

Hoofstuk 1: Inleiding

You all have probably heard the story about Malcolm Forbes, who once got lost floating in one of his famous balloons across miles and finally landed in the middle of a cornfield. He spotted a man coming toward him and asked: "Sir, can you tell me where I am ?" The man said: "Certainly, you are in a basket in a field of corn." Forbes said: "You must be a statistician." The man said: "That's amazing, how did you know that ?" "Easy", said Forbes, "your information is concise, precise and absolutely useless!". (uit: "Looking ahead: Cross-disciplinary opportunities for statistics" door R. Gnanadesiken)

Statistiek is een wetenschap, die met een eigen methodologie complexe verschijnselen helpt verklaren. Hierbij hoort het verzamelen, ordenen, beschrijven, voorstellen en interpreteren van gegevens om, op basis van die informatie, beslissingen te nemen en voorspellingen te maken.

De statistiek richt zich op waarnemingen uit alle denkbare gebieden. Als voorbeeld geven we enkele namen van grote afdelingen in de vereniging voor statistiek van de Verenigde Staten:

- biometrics
- statistical education
- business and economic statistics
- social statistics
- physical and engineering sciences
- statistical computing
- survey research methods
- biopharmaceutical statistics
- statistical graphics
- government statistics
- statistics and the environment
- biostatistics
- chemometrics

1.1. Steekproef en populatie

De groep individuen of objecten waarvan we het kenmerk willen onderzoeken, noemen we de **populatie** (bv. alle Belgen,...). Meestal is het ondoenbaar of onmogelijk om de gehele populatie te onderwerpen aan een onderzoek (te groot, te duur,...). Vaak nemen we daarom een klein gedeelte van de populatie, **een steekproef**. Een kok eet immers ook niet de hele pot soep om uitspraken te doen over de kwaliteit. Wel belangrijk is dat voor het proeven goed wordt geroerd. De eetlepel soep die beoordeeld wordt, moet overeenkomen met het geheel. Een steekproef moet dus **representatief** zijn: een correct beeld geven van de verscheidenheid binnen de populatie (bv: populatie = alle Belgen; steekproef moet zowel mannen als vrouwen van alle leeftijden en van alle landsgedeelten bevatten). De steekproef moet ook **aselect** zijn; elk element van de populatie moet evenveel kans hebben om opgenomen te worden in de steekproef. Alle TV-polls op basis van SMS en alle internet-polls zijn daarom compleet onbetrouwbaar. De steekproef moet tevens **onafhankelijk** zijn: de keuze van éénbepaald element mag de keuze van een ander element niet beïnvloeden.

Het aantal individuen in de steekproef heet de **steekproefgrootte**. Er bestaan verschillende technieken voor het nemen van een steekproef (sampling) zoals: random, systematische, gestratificeerde en cluster sampling. We gaan er hier niet verder op in.

1.2. Soorten statistiek

1.2.1. Verzamelende statistiek.

Het verzamelen van gegevens voor het onderzoek wordt gedaan via waarnemingen, metingen, tellingen of enquêtes bij de elementen van de steekproef. Het is belangrijk dat het vergaren van gegevens op een correcte manier gebeurt, zodanig dat er een antwoord kan gegeven worden op de vraagstelling. (Design van een experiment: de proefopzet)

1.2.2. Beschrijvende statistiek (inventariserende statistiek, descriptive statistics).

De gegevens worden gerangschikt en geordend in tabellen. Ze worden grafisch voorgesteld. Ook worden statistische grootheden zoals gemiddelden, standaardafwijkingen, percentages, vormcoëfficiënten en eventuele correlaties (statistische verbanden) berekend. De gegevens worden dus letterlijk beschreven en samengevat a.d.h.v. een beperkt aantal typerende parameters. We behandelen dit in het volgende hoofdstuk.

1.2.3. Verklarende statistiek (inductieve of inferentiële statistiek)

De verklarende statistiek steunt op de resultaten uit de beschrijvende statistiek en op de kanstheorie om uitspraken te doen over de ganse populatie. Zij heeft dus tot doel het generaliseren van de verzamelde informatie naar een groter geheel.

Vb.: Indien je het gemiddeld gewicht van alle studenten van de VUB wilt kennen, zul je adhv het gewicht van de studenten uit je steekproef, besluiten trekken over het gewicht van de totale populatie. Hoe groter de steekproef (mits goed genomen), hoe groter de betrouwbaarheid. We komen hierop terug in hoofdstuk 6.

Bij de beschrijvende statistiek gaan we dus van veel informatie naar een samenvattend overzicht, terwijl we bij de verklarende statistiek van “weinig” informatie naar algemene beweringen gaan.

Een statistische studie kan een waarnemingsstudie zijn, waarbij we waarnemen wat er (gebeurd) is en hieruit conclusies trekken. Het kan ook een experimentele studie zijn, waarbij we een variabele (de factor of onafhankelijke variabele) manipuleren en kijken hoe dit een andere variabele (afhankelijke variabele, uitkomst) beïnvloedt.

1.3. Gegevens en variabelen

De ruwe gegevens bestaan meestal uit waarnemingen (observations, readings, scores) uitgevoerd op individuen (cases). In veel gevallen zijn individuen ook echt personen, maar het kunnen even goed enzymes, hospitalen, ratten, bossen etc zijn.

Zoals reeds vermeld heet het aantal individuen in de steekproef de steekproefgrootte.

Met een variabele (veranderlijke) bedoelen we het kenmerk dat het onderwerp uitmaakt van een statistisch onderzoek. Deze eigenschap of kenmerk kan verschillende waarden (meestal binnen bepaalde grenzen) aannemen voor de verschillende individuen in de steekproef. Als het gaat over het gewicht van de studenten is “gewicht” een statistische variabele. We noteren statistische variabelen meestal met hoofdletters X,Y,Z.

1.4 Soorten data

De gegevens kunnen onderverdeeld worden volgens verschillende criteria. Het is belangrijk te weten tot welke klasse ze behoren, omdat de te gebruiken statistische technieken dikwijls bepaald worden door de aard van de data.

1.4.1 De schalen

De gegevens kunnen gemeten worden op verschillende types schalen:

- nominale schaal: beschrijving in woorden (vb: zwart, wit). We spreken dan van kwalitatieve of nominale variabelen of attributen. De individuen mogen slechts één van de mogelijke waarden van de variabele krijgen (niet overlappen).
- ordinale schaal: gerangschikte waarden (vb: zeer goed, goed, gemiddeld, slecht). Er is geen precieze afstand tussen de categorieën.
- intervalschaal: gegevens hebben alle eigenschappen van de ordinale schaal, maar ook de afstand tussen de verschillende niveaus is betekenisvol. De gegevens zijn numeriek. De verschillen weerspiegelen een “realiteitsverschil” op een multiplicatieve constante na, die afhangt van de gekozen eenheid. Het verband tussen verschillende eenheden is lineair maar het nulpunt is niet absoluut. Een voorbeeld zijn temperatuursgegevens met als mogelijke eenheden °C en °F ($^{\circ}\text{C} = 5/9 \ ^{\circ}\text{F} - 160/9$). We kunnen op deze intervalschaal naar verschillen kijken, die op een multiplicatieve factor gelijk zijn voor de beide schalen. We kunnen echter geen uitspraken doen over de eigenlijke gegevens, want 2 keer zo warm in °C heeft een totaal andere betekenis dan in °F.
- verhoudingsschaal (ratioschaal): De meetwaarden hebben alle karakteristieken van de intervalschaal, maar bovendien kan de verhouding der opmetingen rechtstreeks geïnterpreteerd worden als een verhouding tussen de werkelijke waarden. Om van de ene meeteenheid naar de andere over te stappen is enkel vermenigvuldigen met een constante toegelaten en bovendien is er een intrinsiek nulpunt op deze schaal. Als voorbeeld kun je het gewicht in kg of in g nemen.

N.B.: In verschillende teksten zal je merken dat men zich soms beperkt tot 3 schalen: nominaal, ordinaal en interval. De intervalschaal dekt voor deze auteurs zowel de interval- als de verhoudingsschaal.

1.4.2. Kwalitatieve en kwantitatieve gegevens

Een andere klassificatie van gegevens wordt uitgedrukt in termen van kwalitatief en kwantitatief. Kwantitatieve resultaten geven een kwantiteit weer. Zij vertellen ons over: hoeveel, hoe lang, hoe zwaar, hoe dikwijls,... Opmetingen op de interval- en de verhoudingsschaal zijn kwantitatief.

Kwalitatieve gegevens, ook wel eens categorische gegevens genoemd, bestaan uit namen of codes voor groepen van gelijkaardige individuen. Opmetingen op de nominale schaal zijn duidelijk kwalitatief. Strikt genomen behoren ook de ordinale opmetingen tot deze categorie, maar toch gebeurt het dat een (min of meer vaag) afstandsbegrip zinvol is bij dergelijke gegevens, zodat ook op deze observaties statistische methodes voor interval of verhoudingsopmetingen met succes worden toegepast. Hier moet dan wel telkens zeer zorgvuldig de interpretatie der resultaten worden bekeken.

Kwantitatieve variabelen kunnen op hun beurt nog onderverdeeld worden in continue en discrete variabelen. Hierbij wordt gekeken naar het aantal verschillende waarden die kunnen worden aangenomen. Men spreekt van discrete gegevens als men te maken heeft met een beperkt aantal mogelijke uitkomsten. Discrete variabelen zijn meestal het gevolg van een telling. Hoewel in feite elke opmeting op een discrete manier gebeurt (men moet toch ergens

afronden) spreekt men over continue gegevens als deze afkomstig zijn van een fenomeen dat in theorie een continuum van waarden kan aannemen.

1.4.3. Univariate en multivariate gegevens

Men maakt ook een onderscheid tussen **univariaat** en **multivariaat**: een dataset is univariaat indien de individuele data afhangen van 1 variabele, ze is multivariaat indien er meer variabelen een rol spelen.

In deze cursus zullen we vrijwel uitsluitend met univariate data van een verhoudingsschaal werken.

Hieronder zie je een typische dataset voor een groep 2^{de} kandidatuur studenten.

We noteren naam (X_0), geboortejaar (X_1), een aantal fysieke kenmerken zoals geslacht (X_2), kleur haar (X_3), kleur ogen (X_4), gewicht (X_5), lengte (X_6), een aantal studiekenmerken zoals studierichting (X_7), gemiddeld examencijfer in de eerste kandidatuur 1K (X_8), gemiddeld examencijfer bij het eindexamen HSO (X_9), en nog veel meer.

