen, krijgen we

$$\chi^2 = \frac{(85-36.1)^2}{36.1} + \frac{(42-73.7)^2}{73.7} + \ldots + \frac{(56-30.8)^2}{30.8} \approx 145.8$$

terwijl we voor een χ^2-verdeling met $(3-1)\cdot(3-1) = 4$ vrijheidsgraden op significantie level $\alpha = 0.001$ de waarde $\chi^2_{4,0.001} = 18.5$ vinden. Het is dus duidelijk dat de resultaten in de twee vakken niet onafhankelijk van elkaar zijn.

Yates-correctie

In het speciaal geval van een 2×2 contingentietabel wordt vaak de *Yates-correctie* toegepast, die rekening ermee houdt, dat de data discreet is, maar de χ^2-verdeling een continue kansverdeling. In het algemeen wordt de χ^2-waarde met Yates-correctie bij l klassen met kansen $p_1, \ldots, p_l$ berekend door

$$\chi^2 := \sum_{i=1}^{l} \frac{(|X_i - np_i| - \frac{1}{2})^2}{np_i}$$

maar dit wordt eigenlijk alleen maar in het geval van 1 vrijheidsgraad toegepast, en dit is juist het geval voor $r = 2$ en $k = 2$.

De Yates-correctie heeft het effect dat de χ^2-waarde die berekent wordt iets lager is dan zonder de correctie. Dit leidt ertoe dat de nulhypothese met Yates-correctie minder snel verworpen wordt dan zonder Yates-correctie.

Voor grote steekproeven maakt de Yates-correctie bijna geen verschil en inmiddels wordt soms aanbevolen, de Yates-correctie *niet* toe te passen. Als alle gevonden aantallen van de cellen kleine zijn (bijvoorbeeld tussen 5 en 10 liggen) is het verstandig om de χ^2-waarde met en zonder Yates-correctie te bepalen. Als de twee manieren tot verschillende conclusies leiden (verwerpen van de nulhypothese bij de ene, niet verwerpen bij de andere), zou men de steekproef moeten vergroten om tot een duidelijke beslissing te kunnen komen.

Voorbeeld: In een proef wordt aan een groep van mensen met een bepaalde ziekte een nieuw medicijn gegeven, terwijl een tweede groep met dezelfde ziekte een placebo krijgt. Er wordt nu gekeken hoe veel van de mensen binnen een bepaalde periode gezond zijn geworden.

	gezond	ziek	totaal
medicijn	75	25	100
placebo	65	35	100
totaal	140	60	200

Als marginale kansen krijgen we hieruit

$$p_{1*} = p_{2*} = \frac{100}{200} = 0.5 \quad \text{en} \quad p_{*1} = \frac{140}{200} = 0.7, \quad p_{*2} = \frac{60}{200} = 0.3.$$

De aanname van onafhankelijkheid betekent in dit geval dat de nieuwe medicijn hetzelfde effect heeft als het placebo. Omdat de groepen even groot zijn, zouden we onder de aanname van onafhankelijkheid verwachten dat in beide groepen $200 \cdot 0.5 \cdot 0.7 = 70$ mensen gezond worden en dat $200 \cdot 0.5 \cdot 0.3 = 30$ ziek blijven.

Zonder Yates-correctie krijgen we hieruit de χ^2-waarde

$$\chi^2 = \frac{(75-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(65-70)^2}{70} + \frac{(35-30)^2}{30} \approx 2.38$$

en met Yates-correctie

$$\chi^2 = \frac{(|75-70|-0.5)^2}{70} + \frac{(|25-30|-0.5)^2}{30} + \frac{(|65-70|-0.5)^2}{70} + \frac{(|35-30|-0.5)^2}{30} \approx 1.93.$$

In beide gevallen kunnen we de nulhypothese op onafhankelijkheid op een level van $\alpha = 0.1$ niet verwerpen, want voor een χ^2-verdeling met 1 vrijheidsgraad vinden we $\chi^2_{1,0.1} = 2.71$. De P-waarde zonder Yates-correctie is 0.123 en de P-waarde met Yates-correctie is 0.165 en dit zijn allebij geen afzonderlijk kleine waarden. Om aan te tonen dat de nieuwe medicijn wel een effect heeft, zijn dus verdere experimenten nodig.

2×2-tabellen

In het voorbeeld hier boven hebben we kunnen zien, dat bij een 2×2-contingentietabel de tellers in de som voor χ^2 alle hetzelfde zijn (in het voorbeeld 5^2). Dit is geen toeval, maar in feite altijd het geval voor 2×2-tabellen en heeft tot gevolg dat we voor dit belangrijke speciaal geval de χ^2-waarde op een veel makkelijkere manier kunnen uitrekenen.

> Het zal geen verrassing zijn, dat een 2×2-tabel een speciaal geval is, want hier gaan we toetsen of twee relatieve frequenties hetzelfde zijn. In de vorige les hebben we gezien, dat we dit voor twee relatieve frequenties p_1 en p_2 kunnen doen, door de z-waarde
>
> $$z := \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}}$$
>
> met $p_0 := \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ te berekenen, die onder de aanname van de nulhypothese $p_1 = p_2$ standaard-normaal verdeeld is. De waarde χ^2 voor de χ^2-toets die we nu gaan berekenen is in dit speciaal geval juist het kwadraat van z.

We noteren de 2-contingentietabel als volgt:

	A	B	totaal
1	a	b	n_1
2	c	d	n_2
totaal	n_A	n_B	n

Hiervoor berekenen we de χ^2-waarde door

$$\begin{aligned}
\chi^2 &= \frac{(a - \frac{n_1 n_A}{n})^2}{\frac{n_1 n_A}{n}} + \frac{(b - \frac{n_1 n_B}{n})^2}{\frac{n_1 n_B}{n}} + \frac{(c - \frac{n_2 n_A}{n})^2}{\frac{n_2 n_A}{n}} + \frac{(d - \frac{n_2 n_B}{n})^2}{\frac{n_2 n_B}{n}} \\
&= \frac{n}{n_1 n_2 n_A n_B}\Bigg(n_2 n_B (a - \frac{n_1 n_A}{n})^2 + n_2 n_A (b - \frac{n_1 n_B}{n})^2 \\
&+ n_1 n_B (c - \frac{n_2 n_A}{n})^2 + n_1 n_A (d - \frac{n_2 n_B}{n})^2 \Bigg) \\
&= \frac{n}{n_1 n_2 n_A n_B}\Bigg(\frac{n_2 n_B}{n^2}(na - n_1 n_A)^2 + \frac{n_2 n_A}{n^2}(nb - n_1 n_B)^2 \\
&+ \frac{n_1 n_B}{n^2}(nc - n_2 n_A)^2 + \frac{n_1 n_A}{n^2}(nd - n_2 n_B)^2 \Bigg).
\end{aligned}$$

Dit ziet nog niet naar een verbetering uit, maar nu vullen we in dat $n = a+b+c+d$, $n_1 = a+b$, $n_2 = c+d$, $n_A = a+c$ en $n_B = b+d$. Dit geeft

$$\begin{aligned} na - n_1 n_A &= (a+b+c+d)a - (a+b)(a+c) \\ &= a^2 + ab + ac + ad - a^2 - ab - ac - bc = ad - bc =: \Delta. \end{aligned}$$

Op een soortgelijke manier zien we in, dat ook

$$nb - n_1 n_B = \Delta, \quad nc - n_2 n_A = \Delta, \quad nd - n_2 n_B = \Delta.$$

Dit is in feite het bewijs, dat we in de tellers van de termen voor χ^2 altijd hetzelfde getal vinden, namelijk $(\frac{\Delta}{n})^2$.

Als we nu nog invullen dat $n_1 + n_2 = n$ en $n_A + n_B = n$, zien we dat $n_2 n_B + n_2 n_A + n_1 n_B + n_1 n_A = n_2(n_B + n_A) + n_1(n_B + n_A) = (n_2 + n_1)n = n^2$ en daarom geldt

$$\frac{n_2 n_B}{n^2}\Delta^2 + \frac{n_2 n_A}{n^2}\Delta^2 + \frac{n_1 n_B}{n^2}\Delta^2 + \frac{n_1 n_A}{n^2}\Delta^2 = \Delta^2 = (ad-bc)^2.$$

Alles bij elkaar genomen, hebben we dus aangetoond dat

$$\chi^2 = \frac{n}{n_1 n_2 n_A n_B}(ad-bc)^2$$

en dit is voor 2×2-contingentietabllen inderdaad veel handiger dan de algemene formule van boven.

5.3 Variantie-analyse

Met de χ^2-toetsen zijn we nagegaan of verschillende steekproeven bij dezelfde verdeling horen. Vaak komt men echter ook de vraag tegen of meerdere verdelingen hetzelfde gemiddelde hebben, bijvoorbeeld als het om verschillende behandelingen van een zekere soort groente gaat. Voor twee steekproeven hebben we hier al naar gekeken, dit konden we met een toets op het verschil van de twee gemiddelden oplossen. Hiervoor hadden we onder de veronderstelling dat de twee steekproeven uit verdelingen met dezelfde variantie komen, gekeken naar de verdeling van de schatter

$$T := \frac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

waarbij $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ de gepoolde variantie van de steekproeven was.

Net zo als we met de χ^2-toets een veralgemening van het vergelijken van 2 relatieve frequenties op relatieve frequenties voor k klassen hebben gevonden, gaan we nu de toets op gelijkheid van gemiddelden op meer dan 2 steekproeven uitbreiden.

Het idee hierbij is, de totale variantie van de steekproeven te analyseren en deze te verdelen in de variantie binnen de enkele steekproeven en de variantie tussen de steekproeven. Daarom heet deze methode dan ook *variantie-analyse* of kort *ANOVA* (voor **AN**alysis **O**f **VA**riance).

We veronderstellen, dat we k steekproeven hebben die afkomstig zijn van normale verdelingen met dezelfde (onbekende) variantie σ^2 en met (onbekende) verwachtingswaarden $\mu_1, \ldots, \mu_k$. De i-de steekproef heeft omvang n_i en wordt met $x_{i1}, \ldots, x_{in_i}$ genoteerd. De totale omvang van alle steekproeven is $n := n_1 + \ldots + n_k$. De nulhypothese is

$$H_0 : \mu_1 = \ldots = \mu_k.$$

We berekenen de steekproefgemiddelden $\overline{x_i}$ en het gemiddelde $\overline{x}$ *en gros* (d.w.z. het gemiddelde over alle steekproeven), dus

$$\overline{x_i} := \frac{1}{n_i} \sum_j x_{ij} \quad \text{en} \quad \overline{x} := \frac{1}{n} \sum_{i,j} x_{ij} = \sum_i \frac{n_i}{n} \overline{x_i}.$$

De totale kwadratische afwijking

$$v := \sum_{i,j} (x_{ij} - \overline{x})^2$$

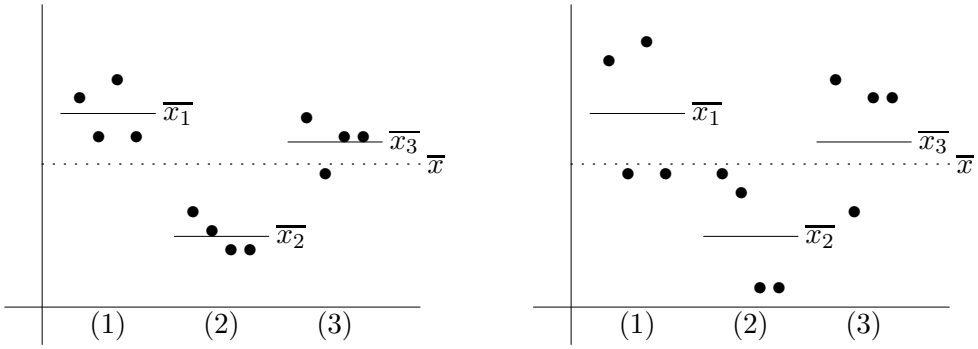
heeft nu twee bronnen, namelijk de kwadratische afwijkingen

$$v_i := \sum_j (x_{ij} - \overline{x_i})^2$$

binnen de steekproeven en de kwadratische afwijking

$$\sum_i (\overline{x_i} - \overline{x})^2.$$

tussen de steekproeven. Het idee achter deze opsplitsing van de kwadratische afwijkingen is in het volgende plaatje te zien:



In beide plaatjes zien we 3 steekproeven met telkens 4 waarden en de steekproefgemiddelden $\overline{x_i}$ zijn in beide gevallen hetzelfde.

In het linkerplaatje liggen de elementen van de steekproeven dicht bij de steekproefgemiddelden, daarom is de bijdrage van de kwadratische afwijkingen binnen de steekproeven in dit geval klein en de totale kwadratische afwijking wordt vooral veroorzaakt door de afwijkingen tussen de steekproefgemiddelden. Dit is sterke evidentie tegen de nulhypothese dat de gemiddelden van de verdelingen gelijk zijn.

In het rechterplaatje zijn de kwadratische afwijkingen binnen de steekproeven veel groter terwijl de kwadratische afwijkingen tussen de steekproefgemiddelden nog steeds hetzelfde zijn. Omdat in dit geval de kwadratische afwijkingen binnen de steekproeven relatief een groter deel bijdragen aan de totale kwadratische afwijking, zou men de nulhypothese moeilijker kunnen verwerpen, want de grote spreiding binnen de steekproeven maakt het plausibel, dat alle steekproeven door een verdeling met hetzelfde gemiddelde voortgebracht zijn.

Om het opsplitsen van de totale kwadratische afwijking binnen en tussen de steekproeven precies te analyseren, maken we weer gebruik van onze succesvolle aanpak, de elementen x_{ij} van de steekproeven als realisaties van onafhankelijke stochasten X_{ij} te zien. Ons uitgangspunt is hierbij, dat $X_{ij} \in \mathcal{N}(\mu_i, \sigma^2)$ is, dus normaal verdeeld met gemiddelde μ_i en variantie σ^2. De schatters $\overline{X_i}$ voor de gemiddelden van de steekproeven en $\overline{X}$ voor het gemiddelde over alle steekproeven zijn dan gegeven door

$$\overline{X_i} := \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \quad \text{en} \quad \overline{X} := \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^{k}\frac{n_i}{n}\overline{X_i}.$$

Er geldt nu

$$\begin{aligned}\sum_{i,j}(X_{ij}-\overline{X})^2 &= \sum_{i,j}((X_{ij}-\overline{X_i})+(\overline{X_i}-\overline{X}))^2\\ &= \sum_{i,j}(X_{ij}-\overline{X_i})^2 + \sum_{i,j}(\overline{X_i}-\overline{X})^2 + 2\sum_{i,j}(X_{ij}-\overline{X_i})(\overline{X_i}-\overline{X})\\ &= \sum_{i,j}(X_{ij}-\overline{X_i})^2 + \sum_{i} n_i(\overline{X_i}-\overline{X})^2 + 2\sum_{i,j}(X_{ij}-\overline{X_i})(\overline{X_i}-\overline{X}).\end{aligned}$$

Maar de laatste som geeft 0, omdat

$$\begin{aligned}\sum_{j}(X_{ij}-\overline{X_i})(\overline{X_i}-\overline{X}) &= (\overline{X_i}-\overline{X})(\sum_{j}(X_{ij}-\overline{X_i}))\\ &= (\overline{X_i}-\overline{X})((\sum_{j}X_{ij}) - n_i\overline{X_i}) = (\overline{X_i}-\overline{X})(n_i\overline{X_i}-n_i\overline{X_i}) = 0.\end{aligned}$$

Dus hebben we aangetoond dat

$$\sum_{i,j}(X_{ij}-\overline{X})^2 = \underbrace{\sum_{i,j}(X_{ij}-\overline{X_i})^2}_{V_b} + \underbrace{\sum_{i} n_i(\overline{X_i}-\overline{X})^2}_{V_t}.$$

We gaan nu de twee stochasten V_b (b voor **b**innen) en V_t (t voor **t**ussen) die gedefinieerd zijn door

$$V_b := \sum_{i,j}(X_{ij}-\overline{X_i})^2 \quad \text{en} \quad V_t := \sum_{i} n_i(\overline{X_i}-\overline{X})^2$$

apart onderzoeken.

We weten dat $S_i^2 = \frac{1}{n_i-1}\sum_j (X_{ij} - \overline{X_i})^2$ een zuivere schatter voor σ^2 is, daarom is $\sum_j (X_{ij} - \overline{X_i})^2$ een zuivere schatter voor $(n_i - 1)\sigma^2$. De som V_b over de kwadratische afwijkingen *binnen* de steekproeven is dus een zuivere schatter voor $\sum_i (n_i - 1)\sigma^2 = (n-k)\sigma^2$ en dus geldt:

$$S_b^2 := \frac{V_b}{n-k} \text{ is een zuivere schatter voor } \sigma^2.$$

Om de variantie tussen de steekproeven te analyseren, schrijven we de stochasten X_{ij} voor de uitkomsten in de steekproeven als $X_{ij} = \mu_i + E_{ij}$, waarbij E_{ij} de afwijking van de verwachtingswaarde μ_i van X_{ij} aangeeft. In het bijzonder is E_{ij} normaal verdeeld met verwachtingswaarde 0 en variantie σ^2.

Omdat de schatters $\overline{X_i}$ verwachtingswaarde μ_i hebben, heeft $\overline{X}$ de verwachtingswaarde

$$\mu := \frac{1}{n}\sum_i n_i \mu_i.$$

We schrijven nu $\mu_i = \mu + \alpha_i$, dan zijn de α_i juist de afwijkingen tussen de gemiddelden van de enkele verdelingen en het gemiddelde over alle verdelingen. In het bijzonder volgt uit $\mu = \frac{1}{n}\sum_i n_i\mu_i$ dat

$$\sum_i n_i\alpha_i = \sum_i n_i(\mu_i - \mu) = (\sum_i n_i\mu_i) - n\mu = 0.$$

Voor de stochast V_t geldt nu:

$$\begin{aligned}
V_t &= \sum_i n_i(\overline{X_i} - \overline{X})^2 = \sum_i n_i((\overline{X_i} - \mu_i) + (\mu - \overline{X}) + (\mu_i - \mu))^2 \\
&= \sum_i n_i(\overline{X_i} - \mu_i)^2 + \sum_i n_i(\mu - \overline{X})^2 + \sum_i n_i(\mu_i - \mu)^2 \\
&+2\sum_i n_i(\overline{X_i} - \mu_i)(\mu - \overline{X}) +2\sum_i n_i(\overline{X_i} - \mu_i)(\mu_i - \mu) +2\sum_i n_i(\mu - \overline{X})(\mu_i - \mu) \\
&= \sum_i n_i(\overline{X_i} - \mu_i)^2 + n(\mu - \overline{X})^2 + \sum_i n_i\alpha_i^2 \\
&+ 2(\mu - \overline{X})\underbrace{\sum_i n_i(\overline{X_i} - \mu_i)}_{n(\overline{X}-\mu)} +2\sum_i n_i(\overline{X_i} - \mu_i)\alpha_i + 2(\mu - \overline{X})\sum_i n_i\alpha_i \\
&= \sum_i n_i(\overline{X_i} - \mu_i)^2 - n(\mu - \overline{X})^2 + \sum_i n_i\alpha_i^2 + 2\sum_i n_i(\overline{X_i} - \mu_i)\alpha_i
\end{aligned}$$

We kijken nu naar de verwachtingswaarde van V_t: Omdat $E[\overline{X_i}] = \mu_i$ geldt, is $E[(\overline{X_i} - \mu_i)^2] = \frac{\sigma^2}{n_i}$ en omdat $E[\overline{X}] = \mu$ is $E[(\overline{X} - \mu)^2] = \frac{\sigma^2}{n}$. Verder hebben we natuurlijk $E[\overline{X_i} - \mu_i] = 0$, daarom geldt

$$\begin{aligned}
E[V_t] &= \sum_i n_i E[(\overline{X_i} - \mu_i)^2] - nE[(\mu - \overline{X})^2] + \sum_i n_i\alpha_i^2 + 2\sum_i n_i\alpha_i E[(\overline{X_i} - \mu_i)] \\
&= \sum_i n_i\frac{\sigma^2}{n_i} - n\frac{\sigma^2}{n} + \sum_i n_i\alpha_i^2 = (k-1)\sigma^2 + \sum_i n_i\alpha_i^2.
\end{aligned}$$

De nulhypothese luidt dat alle μ_i hetzelfde zijn, dus dat alle $\alpha_i = 0$ zijn, de alternatieve hypothese is, dat minstens een $\alpha_i \neq 0$ is. Hieruit volgt:

(1) Onder de aanname van de nulhypothese $\alpha_i = 0$ voor alle i is

$$S_t^2 := \frac{V_t}{k-1} \text{ is een zuivere schatter voor } \sigma^2.$$

(2) Onder de aanname van de alternatieve hypothese $\alpha_i \neq 0$ voor een i is

$$S_t^2 := \frac{V_t}{k-1} \text{ is een zuivere schatter voor } \sigma^2 + \frac{1}{k-1}\sum_i n_i\alpha_i^2 > \sigma^2.$$

Voor gegeven steekproeven berekenen we nu de concrete realisaties s_b^2 en s_t^2 van de schatters S_b^2 en S_t^2 voor σ^2, dus

$$s_b^2 := \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})^2 \quad \text{en} \quad s_t^2 := \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x_i} - \overline{x})^2.$$

Omdat onder de aanname van de nulhypothese S_b^2 en S_t^2 beide zuivere schatters voor σ^2 zijn, kunnen we in dit geval verwachten dat $s_b^2 \approx s_t^2$. Andersom geeft een waarde $s_t^2 \gg s_b^2$ evidentie tegen de nulhypothese. Men kijkt daarom naar de verdeling van de stochast

$$F := \frac{S_t^2}{S_b^2}$$

waarvoor men in het geval van de nulhypothese een waarde rond 1 verwacht. Analoog met de andere toetsen bepaalt men nu weer f-waarden f_α, zo dat onder de aanname van de nulhypothese steekproeven met een waarde van f_α of hoger voor F met kans α optreden, dus

$$P(F > f_\alpha) = \alpha.$$

Omdat men bij de nulhypothese een waarde van F rond 1 verwacht, zullen de $f_\alpha > 1$ zijn. Bij de F-toets met onbetrouwbaarheid α verwerpt men nu de nulhypothese als $\frac{s_t^2}{s_b^2} > f_\alpha$ is.