Deze gegevens kunnen we ordenen in een tabel van de vorm:

X_0	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Naam	Geboorte-Jaar	Geslacht V=0 M=1	Haarkleur bruin = 0 zwart = 1 blond = 2 rood = 3	Kleur ogen bruin=0 blauw = 1 grijs = 2 groen = 3	Gewicht kg	Lengte cm	Studie info=0 nat=1 sch=2 bio=3	Gem. 1K op 20	Gem. HSO op 100
Jan Jansen	1986	1	0	0	83.7	187	0	18	75
Irma Douce	1985	0	2	3	62.1	162	3	12	62
:	:	:	:	:	:	:	:	:	:

Oefening: geef de schalen van de verschillende gegevens.

1.5. De paradox van Russel

Om deze inleiding af te sluiten vermelden we nog een typische paradox.

Aan een universiteit waren vorig jaar bij de eerstejaarsstudenten 35.5% van de meisjesstudenten en 41% van de jongensstudenten geslaagd. Mogen we nu besluiten dat jongens slimmer zijn dan meisjes ?

	ingeschreven	geslaagd	% geslaagd
meisje	1000	355	35.5
jongen	1000	410	41

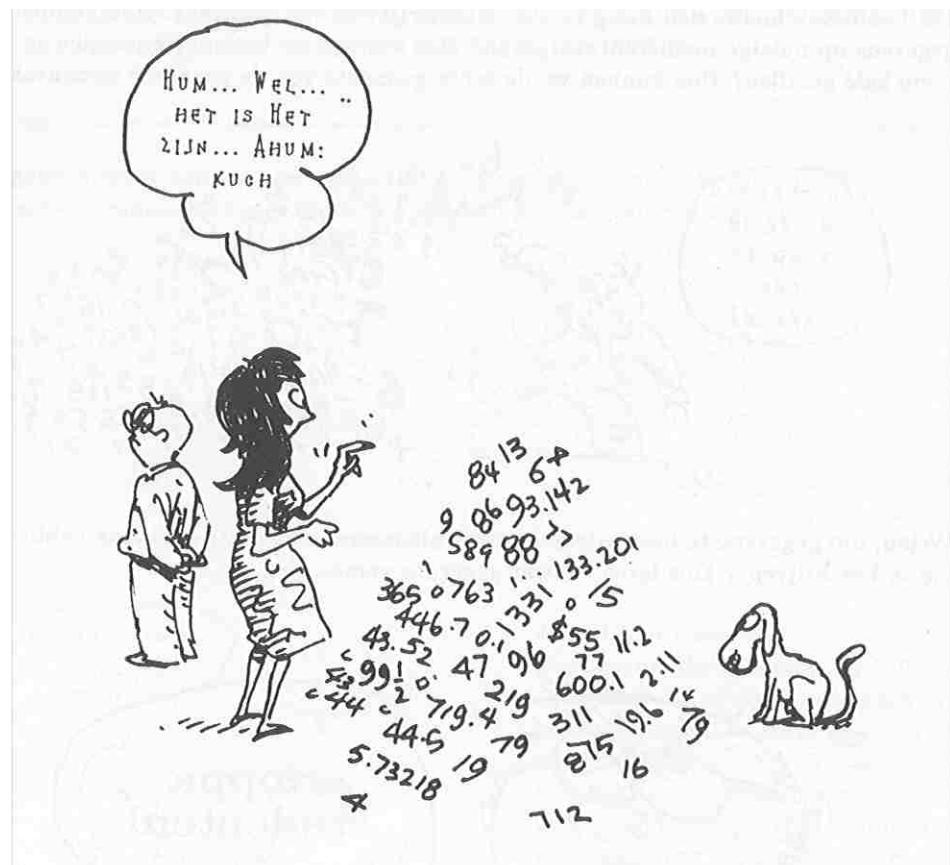
Laat ons eerst de resultaten aan deze universiteit eens iets meer in detail bekijken en op zoek gaan naar de studierichtingen, die het meest verantwoordelijk zijn voor het feit dat er zoveel meisjes falen.

Faculteit	Ingeschreven		Geslaagd			
	Meisje aantal	Jongen aantal	Meisje aantal	percent	Jongen aantal	percent
Economie	300	300	120	40%	120	40%
Geneeskunde	150	100	60	40%	40	40%
Pedagogie	300	50	90	30%	15	30%
Rechten	200	200	60	30%	60	30%
Wetenschappen	50	350	25	50%	175	50%
Totaal	1000	1000	355	35.5	410	41%

Het is duidelijk dat in elke faculteit het slaagpercentage van jongens en meisjes exact hetzelfde is. Hoe komt het nu dat we eerst een verkeerde conclusie zouden trekken ? In deze studie speelt namelijk een verdoken factor (studiekeuze) een belangrijke rol en het effect van deze factor is verstrengeld met het effect van geslacht op het studieresultaat. (Je kan zelfs voorbeelden construeren waarbij het effect omslaat in de andere richting).

Statistiek is dus veel meer dan alleen getallen. Wanneer statistiek vragen uit velerlei domeinen helpt oplossen, dan is het essentieel dat de hele context van de vraagstelling wordt bekeken. Informatie over de proefopzet is cruciaal bij de verdere verwerking van de waarnemingen en beïnvloedt heel sterk het type beoordeling of beslissing die statistisch nog te verantwoorden is. De keuze van een geschikte proefopzet (experimental design) om aan je vraagstelling te kunnen beantwoorden is dus van zeer groot belang. We zullen in deze cursus echter niet expliciet ingaan op dit belangrijk element bij een statistische studie.

Hoofstuk 2: Beschrijvende statistiek



In de beschrijvende statistiek worden de gegevens grafisch voorgesteld en samengevat. We zullen in dit hoofdstuk starten met het geven van methodes om gegevens grafisch voor te stellen. Nadien zullen we kengetallen definiëren om de gegevens samen te vatten.

2.1. Grafische voorstelling:

“One picture is worth a thousand words”

Dagelijks worden we overspoeld door een variëteit van grafische weergaven van kwantitatieve gegevens. Daarbij is het de bedoeling om in één oogopslag een duidelijk en overzichtelijk beeld te krijgen van de belangrijkste karakteristieken die in deze gegevens verweven zitten. Een correcte grafische voorstelling die een helder, eenduidig en waarheidsgetrouw beeld geeft, is lang niet eenvoudig. Vele grafische voorstellingen zijn ofwel te simplistisch, ofwel te gedetailleerd, ofwel ronduit verkeerd.

Bij een statistische studie gaan we de variabele waarin we geïnteresseerd zijn meten met behulp van een steekproef. We beschikken dan over de ruwe data (“raw data”). Na het meten en alvorens berekeningen uit te voeren is het onontbeerlijk je data grafisch voor te stellen. Dit is een stap die je nooit mag vergeten. Ook na een statistische analyse is een grafische voorstelling van de resultaten zeer waardevol.

In dit hoofdstuk bespreken we enkele eenvoudige methoden om data grafisch voor te stellen. We zullen achtereenvolgens de volgende grafieken behandelen:

Het histogram van de cumulatieve, absolute en relatieve frequentie

- Polygonen en ogieven
- De empirische verdelingsfunctie
- Het stengel-en-blad diagram
- Het pareto- en taartdiagram
- Het scatter diagram

2.1.1. Het histogram

Een klassieke manier om een groot aantal kwantitatieve gegevens grafisch samen te vatten is het histogram. We bespreken het opstellen van een histogram aan de hand van een voorbeeld. Stel we hebben de lengten van 100 volwassen Belgische mannen gemeten en de resultaten, afgerond tot op hele centimeters, staan afgebeeld in Tabel 2.1. We hebben dus een steekproef uit de verzameling van lengten van volwassen mannelijke inwoners van België met een steekproefgrootte $n = 100$.

Tabel 2.1

1.74	1.70	1.77	1.68	1.65	1.80	1.77	1.57	1.77	1.50
1.74	1.70	1.77	1.68	1.65	1.80	1.77	1.57	1.77	1.50
1.75	1.75	1.52	1.90	1.88	1.67	1.68	1.77	1.88	1.91
1.65	1.72	1.53	1.58	1.79	1.89	1.81	1.81	1.57	1.57
1.77	1.85	1.70	1.82	1.78	1.96	1.47	1.45	1.79	1.52
1.84	1.81	1.84	1.84	1.75	1.94	1.62	1.57	1.76	1.78
1.50	1.80	1.63	1.47	1.87	1.67	1.62	2.00	1.57	1.90
1.86	1.90	1.67	1.62	1.83	1.99	1.86	1.69	1.66	1.53
1.45	1.59	1.86	1.72	1.87	1.73	1.62	1.91	1.71	1.56
2.01	1.97	1.63	1.62	1.55	1.74	1.86	1.61	1.75	2.15

Erg veel informatie geeft zo'n tabel van ruwe gegevens niet: met name is het uit deze tabel moeilijk af te lezen wat de meest voorkomende lengte is en waar de uitersten liggen. Dezelfde gegevens, maar nu gesorteerd op grootte zoals in Tabel 2.2 geeft veel meer informatie. We zien onmiddellijk dat alle afmetingen in het interval [1.42 , 2.15] liggen en dat waarden in de buurt van 1.75 het meest voorkomen.

Tabel 2.2

1.42	1.53	1.58	1.63	1.68	1.74	1.77	1.81	1.86	1.90
1.45	1.53	1.59	1.64	1.68	1.74	1.77	1.81	1.86	1.91
1.45	1.55	1.61	1.65	1.69	1.75	1.78	1.81	1.86	1.91
1.47	1.56	1.62	1.65	1.70	1.75	1.78	1.82	1.87	1.94
1.47	1.57	1.62	1.66	1.70	1.75	1.79	1.83	1.87	1.96
1.49	1.57	1.62	1.67	1.71	1.75	1.79	1.84	1.88	1.97
1.50	1.57	1.62	1.67	1.72	1.76	1.79	1.84	1.88	1.99
1.50	1.57	1.62	1.67	1.72	1.77	1.79	1.84	1.89	2.00
1.52	1.57	1.62	1.67	1.73	1.77	1.80	1.85	1.90	2.01
1.52	1.58	1.63	1.68	1.73	1.77	1.80	1.86	1.90	2.15

Voor een overzicht is het beter de gegevens in een aantal klassen (meestal tussen 5 en 20) in te delen. Hiervoor bepalen we de hoogste en de laagste waarde in de gegevens en verdelen het spreidingsgebied (= [laagste , hoogste]) in het vooraf vastgelegde aantal klassen:]1.40,1.50] ,]1.50,1.60] ,]1.60,1.70] , ... Je kan ofwel de linkerklassegrens ofwel de rechterklassegrens in het interval meenemen, maar nooit beiden (uitgezonderd voor het eerste of laatste interval). De klassen mogen immers niet overlappen. In ons voorbeeld hebben we dus een klasbreedte

van 0.10. We tellen dan hoeveel maal een meting in een bepaald deelinterval valt. Dit geeft ons de frequentieverdeling of frequentiedistributie.