> De naam *variantie-analyse* voor de F-toets zou inmiddels duidelijk zijn. Men analyseert hoe veel van de totale kwadratische afwijking door de afwijkingen binnen de steekproeven veroorzaakt wordt en hoeveel door de afwijkingen tussen de steekproeven. Als het laatste relatief gezien te veel wordt, geeft dit evidentie tegen de nulhypothese dat de verdelingen van de steekproeven alle hetzelfde gemiddelde hebben.

De verdeling van F heet de *Fisher-verdeling* of *F-verdeling* en wordt afgeleid uit de χ^2-verdelingen.

De F-verdeling van Fisher

We weten dat $\frac{k-1}{\sigma}\,S_t^2$ een χ^2-verdeling χ^2_{k-1} met $k-1$ vrijheidsgraden heeft en $\frac{n-k}{\sigma}\,S_b^2$ een χ^2-verdeling χ^2_{n-k} met $n-k$ vrijheidsgraden. Hieruit volgt dat de F-verdeling gegeven is door

$$F = \frac{S_t^2}{S_b^2} = \frac{\frac{\chi^2_{k-1}}{k-1}}{\frac{\chi^2_{n-k}}{n-k}}$$

dus is F (tot op constanten na) een quotiënt van χ^2-verdeelde stochasten met $k-1$ en $n-k$ vrijheidsgraden. Deze twee aantallen van vrijheidsgraden karakteriseren de F-verdeling en we noteren de F-verdeling met $k-1$ en $n-k$ vrijheidsgraden met $F_{k-1,n-k}$.

Voor de geïnteresseerde lezer vermelden we hier de expliciete dichtheidsfunctie $f_{m,n}$ voor de F-verdeling $F_{m,n}$. Het zal geen verrassing zijn, dat deze op een quotiënt van de dichtheidsfuncties van χ^2-verdelingen lijkt:

$$f_{m,n}(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\,\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1} (n+mx)^{-\frac{m+n}{2}}$$

De verwachtingswaarde en variantie van $F_{m,n}$ zijn

$$E[F_{m,n}] = \frac{n}{n-2} \quad \text{en} \quad Var(F_{m,n}) = \frac{2n^2(n+m-2)}{m(n-2)^2(n-4)}.$$

In het speciaal geval met $k=2$ steekproeven laat zich aantonen dat de verdeling $F_{1,n}$ juist de verdeling van het kwadraat T^2 van een stochast T met Student-t verdeling met n vrijheidsgraden is.

Verder geldt dat voor $n \to \infty$ de verdeling $F_{m,n}$ tegen de verdeling van $\frac{\chi^2_m}{m}$ convergeert en voor $m \to \infty$ gaat $F_{m,n}$ tegen $\frac{n}{\chi^2_n}$.

Variantie-analyse tabellen

De resultaten van een variantie-analyse worden meestal in een bepaalde soort tabellen aangegeven, die er typisch als volgt uit zien:

bron	vrijheids-graden	kwadratische afwijkingen	schattingen voor σ^2	F-waarde	P-waarde
tussen	$k-1$	$\sum_i n_i(\overline{x_i}-\overline{x})^2$	s_t^2	$f = \frac{s_t^2}{s_b^2}$	$P(F_{k-1,n-k} > f)$
binnen	$n-k$	$\sum_{i,j}(x_{ij}-\overline{x_i})^2$	s_b^2		
totaal	$n-1$	$\sum_{i,j}(x_{ij}-\overline{x})^2$			

Voorbeeld: Bij vier leveranciers van een zekere stof worden steekproeven genomen en de zuiverheid van de stof bepaald (die in procent aangegeven wordt). De vraag is, of er evidentie tegen de nulhypothese is, dat de vier leveranciers even zuiver produceren. De steekproeven en hun gemiddelden zijn in de volgende tabel aangegeven:

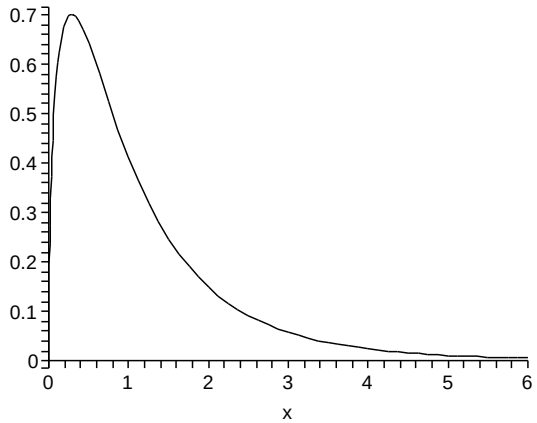
leverancier	steekproeven					n_i	$\overline{x_i}$
1	99.3	99.4	98.8	99.4		4	99.225
2	99.8	97.4	98.9	99.0	98.6	5	98.740
3	98.2	97.2	96.4	98.3		4	97.525
4	98.7	99.6	99.2			3	99.167
totaal						16	98.6375

We hebben $k = 4$ leveranciers en $n = 16$ steekproeven, daarom hebben we de F-verdeling met 3 en 12 vrijheidsgraden nodig. Uit deze gegevens berekent men de volgende variantie-analyse tabel:

bron	vrijheids-graden	kwadratische afwijkingen	schattingen voor σ^2	F-waarde	P-waarde
tussen	3	7.224	2.408	4.726	0.021
binnen	12	6.114	0.509		
totaal	15	13.337			

Afhankelijk van de gebruikte software wordt de P-waarde niet berekend, in dit geval vindt men in de tabellen voor $\alpha = 0.05$ de kritieke waarde $f_{3,12,0.05} = 3.49$ en voor $\alpha = 0.01$ de kritieke waarde $f_{3,12,0.01} = 5.95$. Men zou dus op een onbetrouwbaarheidslevel van 5% de nulhypothese wel kunnen verwerpen, maar op een onbetrouwbaarheidslevel van 1% niet meer. De P-waarde van 0.021 zegt juist, dat onder de aanname van de nulhypothese slechts 2.1% van de steekproeven een F-waarde van 4.726 of groter zouden opleveren.

We zien ook in Figuur 17 dat de gevonden waarde 4.726 van F al redelijk ver in de staart van de F-verdeling ligt, dus zou men in dit geval in ieder geval twijfels hebben of de leveranciers even zuivere stof produceren.



Figuur 17: F-verdeling met 3 en 12 vrijheidsgraden.

Belangrijke begrippen in deze les

- χ^2-aanpassingstoets

- kritieke waarden $\chi^2_{\nu,\alpha}$
- χ^2-toets bij onbekende parameters
- contingentietabel
- χ^2-toets op onafhankelijkheid
- Yates-correcte
- variantie-analyse (ANOVA)
- afwijkingen binnen en tussen steekproeven
- F-verdeling van Fisher
- F-toets

Opgaven

28. Er wordt 120 keer met een dobbelsteen geworpen. De aantallen voor de verschillende uitkomsten zijn:

$$1:12, \quad 2:21, \quad 3:27, \quad 4:22, \quad 5:20, \quad 6:18.$$

Is dit een zuivere dobbelsteen?

29. Bij een reukproef werd aan 50 willekeurig gekozen vrouwen gevraagd of zij parfum A lekkerder vonden dan B of omgekeerd. Aan A gaven 37 vrouwen de voorkeur, de overige vonden B lekkerder. Toets op de significantie level $\alpha = 0.1$ de nulhypothese dat er geen voorkeur voor één van de twee merken bestaat. Voer de toets zonder en met Yates-correctie uit.

30. In een weverij zijn in het verleden gemiddeld 2 weeffouten per $100m^2$ geweven doek opgetreden. Een recente steekproef op 100 stukken doek van $100m^2$ heeft het volgende resultaat opgeleverd:

fouten	0	1	2	3	4	5	6	7	8	9	10
aantal doeken	16	22	28	15	8	3	3	1	2	1	1

(i) Toets op een significantie level van $\alpha = 0.05$ de nulhypothese dat het aantal fouten Poisson-verdeeld met parameter $\lambda = 2$ is.

(ii) Toets op een significantie level van $\alpha = 0.05$ de nulhypothese dat het aantal fouten überhaupt Poisson-verdeeld is.

31. Van 1000 aselect gekozen personen is nagegaan of ze kleurenblind zijn. Van de 480 mannen bleken dit er 38 te zijn, bij de vrouwen was het aantal 6.

(i) Toets op de level $\alpha = 0.1$ of kleurenblindheid onafhankelijk is van het geslacht.

(ii) Wat is het minimale aantal vrouwen dat kleurenblind mag zijn, waarvoor de nulhypothese op level $\alpha = 0.1$ niet verworpen wordt (waarbij we nog steeds van 38 kleurenblinde mannen uit gaan)?

32. Twee groepen A en B van elk 100 patiënten hebben een bepaalde ziekte. Groep A wordt behandeld met een zeker serum, groep B met een ander serum. Na een bepaalde tijd zijn 75 patiënten van groep A en 65 patiënten van groep B genezen. Toets met onbetrouwbaarheid $\alpha = 0.05$ of beide sera evenveel effect hebben.

33. Bij een computerbedrijf wordt in 3 ploegen (ochtend, middag, nacht) op vier verschillende types van computers (A, B, C, D) gewerkt. De manager vraagt zich af of er bij het aantal reboots van computers een samenhang tussen de ploeg en de type computer bestaat. Hij heeft de volgende contingentietabel voor reboots gemaakt:

	type computer			
	A	B	C	D
ochtend	5	3	2	7
middag	7	12	9	16
nacht	1	2	4	2

Wat kan hij op een onbetrouwbaarheidslevel van $\alpha = 0.05$ zeggen?

34. Bij een crash-test met telkens 6 auto's van 3 verschillende merken wordt gekeken, wat de herstelling van de auto's kost. Er worden de volgende resultaten verkregen:

	kosten
A	200€ 50€ 150€ 75€ 100€ 250€
B	75€ 470€ 20€ 140€ 220€ 210€
C	120€ 570€ 600€ 450€ 700€ 350€

Kan op grond van deze waarden de nulhypothese dat de gemiddelde kosten bij iedere merk hetzelfde zijn op een onbetrouwbaarheidslevel van $\alpha = 0.05$ verworpen worden? Hoe zit het met $\alpha = 0.01$? De relevante kritieke waarden voor de F-verdeling zijn $f_{2,15,0.05} = 3.68$ en $f_{2,15,0.01} = 6.36$.

Les 6 Regressie en correlatie

Als we na twee kenmerken van elementen van een populatie kijken, is het een voor de hand liggende vraag of we aan de hand van de waarde van het eerste kenmerk een voorspelling kunnen doen voor de waarde van het tweede kenmerk. Bijvoorbeeld kunnen we ons afvragen of de prijs van een auto een indicatie geeft voor zijn levensduur, of de lengte van de wijsvinger iets zegt over de lengte van een persoon en of de resultaten van een toets in kansrekening iets te maken hebben met de resultaten in een toets over statistiek.

We zullen in deze les naar de samenhang tussen twee stochasten X en Y kijken, die gemeenschappelijke optreden. De belangrijkste vraag is hierbij of er een lineaire samenhang bestaat, dat wil zeggen of we Y kunnen schrijven in de vorm $Y = aX + b$.

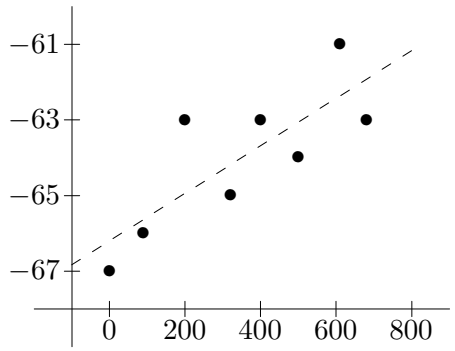
6.1 Regressie

Een vaak gebruikte aanpak om een idee van de samenhang van twee stochasten te krijgen, is een *scatterplot* (spreidingsdiagram), waarin de waarden voor de stochasten X en Y als coördinaten van punten in het 2-dimensionale $x-y$-vlak opgevat worden. Vaak kan men al aan de hand van een scatterplot een samenhang tussen de stochasten vast stellen, bijvoorbeeld als de punten enigszins op een lijn, op een parabool of op een exponentiële functie liggen.

Voorbeeld: Aan de zuidpool is het erg koud en men kan zich afvragen hou de temperatuur veranderd als men zich van de zuidpool verwijderd. Op een dag met $-67°C$ aan de zuidpool meet men in verschillende afstanden ook de temperatuur en krijgt de waarden in de volgende tabel:

i	1	2	3	4	5	6	7	8
afstand/km	0	90	200	320	400	500	610	680
temperatuur/$°C$	-67	-66	-63	-65	-63	-64	-61	-63

Als men de afstanden als x- en de temperaturen als y-coördinaten neemt, krijgt men het volgende plaatje. Daarbij is ook al een poging gedaan om een lijn door de waarden te leggen, die een samenhang tussen afstand en temperatuur zou kunnen beschrijven.



Figuur 18: Scatterplot met regressielijn.

Als we de grafiek van een functie van x door de punten leggen, nemen we een fundamentele beslissing: We maken x de *onafhankelijke* variabele en y de *afhankelijke* variabele. We spreken hierbij van *regressie* (van het Latijnse *regredere* = terugstappen), omdat we de afhankelijke variabele y op de onafhankelijke variabele x terug brengen. Merk op dat we alleen maar een samenhang tussen X en Y willen *beschrijven*, dit heeft niets met een eventuële causaliteit te maken. We maken helemaal geen uitspraak erover of X de *oorzaak* voor Y is of niet.

Als we nu beslissen dat we de grafiek van een functie $f(x)$ door de punten willen leggen, stelt zich de vraag hoe we zo'n grafiek kunnen bepalen. Omdat we y als afhankelijke variabele zien, is het een voor de hand liggende gedachte, de som van de kwadratische afstanden tussen de gevonden y-waarden en de grafiek te minimaliseren. Men spreekt dan van een *best fit* functie. Maar als we geen verdere aannamen over de functie $f(x)$ maken, is dit nog geen goed concept, want er laat zich aantonen dat er voor n punten altijd een veelterm van graad $n-1$ bestaat die precies door de n punten loopt. Hiervoor hebben we wel nodig dat de x-waarden van de punten verschillend zijn. We kunnen dus door n punten altijd een *perfecte fit* bereiken, maar die is vaak niet zo erg zinvol:

(i) De veelterm heeft in het algemeen gigantisch grote coëfficiënten en een grafiek met extreme stijgingen en geeft nauwelijks inzicht in de samenhang tussen de stochasten X en Y.

(ii) De punten die we in de scatterplot hebben getekend zijn vaak een steekproef van een veel grotere populatie (soms zelfs oneindig) en het voorspellingspotentiaal bij zo'n wilde functie is laag. We kunnen niet verwachten dat de y-waarde voor een x tussen of buiten de x_i goed door de functie beschreven wordt.

Voor een *fit* door een scatterplot neemt men daarom meestal geen willekeurige functie, maar een functie van bepaalde type die van een klein aantal parameters afhangt. Typische keuzes hiervoor zijn:

(1) Lineaire functies $f(x) = ax + b$.

(2) Kwadratische functies $f(x) = ax^2 + bx + c$.

(3) Exponentiële functies $f(x) = ae^{bx}$.

Het eenvoudigste maar ook meest belangrijke geval zijn de lineaire functies. Verschillende andere samenhangen laten zich door een transformatie van de variabelen op een lineaire samenhang terugbrengen. De volgende tabel geeft voor een aantal belangrijke relaties de hiervoor benodigde transformaties aan. Merk op dat de beste fit dan de kwadratische afstanden van de getransformeerde variabelen minimaliseert en niet de kwadratische afstanden bij de originele variabelen. Maar vaak is dit juist wenselijk, omdat bijvoorbeeld bij de exponentiële functie ae^{bx} de afwijkingen voor grotere waarden van x zeer groot worden en daarom de fit vooral door de grootste waarde van x bepaald zou worden.

veronderstelde samenhang	transformaties	lineaire samenhang
$y = \frac{a}{x} + b$	$x' = \frac{1}{x}, y' = y$	$y' = ax' + b$
$\frac{1}{y} = ax + b$	$x' = x, y' = \frac{1}{y}$	$y' = ax' + b$
$\frac{1}{y} = \frac{a}{x} + b$	$x' = \frac{1}{x}, y' = \frac{1}{y}$	$y' = ax' + b$
$y = be^{ax}$	$x' = x, y' = \log(y)$	$y' = ax' + \log(b)$
$y = bx^a$	$x' = \log(x), y' = \log(y)$	$y' = ax' + \log(b)$
$y = a\log(x) + b$	$x' = \log(x), y' = y$	$y' = ax' + b$

Om te herkennen dat er een samenhang van de vorm $y = be^{ax}$ bestaat, wordt vaak *logaritmisch* papier gebruikt, waarop de y-waarden al op een logaritmische schaal geplaatst worden. Op zo'n soort papier moeten de punten dan op een lijn liggen. Analoog wordt voor een functie van de vorm $y = bx^a$ *dubbel logaritmisch* papier gebruikt, waarbij zowel de x- en de y-waarden op logaritmische schalen ingetekend worden. Voor het menselijke oog is het veel makkelijker om te herkennen of punten enigszins dicht bij een lijn liggen, dan te zien of ze dicht bij de grafiek van een exponentiële of machtsfunctie liggen.

6.2 De regressielijn

We gaan nu voor een verzameling $(x_1, y_1), \ldots, (x_n, y_n)$ van punten de lijn $y = f(x) = ax + b$ bepalen zo dat de kwadratische afstanden van de y_i van deze lijn minimaal worden. Deze lijn heet de *regressielijn* door de punten (x_i, y_i). De voorwaarde aan de regressielijn is, dat de functie

$$V(a, b) := \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

minimaal wordt. Bij functies van één veranderlijke vindt men een minimum door de nulpunten van de afgeleide te zoeken. Dit lukt bij functies van meer veranderlijken net zo, waarbij men apart de *partiële afgeleiden* naar de verschillende variabelen bekijkt. Bij een partiële afgeleide worden de andere variabelen als constanten behandeld.

In ons geval hebben we

$$\frac{\partial}{\partial a} V(a, b) = \sum_i 2(ax_i + b - y_i)x_i \quad \text{en} \quad \frac{\partial}{\partial b} V(a, b) = \sum_i 2(ax_i + b - y_i).$$

Als we de twee partiële afgeleiden gelijk aan 0 zetten, volgt uit $\frac{\partial}{\partial a} V(a, b) = 0$

$$a \sum_i x_i^2 + b \sum_i x_i = \sum_i x_i y_i$$

en uit $\frac{\partial}{\partial b} V(a, b) = 0$ krijgen we

$$a \sum_i x_i + nb = \sum_i y_i.$$

Dit zijn twee *lineaire vergelijkingen* in de twee onbekenden a en b, die we in principe rechtstreeks zouden kunnen oplossen, maar die enigszins onoverzichtelijke coëfficiënten hebben. We gaan de coëfficiënten daarom met behulp van gemiddelden iets handiger noteren en definiëren hiervoor:

$$\overline{x} = \frac{1}{n}\sum_i x_i, \quad \overline{x^2} = \frac{1}{n}\sum_i x_i^2, \quad \overline{y} = \frac{1}{n}\sum_i y_i, \quad \overline{xy} = \frac{1}{n}\sum_i x_i y_i$$

dan worden de vergelijkingen

$$a\overline{x^2} + b\overline{x} = \overline{xy} \text{ en } a\overline{x} + b = \overline{y}.$$

Uit de laatste vergelijking volgt rechtstreeks dat

$$b = \overline{y} - a\overline{x}$$

en als we dit in de eerste vergelijking invullen, hebben we $a\overline{x^2} + \overline{x}\,\overline{y} - a\overline{x}^2 = \overline{xy}$ of te wel $a(\overline{x^2} - \overline{x}^2) = \overline{xy} - \overline{x}\,\overline{y}$, dus

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$$

en

$$b = \frac{\overline{y}\overline{x^2} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2}$$

Merk op: We hebben gezien dat $\overline{y} = a\overline{x} + b$ geldt, dus gaat de regressielijn in het bijzonder door het zwaartepunt $(\overline{x}, \overline{y})$ van de punten (x_i, y_i).

Voorbeeld: In het voorbeeld met de afstanden en temperaturen hebben we $n = 8$, $\overline{x} = 350$, $\overline{x^2} = 174375$, $\overline{y} = -64$ en $\overline{xy} = -22073.75$. Hieruit volgt $a \approx 0.006289$ en $b = \overline{y} - a\overline{x} \approx -66.201$. Dit zijn natuurlijk juist de coëfficiënten van de lijn die in Figuur 18 ingetekend is.

Ook nu lijken de uitdrukkingen voor de coëfficiënten a en b van de regressielijn nog te ingewikkeld om ze te kunnen onthouden, maar als we ons aan de kansrekening herinneren, kunnen we een heel mooi ezelsbruggetje bouwen.

Als x_i en y_i de waarden van twee uniform verdeelde discrete stochasten X en Y met n mogelijke waarden zijn, dan is $\overline{x}$ juist de verwachtingswaarde van X en $\overline{y}$ is de verwachtingswaarde van Y. Verder geldt dat $\overline{x^2} - \overline{x}^2 = E[X^2] - E[X]^2 = Var(X)$, dus de noemer van a is juist de variantie van X. Voor het product van de stochasten geldt $\overline{xy} - \overline{x}\,\overline{y} = E[X \cdot Y] - E[X] \cdot E[Y] = Cov(X, Y)$, dus is de teller van a de *covariantie* van X en Y. Met deze interpretatie hebben we dus

$$a = \frac{Cov(X, Y)}{Var(X)}$$

en dit ziet er een heel stuk beter uit.