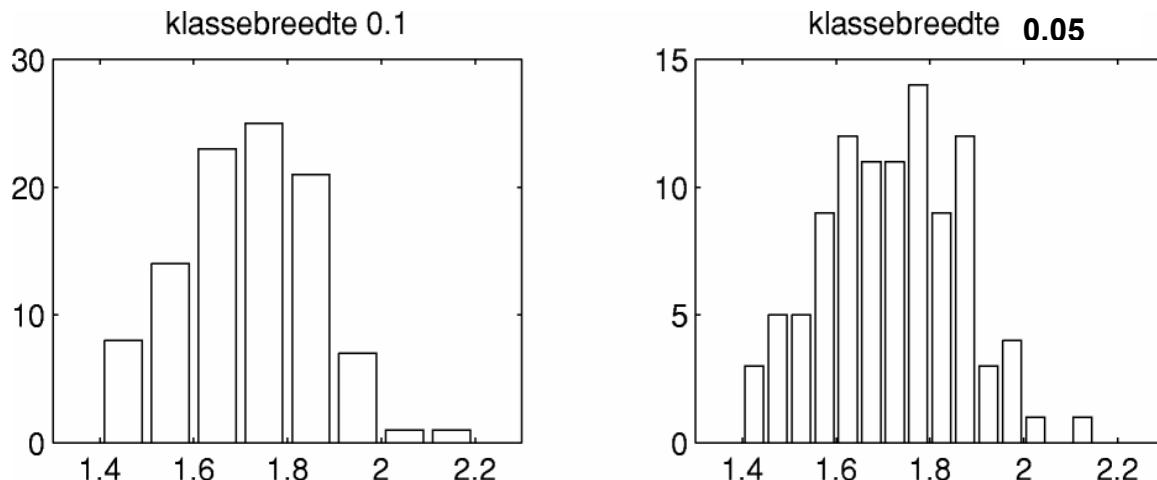
Klasse	Ondergrens	Bovengrens	Frequentie	Relatieve frequentie (%)	Cumulatieve frequentie (%)
1	1.4	1.5	8	8	8
2	1.5	1.6	14	14	22
3	1.6	1.7	23	23	45
4	1.7	1.8	25	25	70
5	1.8	1.9	21	21	91
6	1.9	2.0	7	7	98
7	2.0	2.1	1	1	99
8	2.1	2.2	1	1	100

Tabel 2.3

Grafisch kunnen we deze klasseindeling weergeven met behulp van een histogram. Op ieder deelinterval richten we een rechthoek op waarvan de breedte de klassenbreedte is en dus voor elke rechthoek even groot. De lengte (hoogte) is gelijk aan de frequentie van de desbetreffende klasse (f_i = frequentie van de i^{de} klasse).

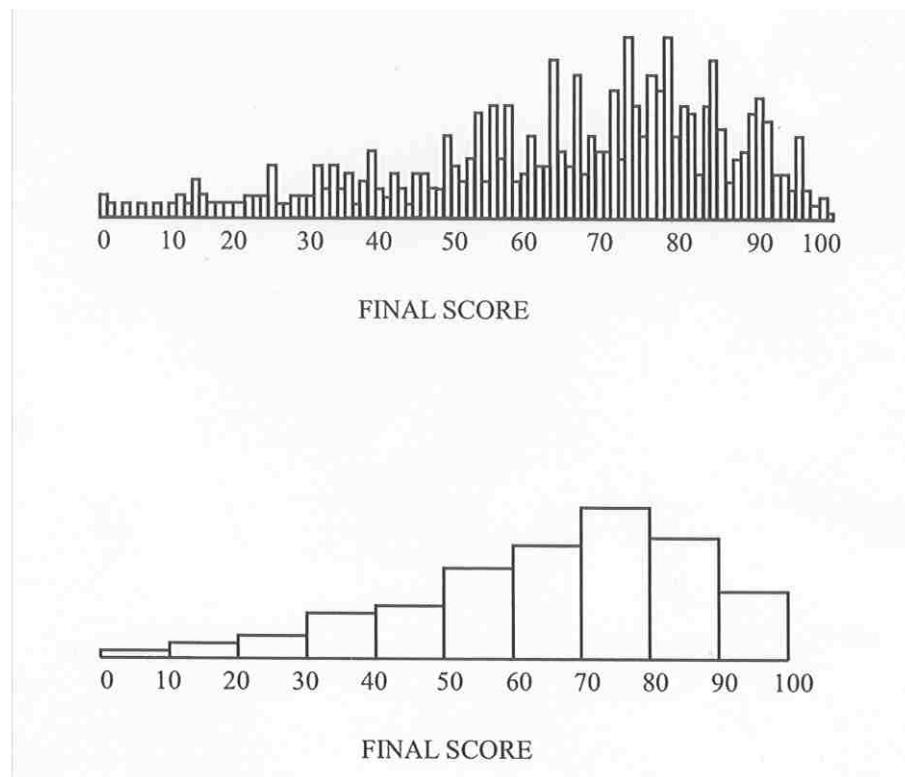
De som van de hoogtes van de rechthoeken komt overeen met de grootte van de steekproef.

In figuur 2.1 wordt het histogram van onze lengtemetingen afgebeeld met een klassenbreedte van 0.1 en 0.05. Zoals je ziet kan de vorm van het histogram vrij sterk van de keuze van de klassenbreedte afhangen.



Figuur 2.1: lengtemetingen afgebeeld met een klassenbreedte van 0.1 en 0.05

Aangezien een histogram bedoeld is om grafisch een eerste samenvattend idee te krijgen over onze waarnemingen, is het beter niet te veel klassen te nemen (meestal tussen 5 en 20) anders krijgen we een te gedetailleerde figuur. Vergelijk ter illustratie de twee onderstaande histogrammen, die dezelfde ruwe data beschrijven.



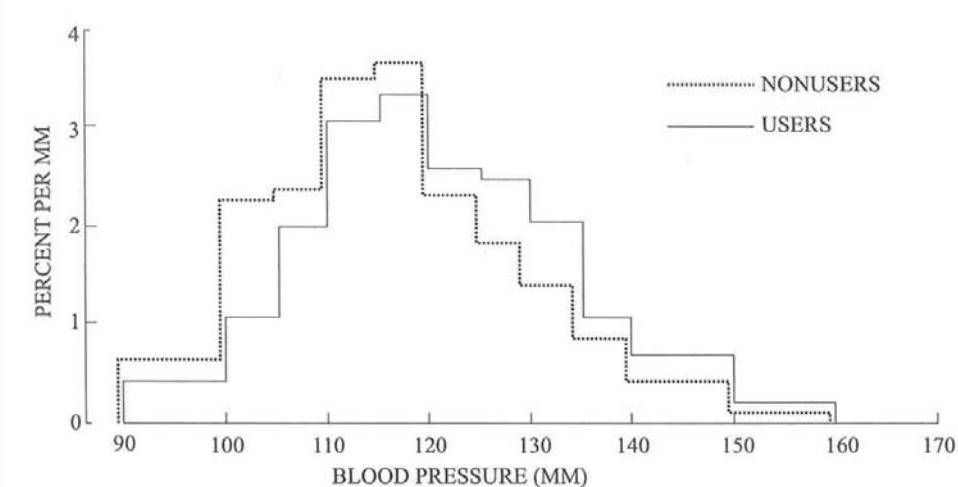
Figuur 2.2

2.1.2. Het histogram van de relatieve frequentie (De dichtheidsschaal)

In de bovenstaande histogrammen werd de absolute frequentie voorgesteld, we kunnen echter analog te werk gaan en de relatieve frequentie (frequentie uitgedrukt in percentages) voorstellen. Dit betekent dat we als vertikale schaal het percentage der waarnemingsgegevens kiezen. Deze verticale as wordt de dichtheidsschaal genoemd.

De hoogte y_i van de i -de rechthoek van het histogram correspondeert in de dichtheidsschaal met: $y_i = 100 * f_i / n$ (relatieve procentuele frequentie) of f_i / n (relatieve frequentie).

De som van de hoogtes van de rechthoeken komt nu overeen met 100% of 1. In ons voorbeeld van de lengtes is de relatieve procentuele frequentie gelijk aan de absolute frequentie, omdat de steekproefgrootte n gelijk aan 100 is. Beide histogrammen zijn dan ook gelijk (zie Tabel 2.3).



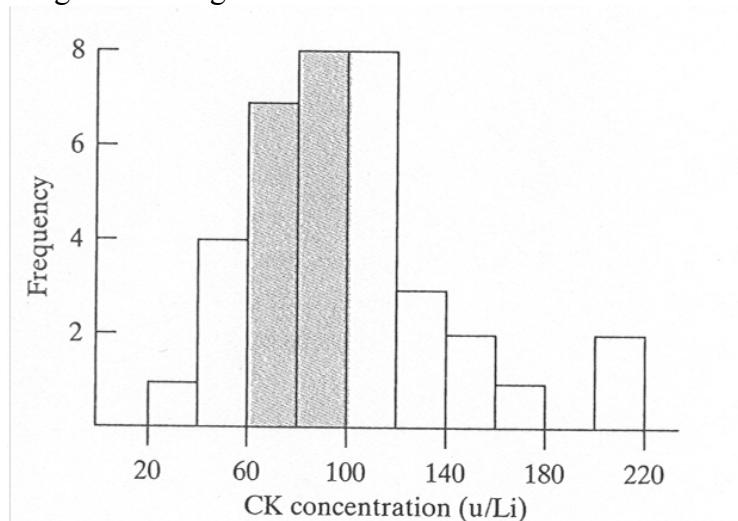
Figuur 2.3

Histogrammen met een dichtheidsschaal kunnen gemakkelijk met elkaar vergeleken worden. In figuur 2.3 vind je histogrammen van de systolische bloeddruk van volwassenen. De volle lijn weerspiegelt de metingen van 1747 volwassenen, die een bepaald geneesmiddel innemen, terwijl de stippellijn de bloeddruk van 3040 niet gebruikers weerspiegelt. Ondanks het verschil in grootte van beide groepen is de vergelijking zinvol.

De interpretatie van oppervlakte bij histogrammen: ongelijke klassenbreedtes

We illustreren dit aan de hand van het volgende voorbeeld.

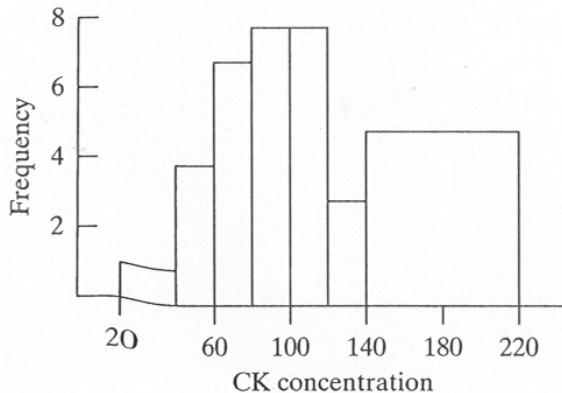
Creatine phosphokinase (CK) is een enzyme dat een rol speelt in de werking van de spieren en de hersenen. In een studie om de natuurlijke variatie van de CK-concentratie in het bloed te bepalen, werd de CK-concentratie in het bloed van 36 mannen gemeten. De resultaten zijn in figuur 2.4 in een histogram samengevat.