Variantie-analyse en de correlatiecoëfficiënt

Oorspronkelijk hadden we de variantie van een uniform verdeelde stochast X met n waarden gedefinieerd door $Var(X) = \frac{1}{n}\sum_i (x_i - \overline{x})^2$ en door dit met $Var(X) = \overline{x^2} - \overline{x}^2$ te vergelijken, krijgen we de volgende identiteit, die ons vaak handig te pas zal komen:

$$\sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n\overline{x}^2 = n\overline{x^2} - n\overline{x}^2.$$

(Natuurlijk kunnen we deze identiteit ook rechtstreeks narekenen, maar dit hebben we in feite al eerder gedaan, en er is geen reden om vervelend werk twee keer te doen.)

Ook voor de covariantie hadden we een andere schrijfwijze gezien, namelijk $Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$. Als we de verwachtingswaarden weer als gemiddelden schrijven, krijgen we hiermee $Cov(X,Y) = \frac{1}{n}\sum_i (x - \overline{x})(y - \overline{y})$ en dit geeft de identiteit

$$\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i x_i y_i - \frac{1}{n}(\sum_i x_i)(\sum_i y_i) = n\overline{xy} - n\overline{x}\,\overline{y}.$$

Bij elkaar genomen kunnen we de stijgingscoëfficiënt a van de regressielijn dus ook schrijven als

$$a = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}.$$

We voeren nu nog eens nieuwe notaties in, om deze uitdrukking voor de stijgingscoëfficiënt van de regressielijn handiger te kunnen schrijven, namelijk

$$v_{xx} := \sum_i (x_i - \overline{x})^2, \quad v_{yy} := \sum_i (y_i - \overline{y})^2, \quad v_{xy} := \sum_i (x_i - \overline{x})(y_i - \overline{y})$$

dan geldt (heel kort)

$$a = \frac{v_{xy}}{v_{xx}}.$$

We hebben tot nu toe altijd gezegd dat y de afhankelijke en x de onafhankelijke variabele is. Maar we kunnen natuurlijk de rollen van x en y omdraaien en bij een regressielijn de kwadratische afstanden tot de x-waarden minimaliseren.

De lijn $x = a'y + b'$ die we als regressielijn voor x afhankelijk van y krijgen, zal in het algemeen afwijken van de lijn $y = ax + b$ die y met betrekking tot x uitdrukt. Er laat zich zelfs aantonen dat de twee lijnen alleen maar overeenkomen als de punten precies op een lijn liggen.

Om de nieuwe regressielijn $x = a'y + b'$ te bepalen, zouden we de hele procedure kunnen herhalen, maar dit is helemaal niet nodig. Als we de coördinaten (x_i, y_i) verruilen en dus de punten (y_i, x_i) bekijken, hoeven we in onze formules alleen maar alle letters x door y te vervangen en andersom. Omdat dit voor v_{xy} geen verschil maakt, krijgen we voor de stijgingscoëfficiënt a' de uitdrukking

$$a' := \frac{v_{xy}}{v_{yy}}.$$

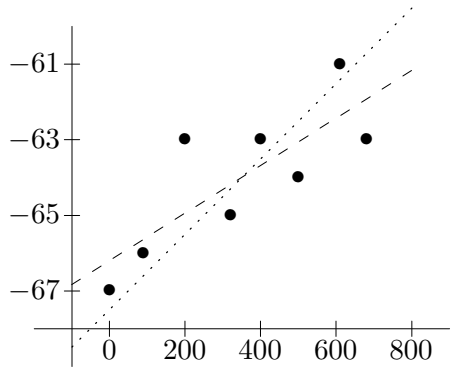
Omdat ook deze regressielijn door het zwaartepunt van de punten moet lopen, geldt verder dat

$$b' = \overline{x} - a'\overline{y}.$$

Om nu de twee regressielijnen met elkaar te kunnen vergelijken, moeten we de vergelijking $x = a'y + b'$ van de nieuwe regressielijn naar y oplossen. Dit geeft

$$y = \frac{1}{a'}x - \frac{b'}{a'} = \frac{v_{yy}}{v_{xy}}x - \frac{b'}{a'}.$$

Voorbeeld: In het voorbeeld hadden we voor x als onafhankelijke variabele de regressielijn $y = 0.006289x - 66.201$ gevonden die de kwadratische afstanden in y-richting minimaliseert. Als we in plaats hiervan naar de regressielijn van x in afhankelijkheid van y kijken, krijgen we de nieuwe regressielijn $y = 0.009962x - 67.487$, die de kwadratische afstanden in x-richting minimaliseert. In Figuur 19 zijn deze twee lijnen te zien, de eerste als stippellijn en de tweede als puntjeslijn. Het is duidelijk dat voor de tweede lijn de afstanden in x-richting kleiner zijn dan voor de eerste lijn.



Figuur 19: Regressielijnen voor x en y als onafhankelijke variabele.

Als we nu de stijgingscoëfficiënten $a = \frac{v_{xy}}{v_{xx}}$ en $\frac{1}{a'} = \frac{v_{yy}}{v_{xy}}$ van de twee regressielijnen vergelijken, krijgen we

$$\frac{a}{\frac{1}{a'}} = a \cdot a' = r^2 \text{ of } \frac{a}{r} = \frac{r}{a'} \quad \text{met} \quad r^2 = \frac{v_{xy}^2}{v_{xx}v_{yy}}.$$

Het getal r^2 is dus een maat voor de afwijking van de stijgingscoëfficiënten van elkaar en hoe dichter de punten bij een lijn liggen, hoe dichter zal r^2 bij 1 liggen.

We definiëren nu

$$r := \frac{v_{xy}}{\sqrt{v_{xx}v_{yy}}}$$

en noemen dit de *correlatiecoëfficiënt* van de punten $(x_1, y_1), \ldots, (x_n, y_n)$.

Een interpretatie van de correlatiecoëfficiënt krijgen we door een redenering die erg op de variantie-analyse lijkt. We kijken hiervoor naar de kwadratische afwijkingen van de y-waarden van het gemiddelde $\overline{y}$ en analyseren hoeveel van deze afwijking door de regressielijn verklaard wordt. Hoe groter het aandeel

van de verklaarde afwijking, hoe beter klopt de aanname van een lineaire samenhang.

Voor iedere waarde y_i noteren we met $\hat{y}_i = ax_i + b$ de volgens de regressielijn verwachte waarde bij x_i, dan heet het verschil $y_i - \hat{y}_i$ tussen de werkelijke waarde en de door de regressielijn voorspelde waarde het *residu* bij x_i. We noemen $(\hat{y}_i - \overline{y})^2$ de (door de regressielijn) *verklaarde afwijking* en $(y_i - \hat{y}_i)^2$ de *onverklaarde afwijking* van y_i van het gemiddelde $\overline{y}$.

We tonen nu aan dat de totale kwadratische afwijking juist de som van de verklaarde en van de onverklaarde afwijking is, d.w.z. we laten zien dat

$$v_{yy} = \sum_i (y_i - \overline{y})^2 = \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{onverklaard}} + \underbrace{\sum_i (\hat{y}_i - \overline{y})^2}_{\text{verklaard}}.$$

Er geldt

$$(y_i - \overline{y})^2 = \big((y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})\big)^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}),$$

dus moeten we laten zien dat

$$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_i (y_i - ax_i - b)(ax_i + b - \overline{y}) = 0.$$

Door uitwerken vinden we dat

$$(y_i - ax_i - b)(ax_i + b - \overline{y}) = ax_i(y_i - ax_i - b) + b(y_i - ax_i - b) - \overline{y}(y_i - ax_i - b)$$

en de lineaire vergelijkingen waaruit we de coëfficiënten a en b hebben gevonden waren juist

$$\sum_i x_i y_i = a \sum_i x_i^2 + b \sum_i x_i \quad \text{en} \quad \sum_i y_i = a \sum_i x_i + bn.$$

Daarom hebben we

$$\sum_i x_i(y_i - ax_i - b) = 0, \quad \text{en} \sum_i (y_i - ax_i - b) = 0$$

en hieruit volgt dat inderdaad

$$\sum_i (y_i - ax_i - b)(ax_i + b - \overline{y}) = 0.$$

We weten nu dat

$$\frac{\sum_i (y_i - \hat{y})^2}{v_{yy}} + \frac{\sum_i (\hat{y}_i - \overline{y})^2}{v_{yy}} = 1,$$

en $\frac{\sum_i (\hat{y}_i - \overline{y})^2}{v_{yy}}$ is juist het aandeel van de totale kwadratische afwijking dat door de regressielijn verklaard wordt. We gaan nu aantonen dat dit aandeel precies het kwadraat r^2 van de correlatiecoëfficiënt is, dus we laten zien dat

$$\frac{\sum_i (\hat{y}_i - \overline{y})^2}{v_{yy}} = \frac{v_{xy}^2}{v_{xx} v_{yy}} = r^2.$$

Om het rekenwerk iets eenvoudiger te houden, gaan we ervan uit dat we de paren (x_i, y_i) zo hebben verschoven dat $\overline{x} = 0$ en $\overline{y} = 0$ is. In dit geval geldt

$$a = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{v_{xy}}{v_{xx}} \quad \text{en} \quad b = 0.$$

Hieruit volgt

$$\frac{\sum_i (\hat{y}_i - \overline{y})^2}{v_{yy}} = \frac{\sum_i (a x_i)^2}{v_{yy}} = \frac{a^2 \sum_i x_i^2}{v_{yy}} = a^2 \frac{v_{xx}}{v_{yy}} = \frac{v_{xy}^2}{v_{xx}^2} \frac{v_{xx}}{v_{yy}} = \frac{v_{xy}^2}{v_{xx} v_{yy}} = r^2.$$

In het bijzonder volgt hiermee ook, dat $0 \leq r^2 \leq 1$ en dus $-1 \leq r \leq 1$. Omdat r^2 het aandeel van de kwadratische afwijking aangeeft dat door de regressielijn verklaard wordt, noemt men r^2 ook de *coëfficiënt van beslistheid* (coëfficiënt of determination).

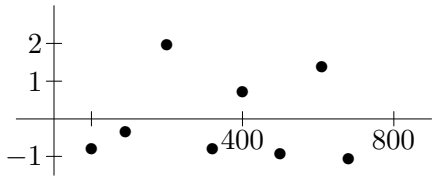
Voorbeeld: In het voorbeeld hadden we voor de twee regressielijnen de stijgingen $a = 0.006289$ en $\frac{1}{a'} = 0.009962$ gevonden, hieruit volgt dat $r^2 = 0.631$ en $r = 0.795$. Dit kunnen we ook uit de kwadratische afwijkingen afleiden:

Er geldt dat $\sum_i (y_i - \overline{y})^2 = 26$ en voor de residuën krijgen we de waarden

i	1	2	3	4	5	6	7	8
afstand	0	90	200	320	400	500	610	680
residu $y_i - \hat{y}_i$	−0.80	−0.36	1.94	−0.81	0.69	−0.94	1.36	−1.08

Voor de door de regressielijn verklaarde kwadratische afwijking krijgen we in dit voorbeeld $\sum_i (\hat{y}_i - \overline{y})^2 = 16.415$ en de onverklaarde afwijking is $\sum_i (y_i - \hat{y}_i)^2 = 9.585$ en we hebben $r^2 = \frac{16.415}{26} = 0.631$.

Om een idee te krijgen of een regressielijn wel een redelijke benadering geeft, is het vaak handig om alleen maar naar de residuën $y_i - \hat{y}_i$ te kijken. Als de residuën een soort van patroon laten zien, is er waarschijnlijk iets mis met de lineaire samenhang tussen de x- en de y-waarden, terwijl een willekeurige verdeling rond de x-as een goed teken is. In Figuur 20 zijn de residuën voor het voorbeeld met de temperaturen in verschillende afstanden van de zuidpool te zien. Er valt in dit geval geen duidelijk patroon op.



Figuur 20: Residuën $y_i - \hat{y}_i$ voor een regressielijn.

Steekproeven

Voor twee stochasten X en Y die twee kenmerken van een populatie beschrijven, heeft de stochast (X, Y) voor de paren van waarden een 2-dimensionale kansverdeling. Net als bij de gewone kansverdelingen kan deze 2-dimensionale kansverdeling discreet of continu zijn.

In het discrete geval is de kansverdeling door de kansen $P(X = x, Y = y)$ bepaald, in het continue geval door een dichtheidsfunctie $f(x, y)$ met verdelingsfunctie

$$F(x, y) := \int_{u \le x, v \le y} f(u, v)\, dv\, du.$$

De integratie over het gebied $-\infty < u \le x, -\infty < v \le y$ wordt hierbij door de twee in elkaar geschakelde gewone integraties berekend, dus

$$F(x, y) := \int_{-\infty}^{x} \big(\int_{-\infty}^{y} f(u, v)\, dv \big) du$$

waarbij (net als bij de partiële afgeleiden) de variabele van de buitenste integratie in de binnenste integraal als constante beschouwd wordt.

De paren $(x_1, y_1), \ldots, (x_n, y_n)$ vatten we nu op als steekproef voor de 2-dimensionale kansverdeling van de stochast (X, Y). Hierbij zijn dan natuurlijk de x_i steekproefwaarden voor X en de y_i steekproefwaarden voor Y, en we definiëren de steekproefvarianties (zo als altijd) door

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2, \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Analoog definiëren we ook een *steekproefcovariantie* door

$$s_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

dan kunnen we de stijgingscoëfficiënt a van de regressielijn door de punten (x_i, y_i) en de correlatiecoëfficiënt r schrijven als

$$a = \frac{s_{xy}}{s_x^2} \quad \text{en} \quad r = \frac{s_{xy}}{s_x s_y}.$$

In het bijzonder hebben we

$$(y - \overline{y}) = a(x - \overline{x}) = \frac{s_{xy}}{s_x^2} (x - \overline{x}) = \frac{s_y}{s_x} r (x - \overline{x}).$$

Als we nu (op de inmiddels welbekende manier) de twee variabelen x en y op z-coördinaten transformeren, zien we dat

$$\frac{y - \overline{y}}{s_y} = \frac{s_{xy}}{s_x^2 s_y} (x - \overline{x}) = \frac{s_{xy}}{s_x s_y} \frac{x - \overline{x}}{s_x} = r \frac{x - \overline{x}}{s_x}$$

dus vinden we ook hier weer een nieuwe interpretatie van de correlatiecoëfficiënt.

6.3 Het lineaire regressie model

Zelfs als we veronderstellen, dat er een lineaire samenhang tussen de stochasten X en Y bestaat, kan men niet verwachten dat de punten (x_i, y_i) van de

steekproef precies op een lijn liggen. De aanname dat de afwijkingen van de lijn toevallige fouten zijn leidt tot het *lineaire regressie model*, waarbij men een lineaire samenhang $Y = \alpha X + \beta$ tussen de stochasten verondersteld, die door een fout term verruisd is. Dit betekent dat de y-waarden y_i in de steekproef voor een gegeven waarde x_i van de vorm

$$y_i = \alpha x_i + \beta + \varepsilon_i$$

zijn, waarbij ε_i een foutterm is. Men neemt verder aan dat voor alle waarden van x_i de fouttermen normaal verdeeld met verwachtingswaarde 0 zijn. In principe zou de variantie van de fouttermen van de x-waarde x_i kunnen afhangen, maar om het model hanteerbaar te houden, gaat men ook hier ervan uit, dat de varianties van alle fouttermen gelijk zijn. De stochast E_i die de foutterm bij x_i aangeeft, heeft dus in het lineaire regressie model een normale verdeling met verwachtingswaarde 0 en variantie σ^2.

Er laat zich aantonen dat het lineaire regressie model juist is als de kansverdeling van (X, Y) een 2-dimensionale normale verdeling is. Voor twee *onafhankelijke* normaal verdeelde stochasten X en Y met $E[X] = \mu_X$, $E[Y] = \mu_Y$, $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$ is de gemeenschappelijke 2-dimensionale normale verdeling van (X, Y) gegeven door de dichtheidsfunctie

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{1}{2}((\frac{x-\mu_X}{\sigma_X})^2+(\frac{y-\mu_Y}{\sigma_Y})^2)}$$

die juist het product van de aparte dichtheidsfuncties voor X en Y is.

Als X en Y niet onafhankelijk zijn, is $\rho := \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$ de correlatiecoëfficiënt van X en Y. In dit geval heeft de 2-dimensionale normale verdeling van (X, Y) de dichtheidsfunctie

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}((\frac{x-\mu_X}{\sigma_X})^2-2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y}+(\frac{y-\mu_Y}{\sigma_Y})^2)}.$$

Ook als het paar (X, Y) geen 2-dimensionale normale verdeling heeft, biedt het lineaire regressie model in veel gevallen een redelijke aanpak, omdat (net als bij de gewone kansverdelingen) de verdeling van (X, Y) vaak goed door een normale verdeling benaderd wordt en dus ook het lineaire regressie model bij benadering klopt.

Schatters

De coëfficiënten a en b van de regressielijn door de punten (x_i, y_i) zien we nu als schatting voor de parameters α en β van het lineaire regressie model. Om dit verder te analyseren, noemen we de schatters, die op een concrete steekproef de waarden a en b geven A en B en de stochast die de verdeling van de y-waarden voor een vaste x_i beschrijft, noemen we Y_i. Volgens onze veronderstelling heeft dan Y_i een normale verdeling met verwachtingswaarde $\mu_i := \alpha x_i + \beta$ en variantie σ^2. Verder krijgen we ook een schatting e_i voor de foutterm ε_i, door $e_i := y_i - (ax_i + b)$ te definiëren.

We gaan nu de verwachtingswaarden en varianties van de schatters A en B bepalen. Hiervoor merken we eerst op dat uit de definitie van het gemiddelde $\overline{x}$ volgt, dat $\sum_i (x_i - \overline{x}) = 0$. Hieruit krijgen we

$$v_{xx} = \sum_i (x_i - \overline{x})(x_i - \overline{x}) = \sum_i (x_i - \overline{x})x_i - \overline{x} \underbrace{\sum_i (x_i - \overline{x})}_{=0} = \sum_i (x_i - \overline{x})x_i$$

en

$$v_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i (x_i - \overline{x})y_i - \overline{y} \underbrace{\sum_i (x_i - \overline{x})}_{=0} = \sum_i (x_i - \overline{x})y_i.$$

Uit het laatste volgt voor de stijgingscoëfficiënt a van de regressielijn, dat

$$a = \frac{v_{xy}}{v_{xx}} = \sum_i \frac{x_i - \overline{x}}{v_{xx}} y_i$$

en dus kunnen we de schatter A schrijven als

$$A = \sum_i \frac{x_i - \overline{x}}{v_{xx}} Y_i.$$

Voor de verwachtingswaarde $E[A]$ krijgen we hiermee

$$\begin{aligned} E[A] &= \sum_i \frac{x_i - \overline{x}}{v_{xx}} E[Y_i] = \sum_i \frac{x_i - \overline{x}}{v_{xx}} (\alpha x_i + \beta) \\ &= \frac{\alpha}{v_{xx}} \sum_i (x_i - \overline{x})x_i + \frac{\beta}{v_{xx}} \underbrace{\sum_i (x_i - \overline{x})}_{=0} = \frac{\alpha}{v_{xx}} v_{xx} = \alpha, \end{aligned}$$

dus is A een zuivere schatter voor α.

Omdat de Y_i onafhankelijk zijn en variantie σ^2 hebben, krijgen we voor de variantie $Var(A)$ dat

$$Var(A) = \sum_i (\frac{x_i - \overline{x}}{v_{xx}})^2 Var(Y_i) = \frac{v_{xx}}{v_{xx}^2} \sigma^2 = \frac{\sigma^2}{v_{xx}}.$$

Merk op dat de schatting a van α beter wordt naarmate de spreiding v_{xx} van de x-waarden in de steekproef groter wordt. Dit zou men ook verwachten, want de stijging van een regressielijn door punten met sterk verspreide x-waarden is minder gevoelig tegen schommelingen in de y-waarden van de punten dan een lijn door punten met x-waarden die dicht bij elkaar liggen.