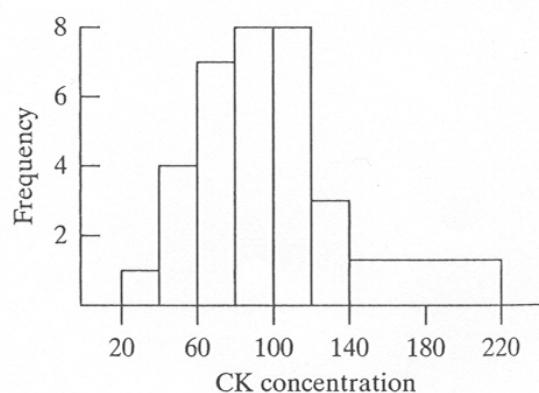
Figuur 2.4

We kunnen dit histogram op 2 manieren bekijken. De hoogtes van de rechthoeken geven een beeld van de distributie van de gegevens, maar ook hun oppervlaktes hebben een betekenis. De oppervlakte van elke rechthoek is evenredig met de corresponderende frequentie. Bijgevolg kunnen ook de oppervlaktes van één of meerdere rechthoeken geïnterpreteerd worden als de frequentie van de metingen in de bewuste klasse(n). In figuur 2.4 is de gearceerde oppervlakte gelijk aan 42% van de totale oppervlakte (som van de oppervlakte van alle rechthoeken). Dit betekent dat 42% van de metingen tussen 60 en 100 liggen.

In het bovenstaande namen we steeds klassen met een gelijke breedte, om geen vertekend beeld van de data te krijgen. Toch zal men sporadisch histogrammen maken met ongelijke klassenbreedtes. Men zal soms naburige klassen met een lage frequentie tezamen nemen. In ons voorbeeld kunnen we de 4 laatste klassen van 140 tot 220 tezamen nemen. Deze grote klasse zou dan een frequentie van 5 hebben. Indien we dit voorstellen krijgen we het histogram van figuur 2.5(a). Het is duidelijk dat dit histogram een vertekend beeld van de distributie geeft. Tevens is de oppervlakte niet meer evenredig met de frequentie. Dit probleem kan opgelost worden door de frequentie van de grote klasse te delen door 4, omdat deze klasse 4 keer zo breed is als de andere klassen (Figuur 2.5(b)). Merk op dat de hoogte van de grote klasse nu het gemiddelde is van de frequentie van de 4 oorspronkelijke klassen. Op deze manier worden zowel de vorm van het histogram als de evenredigheid tussen oppervlakte en frequentie bewaard.



(a)

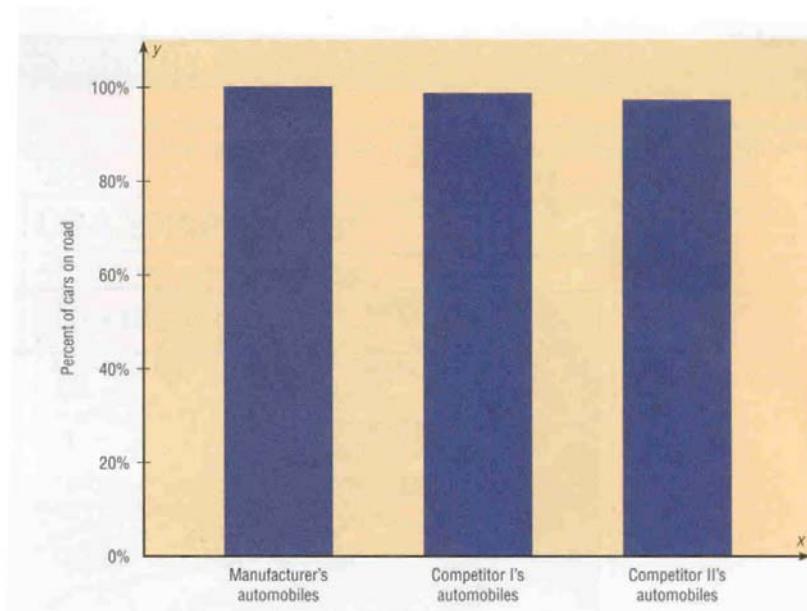


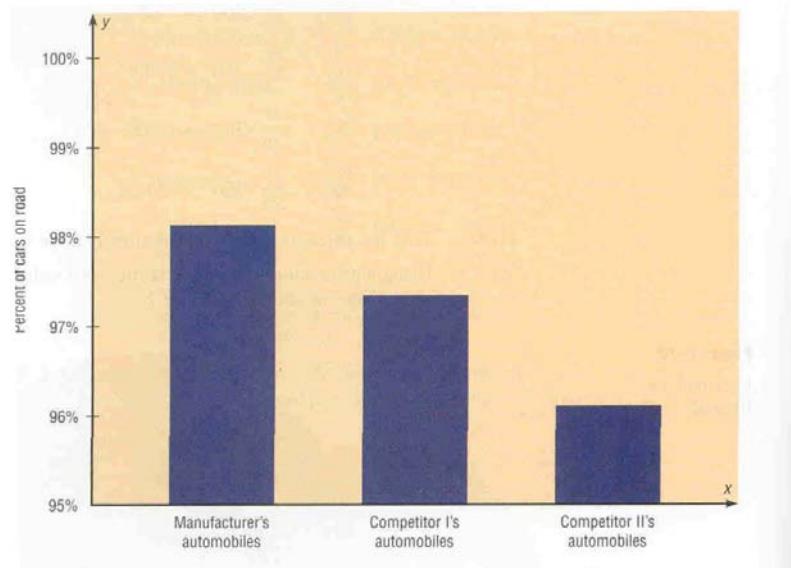
(b)

Figuur 2.5

Misleidende diagrammen

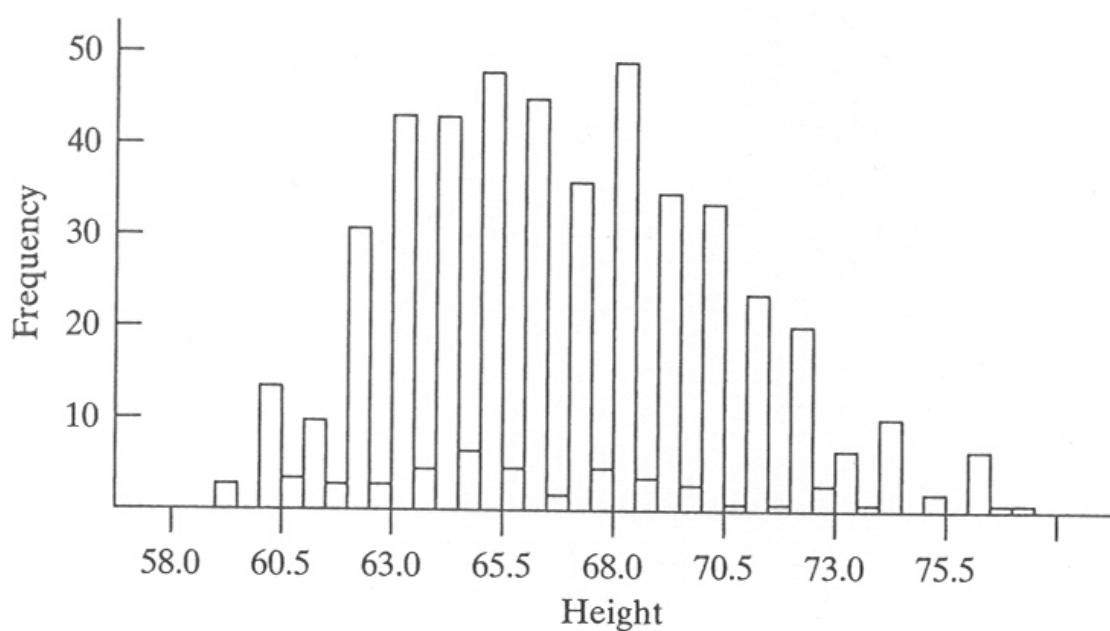
1. Indien de klassenbreedtes niet gelijk zijn voor elke klasse, kunnen we het histogram aanpassen zoals in figuur 2.5(b). De hoogte van elke balk correspondeert nu niet meer met de frequentie van de desbetreffende klasse. Dit geldt wel indien we alle klassen, ondanks hun verschil in breedte, grafisch even breed zouden voorstellen.
2. Het kan echter gebeuren dat we een kleinste en/of grootste klasse hebben met een open einde; we kijken dan naar alle data respectievelijk kleiner en/of groter dan de andere klassengrens. Indien we deze klasse op het histogram met een gelijke breedte voorstellen als de andere klassen, blijft de hoogte van elke balk evenredig met de frequentie.
3. Let op de schaal. We geven hier een grafisch voorbeeld, waarbij het percentage van het aantal wagens, dat nog in het verkeer is na tien jaar, voor één fabrikant en twee van zijn concurrenten wordt voorgesteld. We doen dit m.b.v. twee histogrammen met een verschillende schaal. Welke conclusie trek je uit beide histogrammen ?





Figuur 2.6

4. Soms kan de aard van je data zodanig zijn dat je een vertekend histogram krijgt. In figuur 2.7 wordt de lengte van 510 studenten (in inches) voorgesteld met behulp van 37 klassen met een breedte van 0.5 inch. De lengtes werden niet gemeten, maar er werd aan de studenten gevraagd hun lengte mede te rapporteren.