Voor de schatter B gebruiken we nu dat a en b samenhangen door de relatie $\overline{y} = a\overline{x} + b$. Omdat de x-waarden x_i vast gekozen zijn, is $\overline{x}$ bij alle steekproeven hetzelfde en de verdeling van $\overline{y}$ wordt beschreven door de stochast $\frac{1}{n}\sum_i Y_i$. Voor de schatters A en B geldt dus de relatie

$$\frac{1}{n} \sum_i Y_i = A\overline{x} + B.$$

en hieruit volgt voor $B = \frac{1}{n}\sum_i Y_i - A\overline{x}$:

$$E[B] = \frac{1}{n}\sum_i E[Y_i] - E[A]\overline{x} = \frac{1}{n}\sum_i(\alpha x_i + \beta) - \alpha\overline{x} = \alpha\overline{x} + \frac{1}{n}\,n\beta - \alpha\overline{x} = \beta.$$

Ook B is dus een zuivere schatter voor de coëfficiënt β van het lineaire regressie model. Voor de variantie $Var(B)$ geldt:

$$Var(B) = \frac{1}{n^2}\sum_i Var(Y_i) + \overline{x}^2 Var(A) = \frac{1}{n^2}n\sigma^2 + \frac{\overline{x}^2}{v_{xx}}\sigma^2 = (\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}})\sigma^2.$$

Omdat $v_{xx} = n\overline{x^2} - n\overline{x}^2$, geldt $\overline{x}^2 = \overline{x^2} - \frac{1}{n}v_{xx}$ en dus kunnen we $Var(B)$ ook schrijven als

$$Var(B) = (\frac{\overline{x^2}}{v_{xx}})\sigma^2.$$

Ten slotte gaan we de schatter $C = \sum_i E_i^2$ analyseren, die als schatting de som $\sum_i e_i^2$ van de kwadraten van de residuën $e_i = y_i - \hat{y}_i$ geeft. We hadden al gezien dat $v_{yy} = \sum_i(y_i - \overline{y})^2 = \sum_i(y_i - \hat{y}_i)^2 + \sum_i(\hat{y}_i - \overline{y})^2$ is. Hieruit volgt dat

$$\sum_i e_i^2 = \sum_i(y_i - \hat{y}_i)^2 = v_{yy} - \sum_i(\hat{y}_i - \overline{y})^2 = v_{yy} - \frac{v_{xy}^2}{v_{xx}} = v_{yy} - a^2 v_{xx}.$$

Als we $\overline{Y} := \frac{1}{n}\sum_i Y_i$ schrijven, volgt hieruit voor de schatter $C = \sum_i E_i^2$ dat

$$C = \sum_i(Y_i - \overline{Y})^2 - v_{xx}A^2 = \sum_i Y_i^2 - n\overline{Y}^2 - v_{xx}A^2.$$

Met de relatie $Var(X) = E[X^2] - E[X]^2$, dus $E[X^2] = Var(X) + E[X]^2$, volgt hieruit

$$\begin{aligned}
E[C] &= \sum_i E[Y_i^2] - nE[\overline{Y}^2] - v_{xx}E[A^2] \\
&= \sum_i(Var(Y_i) + E[Y_i]^2) - n(Var(\overline{Y}) + E[\overline{Y}]^2) - v_{xx}(Var(A) + E[A]^2) \\
&= \sum_i(\sigma^2 + (\alpha x_i + \beta)^2) - n(\frac{1}{n}\sigma^2 + (\alpha\overline{x} + \beta)^2) - v_{xx}(\frac{1}{v_{xx}}\sigma^2 + \alpha^2) \\
&= n\sigma^2 + \sum_i(\alpha x_i + \beta)^2 - \sigma^2 - n(\alpha\overline{x} + \beta)^2 - \sigma^2 - v_{xx}\alpha^2 \\
&= (n-2)\sigma^2 + \sum_i(\alpha^2 x_i^2 + 2\alpha\beta x_i + \beta^2) - n\alpha^2\overline{x}^2 - 2\alpha\beta n\overline{x} - n\beta^2 - v_{xx}\alpha^2 \\
&= (n-2)\sigma^2 + \alpha^2(n\overline{x^2} - n\overline{x}^2 - v_{xx}) \\
&= (n-2)\sigma^2.
\end{aligned}$$

In de laatste stap hebben we hierbij gebruik ervan gemaakt dat $v_{xx} = \sum_i(x_i - \overline{x})^2 = n\overline{x^2} - n\overline{x}^2$. Uit $E[C] = (n-2)\sigma^2$ volgt in het bijzonder:

$$\frac{1}{n-2}\sum_i E_i^2 \text{ een zuivere schatter voor } \sigma^2.$$

De reden dat we in dit geval 2 vrijheidsgraden verliezen is, dat we de twee coëfficiënten a en b uit de steekproefwaarden hebben geschat.

Betrouwbaarheidsintervallen voor de parameters van het model

Volgens onze succesvolle strategie analyseren we nu weer onze schatters A en B om betrouwbaarheidsintervallen rond de schattingen a en b voor de coëfficiënten α en β van het lineaire regressie model te vinden.

Zo als altijd verschuiven we de schatter zo dat zijn verwachtingswaarde 0 wordt en delen vervolgens door zijn standaardafwijking, dus we kijken naar $Z := \frac{A-E[A]}{\sqrt{Var(A)}}$. Er laat zich aantonen dat de zo verkregen stochast

$$Z := \frac{A-\alpha}{\frac{\sigma}{\sqrt{v_{xx}}}} = \frac{A-\alpha}{\sigma}\sqrt{v_{xx}}$$

een standaard-normale verdeling heeft, dus kunnen we met de z-waarden betrouwbaarheidsintervallen definiëren. Er geldt dat

$$\begin{aligned} P(|Z| \leq z_{\frac{1-\gamma}{2}}) &= P(\alpha - z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}} \leq A \leq \alpha + z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}}) \\ &= P(A - z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}} \leq \alpha \leq A + z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}}) = \gamma. \end{aligned}$$

De stijgingscoëfficiënt a die we uit de steekproef hebben berekend, levert dus op betrouwbaarheidslevel γ (let op: α is nu de parameter van het lineaire regressie model) voor α het betrouwbaarheidsinterval

$$\left[a - z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}}, a + z_{\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{v_{xx}}}\right].$$

In de meeste gevallen zullen we de variantie σ^2 van de fouttermen niet kennen, in dit geval moeten we σ door de schatting

$$s := \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

vervangen. Hiermee krijgen we de stochast

$$T_\alpha := \frac{A-\alpha}{\frac{s}{\sqrt{v_{xx}}}} = \frac{A-\alpha}{s}\sqrt{v_{xx}}$$

en er laat zich aantonen dat T_α een Student-t verdeling met $n-2$ vrijheidsgraden heeft.

We hoeven dus in het boven gevonden betrouwbaarheidsinterval voor α alleen maar de kritieke z-waarde $z_{\frac{1-\gamma}{2}}$ door de kritieke t-waarde $t_{\frac{1-\gamma}{2}}$ van een Student-t verdeling met $n-2$ vrijheidsgraden te vervangen en de standaardafwijking σ door de schatting s en krijgen zo (bij onbekende variantie σ^2) op betrouwbaarheidslevel γ het volgende betrouwbaarheidsinterval voor α:

$$\left[a - t_{\frac{1-\gamma}{2}}\frac{s}{\sqrt{v_{xx}}}, a + t_{\frac{1-\gamma}{2}}\frac{s}{\sqrt{v_{xx}}}\right].$$

De stochast T_α laat ook weer het verband tussen de regressielijn en de variantie-analyse zien. Men kan aantonen dat T_α^2 een F-verdeling met 1 en $n-2$ vrijheidsgraden heeft.

Met behulp van de betrouwbaarheidsintervallen kunnen we nu ook toetsen voor de coëfficiënt α van het lineaire regressie model definiëren. De meest belangrijke vraag is hierbij meestal, of de steekproef (x_i, y_i) evidentie geeft tegen de nulhypothese

$$H_0 : \alpha = 0.$$

Bij een tweezijdige toets zullen we de nulhypothese $\alpha = 0$ op een significantie level van $1-\gamma$ verwerpen, als

$$|a| > z_{\frac{1-\gamma}{2}} \frac{\sigma}{\sqrt{v_{xx}}} \text{ bij bekende } \sigma^2 \quad \text{en} \quad |a| > t_{\frac{1-\gamma}{2}} \frac{s}{\sqrt{v_{xx}}} \text{ bij onbekende } \sigma^2.$$

Voorbeeld: In ons voorbeeld hadden we $a = 0.00629$ gevonden. We berekenen verder dat $v_{xx} = 415000$ en $\sum_i e_i^2 = 9.585$. Hieruit krijgen we de schatting $s^2 = 1.598$ voor σ^2 en dus voor σ de schatting $s = 1.264$. Om de nulhypothese $\alpha = 0$ te toetsen, moeten we $\frac{a}{s}\sqrt{v_{xx}}$ met de t-waarden van een Student-t verdeling met $8-2 = 6$ vrijheidsgraden vergelijken. Voor $\gamma = 95\%$ hebben we $t_{\frac{1-\gamma}{2}} = t_{6,0.025} = 2.45$ en voor onze waarden geldt dat $\frac{a}{s}\sqrt{v_{xx}} = 3.21$, dus kunnen we op een significantielevel van 5% de nulhypothese $\alpha = 0$ verwerpen. Het betrouwbaarheidsinterval dat we op level 5% voor α vinden is

$$\left[a - 2.45 \cdot \frac{s}{\sqrt{v_{xx}}}, a + 2.45 \cdot \frac{s}{\sqrt{v_{xx}}}\right] = [0.00148, 0.01110].$$

De reden dat we hier zo een relatief groot interval voor α krijgen ligt in het feit dat er bij een regressielijn door slechts 8 punten geen grote zekerheid over de stijging kan bestaan.

Minder belangrijk dan A is de schatter B voor de verschuiving β op de y-as. Ook hier krijgt men door de transformatie $Z := \frac{B-E[B]}{\sqrt{Var(B)}}$ en standaard-normaal verdeelde stochast, namelijk

$$Z := \frac{B-\beta}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}}.$$

Dit geeft op onbetrouwbaarheidslevel $1-\gamma$ voor β het betrouwbaarheidsinterval

$$\left[b - z_{\frac{1-\gamma}{2}}\,\sigma\,\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}, b + z_{\frac{1-\gamma}{2}}\,\sigma\,\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}\right].$$

Ook in dit geval geeft vervangen van σ door de schatting $s = \sqrt{\frac{\sum_i e_i^2}{n-2}}$ een stochast T_β met een Student-t verdeling met $n-2$ vrijheidsgraden, namelijk

$$T_\beta := \frac{B-\beta}{s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}}.$$

Hieruit volgt bij onbekende variantie σ^2 het betrouwbaarheidsinterval

$$\left[b - t_{\frac{1-\gamma}{2}}\, s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}, b + t_{\frac{1-\gamma}{2}}\, s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}}}\right]$$

voor β.

Betrouwbaarheidsintervallen voor de waarden

Uit de schatters A en B kunnen we voor een willekeurige waarde x door

$$M_x := Ax + B$$

een schatter voor het gemiddelde μ_x van de y-waarden voor de x-waarde x maken. Merk op dat we volgens het lineaire regressie model nog steeds ervan uit gaan dat de y-waarden voor x een normale verdeling rond $\mu_x = \alpha x + \beta$ hebben.

De verwachtingswaarde van M_x is $E[M_x] = \alpha x + \beta = \mu_x$ en dus is M_x een zuivere schatter voor μ_x.

Om de variantie van M_x te bepalen, merken we eerst op dat we hadden gezien dat $\overline{y} = a\overline{x} + b$, waaruit volgt dat voor een punt (x, y) op de regressielijn geldt dat $y - \overline{y} = a(x - \overline{x})$. Voor de schatter M_x volgt hieruit, dat

$$M_x - \overline{Y} = A(x - \overline{x}) \text{ en dus } M_x = \overline{Y} + A(x - \overline{x})$$

waarbij $\overline{Y}$ en A onafhankelijke stochasten zijn. Voor de variantie van M_x volgt hieruit

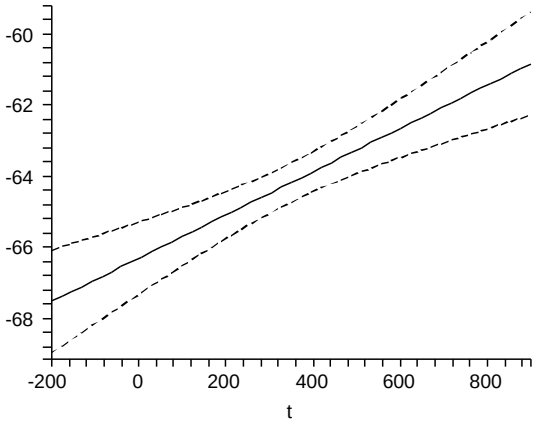
$$Var(M_x) = Var(\overline{Y}) + (x-\overline{x})^2 Var(A) = \frac{\sigma^2}{n} + (x-\overline{x})^2 \frac{\sigma^2}{v_{xx}} = (\frac{1}{n} + \frac{(x-\overline{x})^2}{v_{xx}})\sigma^2.$$

Omdat B juist de schatter M_x voor $x = 0$ is, moeten we hier voor $x = 0$ de variantie van B terugvinden, en dit is inderdaad het geval, want we hadden gevonden dat $Var(B) = (\frac{1}{n} + \frac{\overline{x}^2}{v_{xx}})\sigma^2$.

Merk op dat een betrouwbaarheidsinterval voor μ_x afhangt van de afstand tussen x en $\overline{x}$. Voor x dicht bij het gemiddelde is het betrouwbaarheidsinterval minder groot dan voor verder verwijderde. In Figuur 21 is dit voor ons voorbeeld duidelijk te zien, de twee krommen rond de regressielijn geven de grenzen van het betrouwbaarheidsinterval voor μ_x op een betrouwbaarheidslevel van 90% aan.

Als we nu een interval voor de y-waarden voor een zekere x-waarde willen schatten, weten we dat we volgens het lineaire regressie model de schatting $y = ax + b + e$ hebben, waarbij de foutterm e normaal verdeeld met variantie σ^2 is. Als we de schatter die de verdeling van de y-waarden voor de x-waarde x beschrijft met Y_x noteren, volgt hieruit dat

$$Y_x = Ax + B + E_x = M_x + E_x$$



Figuur 21: Betrouwbaarheidsintervallen voor de gemiddelden μ_x.

waarbij E_x normaal verdeeld met verwachtingswaarde 0 en variantie σ^2 is. Hieruit krijgen we

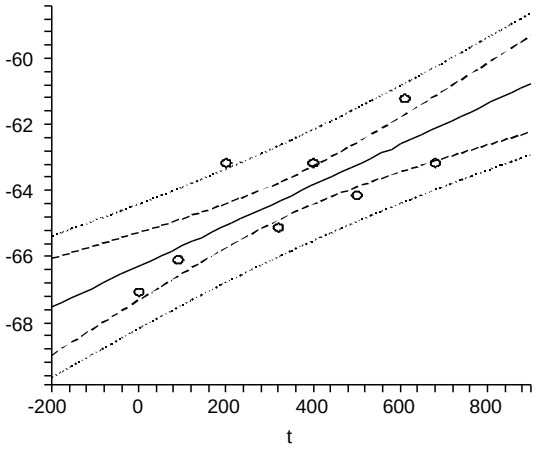
$$E[Y_x] = E[M_x] + E[E_x] = \mu_x,$$

dus is ook Y_x een zuivere schatter voor $\mu_x = \alpha x + \beta$.

Voor de variantie geldt

$$\begin{aligned} Var(Y_x) &= Var(M_x) + Var(E_x) = (\frac{1}{n} + \frac{(x-\overline{x})^2}{v_{xx}})\sigma^2 + \sigma^2 \\ &= (1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{v_{xx}})\sigma^2. \end{aligned}$$

Het zal geen verrassing zijn, dat de schatter Y_x voor de y-waarden een groter interval rond μ_x geeft dan de schatter M_x voor het gemiddelde van de y-waarden.



Figuur 22: Betrouwbaarheidsintervallen voor de y-waarden.

In Figuur 22 zijn voor ons voorbeeld de betrouwbaarheidsintervallen op betrouwbaarheidslevel 90% voor de y-waarden samen met de (kleinere) betrouw-

baarheidsintervallen voor de μ_x te zien. Omdat ook de punten voor de paren (x_i, y_i) aangegeven zijn, kunnen we in het bijzonder herkennen dat 7 van de 8 waarden binnen het betrouwbaarheidsinterval voor de y-waarden liggen, zo als we dat bij een onbetrouwbaarheid van 10% zouden kunnen verwachten.

Correlatie

Als we stochasten X en Y met zekere kansverdelingen en met een gemeenschappelijke kansverdeling voor (X, Y) veronderstellen, dan is de correlatiecoëfficiënt gedefinieerd door

$$\rho := \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

Als we de paren (x_i, y_i) als steekproef zien, waarbij de y-waarden voor de x-waarde x_i door de stochast Y_i worden beschreven, krijgen we de schatter

$$R := \frac{\sum_i (x_i - \overline{x})(Y_i - \overline{Y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (Y_i - \overline{Y})^2}}$$

die voor een concrete steekproef de correlatiecoëfficiënt

$$r := \frac{v_{xy}}{\sqrt{v_{xx} v_{yy}}}$$

als schatting voor ρ geeft.

We gaan er nu van uit dat het lineaire regressie model inderdaad van toepassing is, dus dat (X, Y) een 2-dimensionale normale verdeling heeft. Dan kunnen we uit r iets over ρ concluderen, als we op R de *Fisher transformatie* toepassen:

$$V := \frac{1}{2} \log(\frac{1+R}{1-R})$$

heeft (bij benadering) een normale verdeling met

$$E[V] = \frac{1}{2} \log(\frac{1+\rho}{1-\rho}) \quad \text{en} \quad Var(V) = \frac{1}{n-3}.$$

Hiermee kunnen we betrouwbaarheidsintervallen voor ρ definiëren en kunnen de nulhypothese $\rho = 0$ toetsen. Er geldt

$$P\left(\frac{1}{2} \log(\frac{1+\rho}{1-\rho}) - \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}} \leq \frac{1}{2} \log(\frac{1+r}{1-r}) \leq \frac{1}{2} \log(\frac{1+\rho}{1-\rho}) + \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}\right) = \gamma,$$

dus is

$$\left[\frac{1}{2} \log(\frac{1+r}{1-r}) - \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}, \frac{1}{2} \log(\frac{1+r}{1-r}) + \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}\right]$$

een betrouwbaarheidsinterval op significantie level $1-\gamma$ voor $\frac{1}{2} \log(\frac{1+\rho}{1-\rho})$.

Om een betrouwbaarheidsinterval voor ρ te krijgen, bepalen we de inverse functie van $f(x) = \frac{1}{2} \log(\frac{1+x}{1-x})$, d.w.z. we lossen $y = \frac{1}{2} \log(\frac{1+x}{1-x})$ naar x op. Er

geldt $e^{2y} = \frac{1+x}{1-x}$, dus is $e^{2y}-1 = \frac{1+x}{1-x} - \frac{1-x}{1-x} = \frac{2x}{1-x}$ en $e^{2y}+1 = \frac{1+x}{1-x} + \frac{1-x}{1-x} = \frac{2}{1-x}$. Hieruit volgt dat

$$x = \frac{e^{2y}-1}{e^{2y}+1} \Leftrightarrow y = \frac{1}{2}\log(\frac{1+x}{1-x}).$$

Met

$$v_1 := \frac{1}{2}\log(\frac{1+r}{1-r}) - \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}} \quad \text{en} \quad v_2 := \frac{1}{2}\log(\frac{1+r}{1-r}) + \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}$$

vindt men zo dat

$$P\left(\frac{e^{2v_1}-1}{e^{2v_1}+1} \leq \rho \leq \frac{e^{2v_2}-1}{e^{2v_2}+1}\right) = \gamma$$

en dit geeft

$$\left[\frac{e^{2v_1}-1}{e^{2v_1}+1}, \frac{e^{2v_2}-1}{e^{2v_2}+1}\right]$$

als betrouwbaarheidsinterval op significantielevel $1-\gamma$ voor ρ.

Omdat voor $\rho = 0$ geldt dat $\frac{1}{2}\log(\frac{1+\rho}{1-\rho}) = 0$ kunnen we de nulhypothese

$$H_0 : \rho = 0$$

op onbetrouwbaarheidslevel $1-\gamma$ heel makkelijk toetsen: De nulhypothese wordt verworpen als

$$|\frac{1}{2}\log(\frac{1+r}{1-r})| > \frac{z_{\frac{1-\gamma}{2}}}{\sqrt{n-3}}.$$

Ten slotte merken we op dat onder de veronderstelling dat (X, Y) een 2-dimensionale normale verdeling heeft, de toets op de nulhypothese $\rho = 0$ equivalent is met de toets dat de stijgingscoëfficiënt $\alpha = 0$ is. Er geldt namelijk dat de stochast T_α geschreven kan worden als

$$T_\alpha = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}.$$

Dit zien we als volgt in: We hebben gezien dat voor een gegeven steekproef (x_i, y_i) de schatter T_α de schatting

$$t = \frac{a-\alpha}{\sqrt{\frac{\sum_i e_i^2}{n-2}}}\sqrt{v_{xx}}$$

geeft. Onder de veronderstelling van de nulhypothese $\alpha = 0$ en door invullen van $a = \frac{v_{xy}}{v_{xx}}$ laat zich dit herschrijven tot

$$t = \frac{a}{\sqrt{\sum_i e_i^2}}\sqrt{v_{xx}(n-2)} = \frac{v_{xy}}{v_{xx}}\sqrt{\frac{v_{xx}(n-2)}{\sum_i e_i^2}} = \frac{v_{xy}}{\sqrt{v_{xx}}}\sqrt{\frac{(n-2)}{\sum_i e_i^2}}.$$

Aan de andere kant geldt dat r^2 het aandeel van de door de regressielijn verklaarde kwadratische afwijking v_{yy} is, daarom is $1-r^2$ juist het onverklaarde aandeel, dus

$$1 - r^2 = \frac{\sum_i e_i^2}{v_{yy}}.$$

Met $r = \frac{v_{xy}}{\sqrt{v_{xx}v_{yy}}}$ volgt hieruit, dat

$$\frac{r}{\sqrt{1-r^2}} = \frac{v_{xy}}{\sqrt{v_{xx}v_{yy}}}\sqrt{\frac{v_{yy}}{\sum_i e_i^2}} = \frac{v_{xy}}{\sqrt{v_{xx}}}\sqrt{\frac{1}{\sum_i e_i^2}},$$

dus geldt inderdaad dat

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

We kunnen dus de nulhypothese

$$H_0 : \alpha = 0$$

ook met behulp van de correlatiecoëfficiënt r van de steekproef (x_i, y_i) toetsen en we verwerpen de nulhypothese op een significantielevel $1-\gamma$ als

$$|\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}| > t_{\frac{1-\gamma}{2}}.$$

Belangrijke begrippen in deze les

- scatterplot
- regressie
- beste fit
- regressielijn
- correlatiecoëfficiënt
- lineair regressie model
- Fisher transformatie