Figuur 2.7

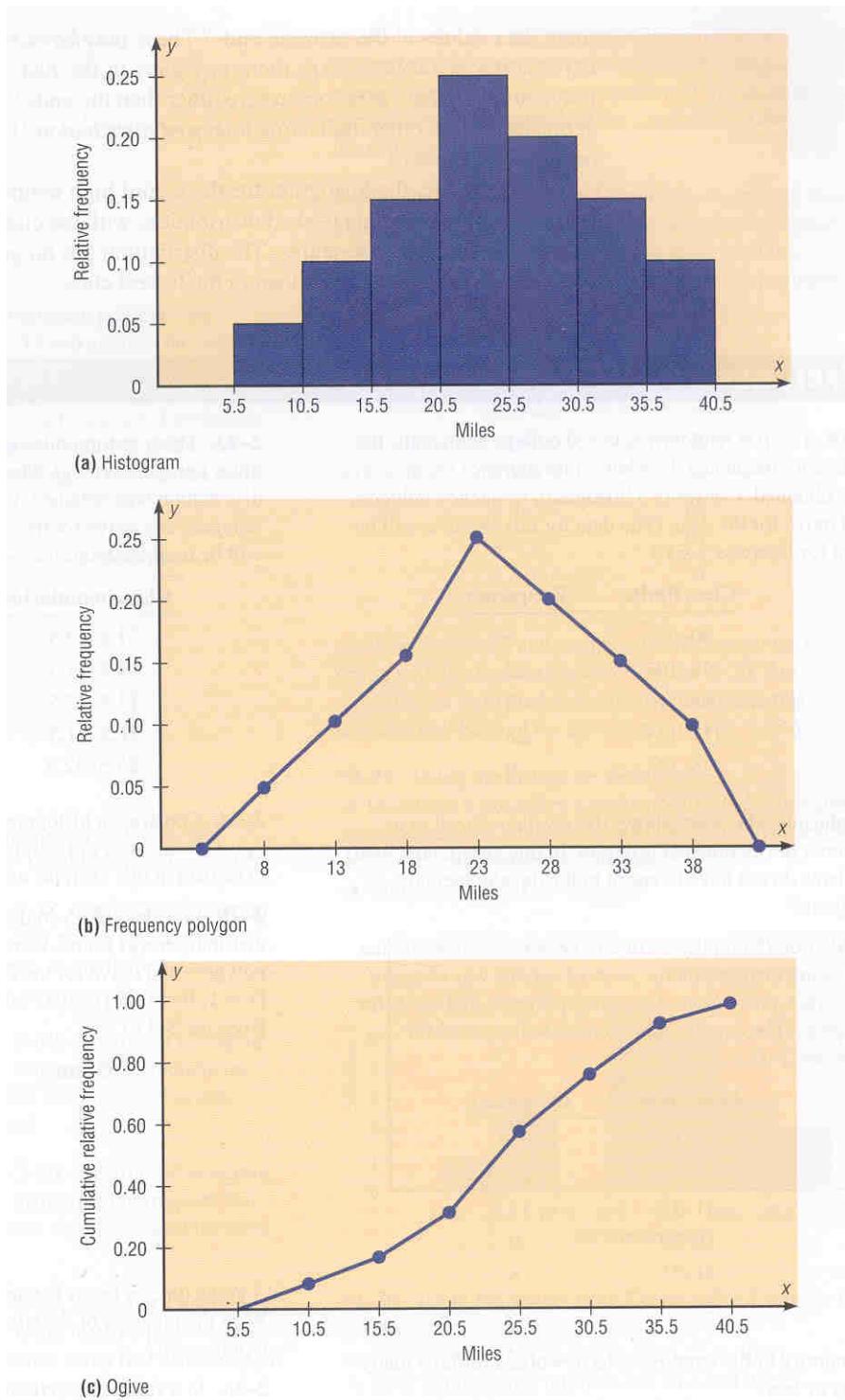
Verklaar de vorm van dit histogram

2.1.3. Cumulatief frequentiediagram

We kunnen de tabel met de gegevens van de lengtes uitbreiden met de cumulatieve frequentie (Tabel 2.3) en deze frequentie met behulp van een histogram voorstellen. We kunnen dit zowel doen met de absolute als met de relatieve frequentie. Meestal echter zullen cumulatieve frequenties met behulp van ogen worden voorgesteld.

2.1.4. Polygonen en ogieven

Indien we de toppen van een histogram ter hoogte van de klassemiddens verbinden door lijnstukken spreken we van een polygoon. Indien we dit doen ter hoogte van de rechterklassengrenzen voor het cumulatief frequentiehistogram spreken we van een ogief. Hieronder is een voorbeeld gegeven. Het aantal km, dat 20 random gekozen joggers per dag lopen, wordt voorgesteld m.b.v. een histogram, een polygoon en een ogief. Dit gebeurt telkens voor de relatieve frequenties. De ogief zal voor grote waarden van x steeds naar 1 gaan. De grafieken moeten er voor de absolute frequentie, op de schaal van de Y-as na, identiek uitzien.



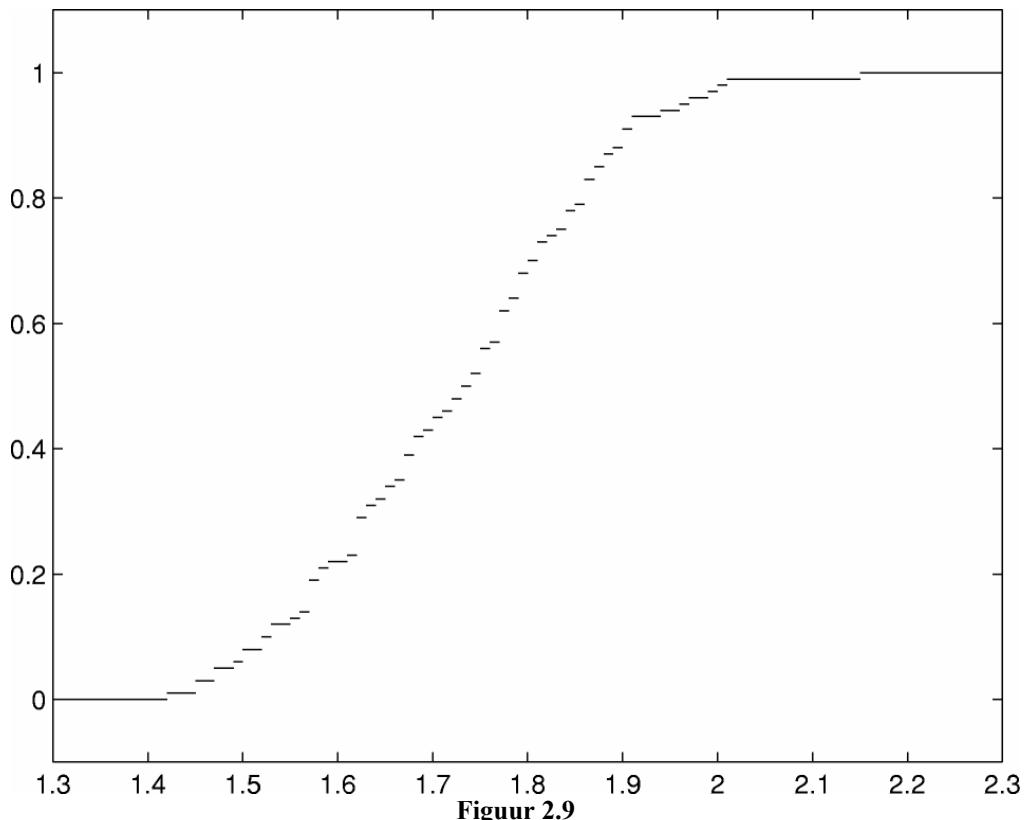
Figuur 2.8

2.1.5. Empirische verdelingsfunctie

Het cumulatief frequentiehistogram benadert de grafiek van de empirische verdelingsfunctie F_n , gedefinieerd als:

$$F_n(x) = \frac{\text{aantal metingen} \leq x}{n}$$

Deze functie $F_n(x)$ stelt dus het aantal metingen voor met een waarde kleiner dan of gelijk aan x gedeeld door het totale aantal metingen. In figuur 2.9 is deze weergegeven voor de data van Tabel 2.1. Deze functie is een trapfunctie die in de datapunten x_i een sprong maakt. De hoogte van de sprong is een veelvoud van $1/n$.



Figuur 2.9

Deze functie $F_n(x)$ benadert de kans om een volwassen mannelijke inwoner van België aan te treffen met lengte kleiner dan of gelijk aan x .

2.1.6. Stengel-en-blad diagram (“stem-and-leaf plot”)

De statisticus John Tukey vond een manier om snel gegevens samen te vatten en/of te ordenen en toch de individuele gegevenspunten te houden. Deze heet het stengel-en-blad diagram.

Om een stengel-en-blad figuur te maken moeten we beginnen met elk waarnemingsgetal op te splitsen in 2 delen: een stam en een blad. De plaats waar we het getal splitsen (op de eenheid, het tiental,...) hangt af van verschillende factoren zoals het aantal gegevens, het bereik, de eventuele aanwezigheid van uitschieters en eigen ervaring en oordeel. We illustreren de methode met het voorbeeld van de lichaamslengte.

De gegevens splitsen we tussen de tienden en de hondersten. Zo wordt bijvoorbeeld 1.68 gesplitst in 1.6 (de stengel) en 8 (het blad). De kleinste stengel is dus 1.4 en de grootste 2.1.

We maken nu een kolom met alle stengels; voeg nadien voor elk datapunt het laatste cijfer (het blad), aan de bijbehorende rij (de stengel) toe.

1.4	:	2 5 5 7 7 9
1.5	:	0 0 2 2 3 3 5 6 7 7 7 7 8 8 9
1.6	:	1 2 2 2 2 2 3 3 4 5 5 6 7 7 7 7 8 8 9
1.7	:	0 0 1 2 2 3 3 4 4 5 5 5 6 7 7 7 7 8 8 9 9 9
1.8	:	0 0 1 1 1 2 3 4 4 4 5 6 6 6 7 7 8 8 9
1.9	:	0 0 0 1 1 4 6 7 9
2.0	:	0 1
2.1	:	5

Figuur 2.10

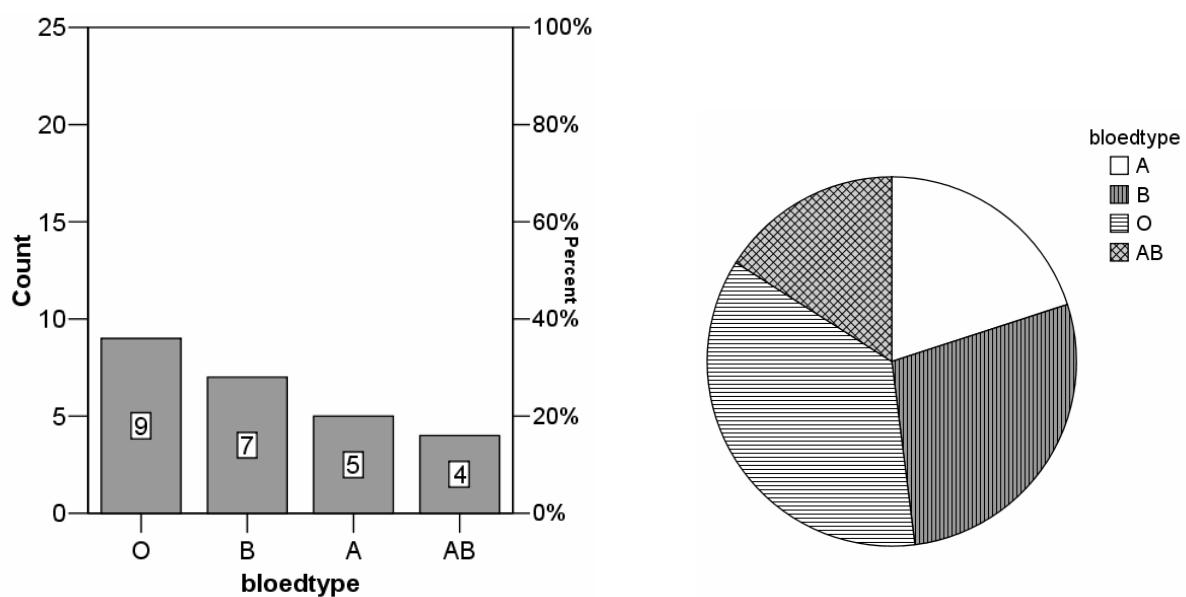
Merk op dat de vorm van deze tabel aan een “liggend” histogram doet denken.

Wanneer de oorspronkelijke data niet geordend zijn, zullen in eerste instantie de bladeren in willekeurige volgorde voorkomen. Het is dan nuttig de bladeren te ordenen volgens grootte.

2.1.7. Paretodiagram en taartdiagram

Indien we met kwalitatieve of categorische data werken, kunnen we natuurlijk ook een frequentieverdeling en een histogram tekenen, waarin nu de verschillende klassen van het histogram (nu ook wel Paretodiagram genoemd) corresponderen met de mogelijke waarden van de variabele. De klassen worden gerangschikt zodat de klasse met de hoogste frequentie links staat en deze met de laagste frequentie rechts. Ook wordt in dit geval veelvuldig met een taartdiagram (“pie graph”) gewerkt, waarvan de oppervlakte van de taartpunten met de frequentie correspondeert. Hieronder staat een voorbeeld van een histogram voor het bloedgroep type van 25 soldaten.

Bloedtype	Freq.	Rel. freq. (%)
A	5	20
B	7	28
O	9	36
AB	4	16
	25	100



Figuur 2.11

2.1.8 Scatterdiagram

Indien we metingen hebben, die van twee variabelen afhangen of indien we een experimentele studie hebben uitgevoerd met 1 afhankelijke en 1 onafhankelijke variabele, zullen we de data meestal voorstellen in een scatterdiagram. Dit scatterdiagram geeft ons dadelijk informatie over de relatie tussen de twee voorgestelde variabelen. We behandelen dit in paragraaf 2.4, waar figuur 2.22 een voorbeeld van een scatterdiagram is.

2.2. Data samenvatten

In het tweede deel van dit hoofdstuk over beschrijvende statistiek zullen we enkele klassieke methodes beschrijven om kwantitatieve gegevens samen te vatten met behulp van getallen. We behandelen hier de volgende thema's:

- Kengetallen van centrale tendens of centrummaten
- Kengetallen van variantie of spreidingsmaten
- Kengetallen voor de positie van een datapunt
- Symmetrie van de verdeling
- Technieken van "Exploratory Data Analysis" (EDA):
 - de 5-getallen samenvatting en de boxplot
- Meerdimensionale data: de covariantiematrix

We vermelden hier het verschil tussen een statistiek en een parameter:

- Een statistiek = karakteristiek of grootheid bepaald met de data uit de steekproef. We noemen dit empirische grootheden
 - Een parameter = karakteristiek of grootheid bepaald met de data van de totale populatie
- Een statistiek zullen we voorstellen door een romeinse letter en een parameter door de overeenkomstige griekse letter. Zo zal het populatiegemiddelde aangeduid worden met μ en het steekproefgemiddelde met m . Dit wordt later wel duidelijk.

In dit hoofdstuk noteren we onze metingen na ordening $x_1, x_2, x_3, x_4, \dots, x_n$. De kleinste waarde van onze metingen is x_1 en de grootste x_n . n is de steekproefgrootte.

We beschouwen enerzijds de dataset van Tabel 2.2, de ruwe data, die we dataset 1 noemen, als de gegroepeerde data (tabel 2.3), die we dataset 2 noemen evenals de volgende dataset:

dataset 3: 8 8 8 9 9 10 11 12 12 13 1000

Bemerkt dat dataset 3 één afwijkende waarde bevat, namelijk $x_{11} = 1000$.

2.2.1. Kengetallen van centrale tendens of centrummaten

"Average" when you think of it, it is a funny concept. Although it describes all of us, it describes none of us... While none of us wants to be the average man, we all want to know about him or her.

2.2.1.1. Het gemiddelde

Het meest gebruikte kengetal om het "centrum" van gegevens te karakteriseren is het rekenkundig gemiddelde:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Voor gegroepeerde data nemen we:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_{mi}$$

waarbij x_{mi} het middelpunt en f_i de frequentie van het i -de klasse is; k is het aantal klassen en n de steekproefgrootte.

Dit levert voor dataset 1: $\bar{x} = 1.724$, dataset 2 : $\bar{x} = 1.717$, en dataset 3: $\bar{x} = 100$

Het spreekt voor zich dat \bar{x} een goede schatting voor het populatiegemiddelde μ is (zie later).

2.2.1.2. De mediaan

De mediaan is het middelste datapunt, na ordening van de data volgens grootte. Bij een even steekproefgrootte is de mediaan het gemiddelde van de middelste twee datapunten.

$$\text{Dus: } med = \begin{cases} \frac{x_{\frac{n+1}{2}}}{2} & n \text{ oneven} \\ \frac{1}{2} \left(x_{\frac{n}{2}} + x_{\frac{n+1}{2}} \right) & n \text{ even} \end{cases}$$

Bij gegroepeerde gegevens zoeken we eerst de mediaanklasse: de klasse waarin het middelste datapunt ligt. De mediaan zelf wordt dan gedefinieerd met mbv percentielen (zie verder). Dit levert voor dataset 1: $med = 1.735$; dataset 2: mediaanklasse = $[1.70, 1.80]$; en dataset 3: $med = 10$.

2.2.1.3. De modus

De modus is de waarde die het meest voorkomt. Bij gegroepeerde gegevens spreken we van de modale klasse, de klasse met de hoogste frequentie. Deze modale klasse kan ook gevonden worden bij kwantitatieve of categorische data.

Dataset 1: $mod = 1.62$; dataset 2: modale klasse = $[1.70, 1.80]$; dataset 3: $mod = 8$

2.2.1.4. De “midrange”

De midrange is het gemiddelde van de hoogste en de laagste waarde. Dus:

$$\text{midrange} = \frac{x_1 + x_n}{2}$$

Dit levert voor dataset 1: midrange = 1.785 en voor dataset 3: midrange = 504

2.2.1.5. Het gewogen gemiddelde

Bij het berekenen van het examengemiddelde voor elke student, zal er rekening gehouden worden met de studiepunten voor elk vak. In dit geval spreken we van een gewogen gemiddelde. Dit gewogen gemiddelde wordt gedefinieerd als:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i w_i$$

of voor gegroepeerde data:

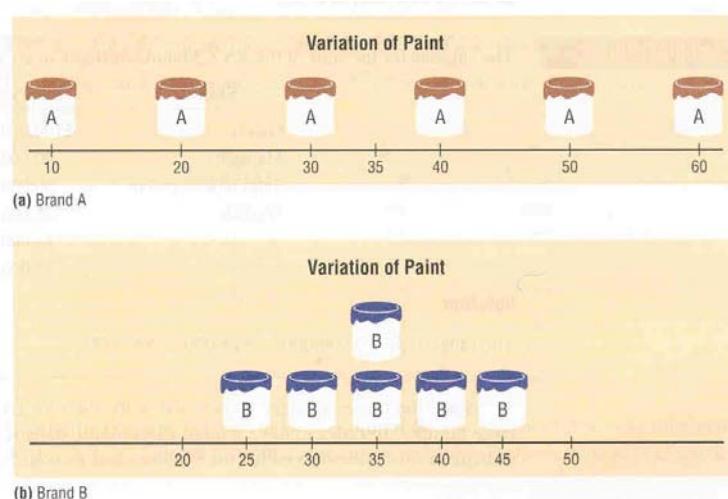
$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_{mi} w_i$$

waarbij w_i het gewicht van de i -de meting (of klasse) is. Er geldt $\sum_{i=1}^n w_i = n$. We zullen gewogen gemiddeldes en sommen later ook gebruiken, wanneer we bijvoorbeeld metingen hebben waarvan er een gedeelte veel accurater gemeten is dan een ander gedeelte.

N.B.: Als we in dataset 3 de uitschieter 1000 weglaten is het gemiddelde 10. De mediaan evenals de modus worden het minst beïnvloed door deze meting. We noemen dit dan ook robuuste parameters.

2.2.2. Kengetallen voor variantie of spreidingsmaten

In figuur 2.12 wordt het aantal maanden vooraleer outdoor verf verkleurt voorgesteld en dit voor 2 merken (merk A en merk B). Alhoewel de centrale tendens voor beide merken dezelfde is, zijn beide datasets verschillend qua spreiding.



Figuur 2.12

Naast een maat voor de centrale tendens wil men ook een maat voor de spreiding in de data. We bespreken hier de range, de variantie, de standaardafwijking en de gemiddelde en mediaan absolute afwijking.

2.2.2.1. Het bereik of de range

Range = $R = \text{hoogste} - \text{laagste waarde} = x_n - x_1$.

Dit levert voor dataset 1: $R = 0.73$ en voor dataset 3: $R = 992$

Bemerk dat uitschieters de range zeer sterk beïnvloeden

2.2.2.2. Variantie en standaardafwijking

De variantie is het gemiddelde van de kwadratische afwijkingen van het gemiddelde.

De **variantie** van een populatie wordt gedefinieerd als:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N} \quad \text{waarbij } \mu \text{ het populatiegemiddelde is en } N \text{ de populatiegrootte}$$

Voor de ruwe data van een steekproef met steekproefgrootte n definiëren we analoog

$$s_n^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Bemerk dat we hier door $n-1$ delen in plaats van door n . Zoals we later zullen zien levert dit een betere schatter voor de populatievariantie op.

De **standaardafwijking** is de wortel uit de variantie.

Dit levert voor dataset 1: $s_n^2 = 0.022$; $s_n = 0.15$ en voor dataset 3: $s_n^2 = 89103$; $s_n = 299$

Dikwijls worden voor beide grootheden de volgende verkorte formules gebruikt:

$$s_n^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} \quad \text{en} \quad s_n = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}}$$

In sommige boeken wordt deze formule zelfs geprezen omdat er minder optellingen voor nodig zijn. Het resultaat kan echter zeer onnauwkeurig zijn ten gevolge van afronding.

Voor gegroepeerde data definiëren we:

$$s_n^2 = \frac{\sum f_i x_{mi}^2 - n\bar{x}^2}{n-1} \quad \text{en} \quad s_n = \sqrt{\frac{\sum f_i x_{mi}^2 - n\bar{x}^2}{n-1}}$$

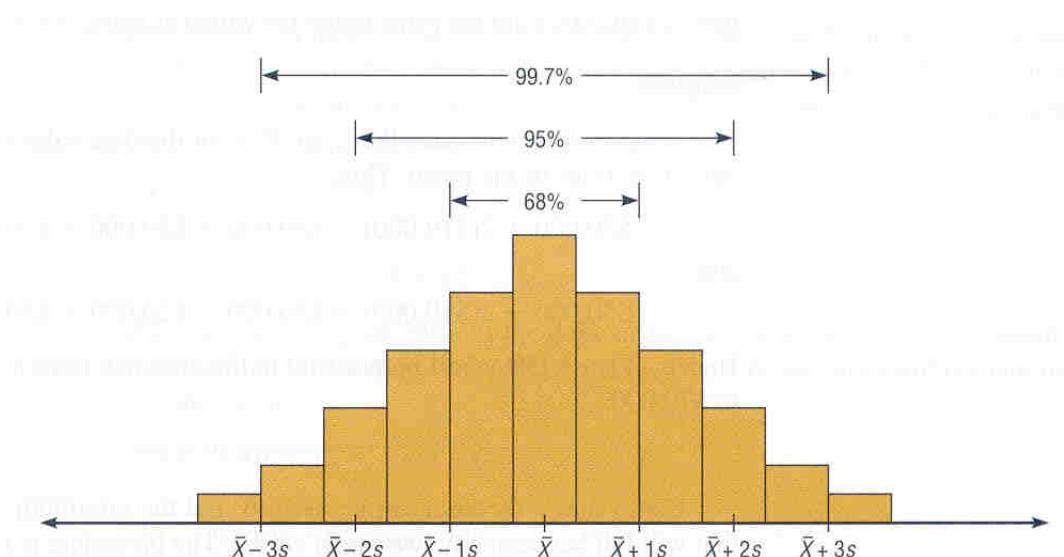
N.B.: Let wel dat als bijvoorbeeld onze gegevens lengten zijn uitgedrukt in "inch", s (de standaardafwijking) een lengte is eveneens uitgedrukt in "inch" terwijl de variantie dan een oppervlakte is; als we de gegevens vervolgens herschalen naar "cm" door ze te vermenigvuldigen met 2.54 moeten we s met dezelfde factor vermenigvuldigen, terwijl de variantie met het kwadraat van 2.54 vermenigvuldigd moet worden.

Chebyshev's theorema

De fractie van de data die dichter bij het gemiddelde ligt dan k keer de standaardafwijking is minstens $(1 - 1/(k^2))$.

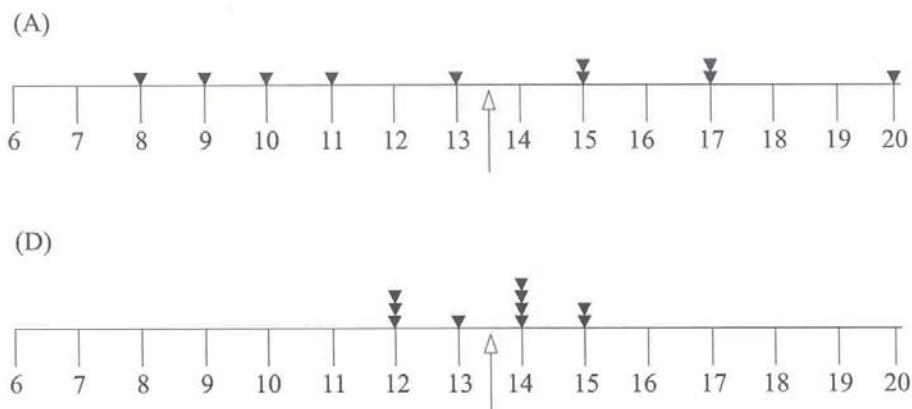
Dit betekent dat minstens 75% van de data op minder dan 2 standaardafwijkingen van het gemiddelde ligt en 89% op minder dan 3 standaardafwijkingen.

Voor data die normaal verdeeld zijn kunnen we dit nog verbeteren. (Zie later en figuur 2.13)



Figuur 2.13

N.B.: Fysisch houdt het begrip variantie verband met het traagheidsmoment en het begrip gemiddelde met het massacentrum.



Figuur 2.14

2.2.2.3. De mediaan en de gemiddelde absolute afwijking (median absolute deviation of MAD en mean absolute deviation of MeanAD)

De MAD en MeanAD worden gedefinieerd als de mediaan respectievelijk het gemiddelde van de absolute afwijkingen t.o.v. de steekproefmediaan:

$$\text{MAD} = \text{mediaan } \{ |x_i - \text{med}| \}_{i=1}^n \quad \text{en} \quad \text{MeanAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \text{med}|$$

De helft van de metingen liggen tussen med-MAD en med+MAD.

Dit levert voor dataset 1: MAD= 0.11 ; MeanAD = 0.12 en voor dataset 3: MAD = 2; MeanAD = 91

N.B.: Soms wordt er ook een coëfficiënt van variatie gedefinieerd: $\text{CVar} = \frac{s}{\bar{x}} \cdot 100\%$. We gaan hier niet verder op in.

2.2.3. Kengetallen voor de positie van een datapunt

Stel dat een student 14 heeft op statistiek en 13 op wiskunde. Het is moeilijk deze scores met elkaar te vergelijken. Met behulp van de hieronder gedefinieerde getallen gaan we dit toch proberen.

Voor elk datapunt zullen we getallen definiëren, die ons een idee geven van de positie van dat punt binnen de totale verzameling van datapunten. We behandelen hier de meest gebruikte, namelijk de z-score en de percentielen.

2.2.3.1. De standaard score of z-score

De z-score van een datapunt x_i wordt gedefinieerd als:

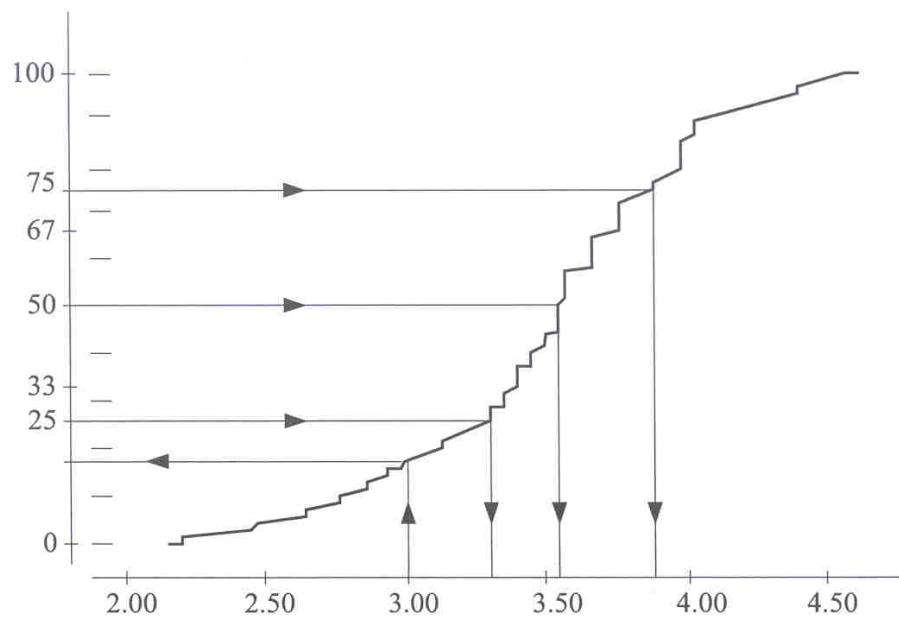
$$z = \frac{x_i - \bar{x}}{s} \quad (\text{voor populaties : } Z = \frac{X - \mu}{\sigma})$$

Indien voor de vakken statistiek en wiskunde het groepsgemiddelde respectievelijk 13 en 12 was en de standaardafwijking respectievelijk 2 en 1, levert dit voor onze student een z-score van respectievelijk 0.5 en 1. Ondanks een lager cijfer op het vak wiskunde is de score voor dit vak in vergelijking tot de totale groep toch beter dan de score voor statistiek.

Als alle datapunten worden omgezet in z-scores zullen de verkregen getallen een gemiddelde 0 hebben en een variantie van 1.

2.2.3.2. Percentielen, kwantielen en het interkwartiel

In één van de vorige paragrafen behandelden we reeds de empirische verdelingsfunctie F_n , die het aantal metingen weergeeft met een waarde kleiner dan of gelijk aan x gedeeld door het totale aantal metingen. In de praktijk willen we vaak een antwoord op de omgekeerde vraag: "Voor welke waarde van x is 25% (of 50% of 90%) van de metingen kleiner dan of gelijk aan x ". We maken bijvoorbeeld deuren zo hoog dat minstens 99.9% van de mensen hun hoofd niet zal stoten en we moeten dus weten waar die grens ligt. Bij gegeven α zouden we het $\alpha\%$ -percentiel met $0 < \alpha < 100$ willen definiëren als het punt ξ_α waaronder $\alpha\%$ van de metingen liggen: we zoeken dus de inverse functie van F_n . Omdat we maar een eindig aantal metingen hebben en F_n constant is tussen ieder tweetal opeenvolgende metingen, bestaat zo'n inverse functie echter niet (of niet overal).



Figuur 2.15

Als we bijvoorbeeld de mediaan (dit is het 50%-percentiel) van de 100 metingen van tabel 2.2 willen bepalen, dan vinden we $x_{50}=1.73$ en $x_{51}=1.74$, zodat voor iedere x tussen deze twee waarden het percentage metingen ter linker zijde gelijk is aan 50%.

Voor een eenduidige waarde kozen we in dit geval het midden tussen deze twee punten reeds als mediaan. Als we echter in deze dataset de laatste meting $x_{100}=2.15$ schrappen, omdat deze lengte zeer uitzonderlijk is en dus waarschijnlijk een meet- of tikfout is, dan houden we 99 metingen over; 50% ervan geeft aanleiding tot het beschouwen van schimmige "halve" metingen. Bovendien is er geen punt x_i te vinden zodat precies 50% van de metingen kleiner dan of gelijk aan x_i is en ook 50% groter dan x_i . In dit geval definieerden we de mediaan als het punt waar de sprong van kleiner dan 50% naar groter dan 50% gemaakt wordt.

Voor een definitie van het (empirische) $\alpha\%$ -percentiel ξ_α doen we in feite hetzelfde. We definiëren dit als de $\frac{\alpha}{100}(n+1)$ -de waarneming en berekenen het als volgt. Als p het geheel

deel van het getal $\frac{\alpha}{100}(n+1)$ voorstelt en ρ de overschat (d.w.z. $\frac{\alpha}{100}(n+1) = p+\rho$), dan zal ξ_α ergens tussen x_p en x_{p+1} liggen. Via interpolatie zullen we ξ_α dan als volgt definiëren:

$$\xi_\alpha = x_p + \rho(x_{p+1} - x_p)$$

N.B.: Merk op dat het geen zin heeft om te spreken van een $\alpha\%$ -percentiel met $\alpha < \frac{100}{n+1}$ of $\alpha > \frac{100n}{n+1}$.

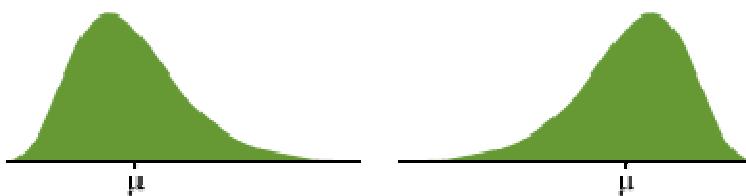
Tenslotte kunnen we in plaats van procenten ook fracties (tussen nul en een) beschouwen; we spreken dan van kwantielen. Het 0.2-kwantiel is dus het 20%-percentiel.

De meest gebruikte percentielen zijn die op 50% (de mediaan), 25% (**het linker- of eerste kwartiel: Q_1**) en 75% (**het rechter- of derde kwartiel: Q_3**)

Het **interkwartiel (IQR=interquartile range)** is het verschil tussen het derde (Q_3) en het eerste (Q_1) kwartiel: $IQR = Q_3 - Q_1$. Dit getal wordt tevens gebruikt als een spreidingsmaat. Bemerk dat de interkwartielafstand ongevoelig is voor uitschieters.

2.2.4. De symmetrie van een verdeling

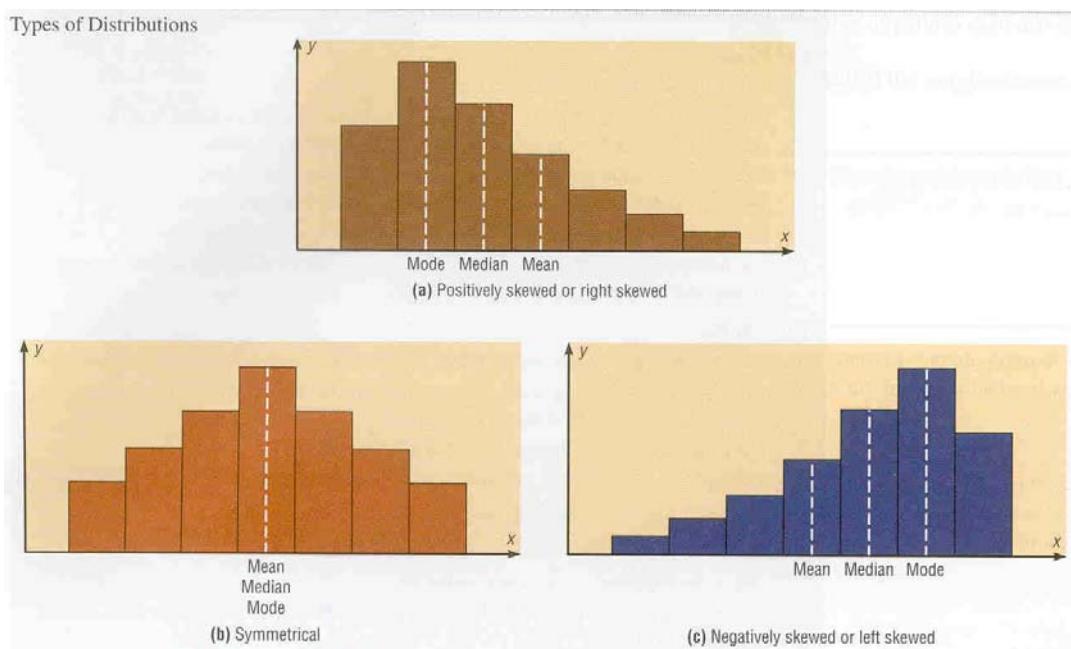
We kunnen ook getallen definiëren die de mate van symmetrie van de dataset weergeven. We noemen deze parameter de skewness (scheefheid).



Figuur 2.16 2 verdelingen met hetzelfde gemiddelde en standaardafwijking
links: positief of rechts skewed; rechts: negatief of links skewed

We gaan er pas later op in. We vermelden alleen de volgende eigenschappen.

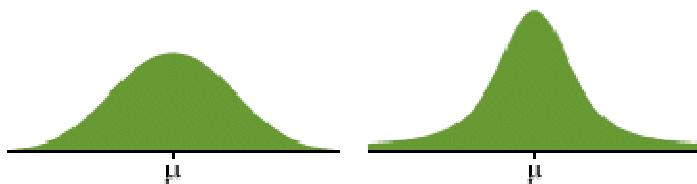
Een symmetrische verdeling is een verdeling waarbij de data langs beide kanten van het gemiddelde identiek verdeeld zijn. In dit geval zal het gemiddelde gelijk zijn aan de mediaan.



Figuur 2.17

Bij een positieve of rechts “skewed” verdeling zullen het merendeel van de data kleiner zijn dan het gemiddelde. Bij een negatieve of links “skewed” verdeling zullen het merendeel van de data daarentegen groter zijn dan het gemiddelde. Bovenstaande histogrammen illustreren de positie van gemiddelde, mediaan en modus in de drie gevallen.

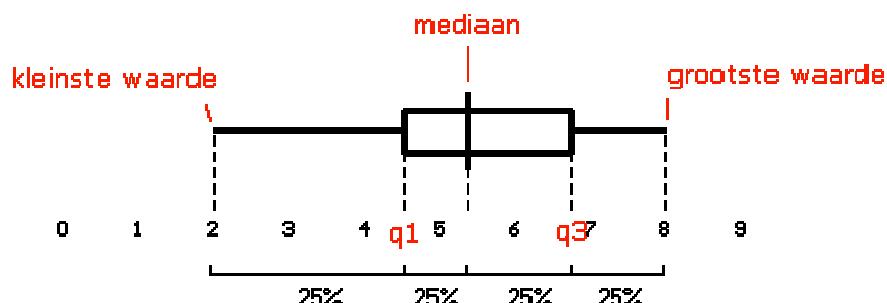
Een ander getal, namelijk de kurtosis, geeft een indicatie van de vorm van de distributie van de datapunten; hoe groter dit getal hoe meer gepiekt onze distributie. Hieronder vind je de afbeelding van een verdeling met een kleine (links) en een grote (rechts: meer gepiekt en “dikkere staart”) kurtosis.



Figuur 2.18

2.3. Exploratory Data Analysis (EDA): de 5-getallen samenvatting en de boxplot (box and whisker diagram)

In EDA worden de ruwe data meestal samengevat in 5 getallen: het minimum, het maximum, de mediaan, Q_1 en Q_3 . Deze getallen worden voorgesteld in een boxplot, zoals hieronder afgebeeld.



Figuur 2.19

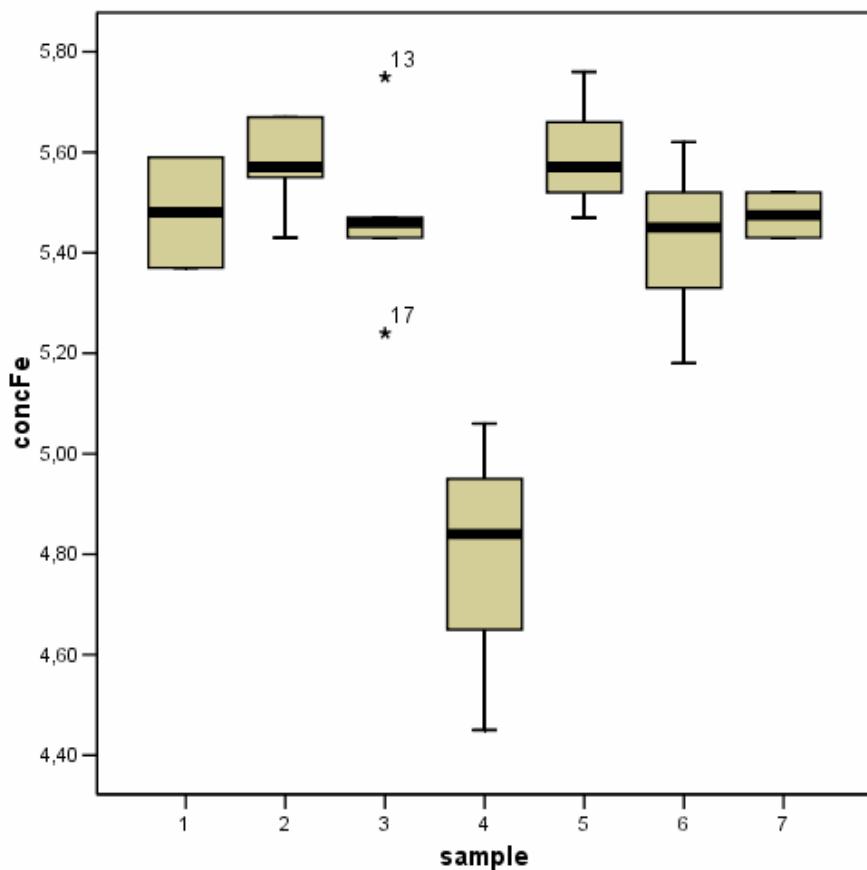
Hierbij wordt (horizontaal of verticaal) een as getekend gaande van de kleinste naar de grootste meting (het bereik). Op deze as worden de mediaan, het eerste en het derde kwartiel aangegeven met een dwarse streep. Van het stuk tussen het eerste en het derde kwartiel wordt een doosje (box) gemaakt.

Indien de mediaan zich in het midden van de box bevindt en de as langs beide kanten van de box even lang is, zal de verdeling der datapunten ongeveer symmetrisch zijn.

De boxplot wordt soms verfijnd om rekening te houden met uitschieters (outliers, aberante metingen). De identificatie en behandeling van uitschieters is niet eenduidig. In een boxplot zal men de datapunten die kleiner zijn dan $Q_1 - 1.5 \cdot IQR$ of groter dan $Q_3 + 1.5 \cdot IQR$ als uitschieters aanduiden met een sterretje. De as gaat dan van de kleinste naar de grootste meting, die geen uitschieter is.

Als we verscheidene datasets met elkaar willen vergelijken zoals in figuur 2.20 kan de boxplot een goed visueel hulpmiddel zijn. In deze figuur worden de resultaten van een bepaling van de Fe-concentratie op een multivitamine complex samengevat voor verschillende samples. Elk sample werd meermaals gemeten. Sample 4 heeft duidelijk een

andere samenstelling dan de andere samples. Ook is de variabiliteit van meetwaarden in de verschillende samples ongelijk. Misschien is de gebruikte meettechniek voor elk sample verschillend. In dat geval is de meettechniek gebruikt voor sample 3 duidelijk superieur. Deze techniek gaf wel twee aberante meetresultaten.



Figuur 2.20: Fe-concentratiebepaling in een multivitaminecomplex op 7 verschillende samples

2.4. Meerdimensionale data

Hieronder staan de cijfers die studenten Informatica (I1--I31) en Biotech (B1--B22) behaalden bij het schriftelijk examen en bij de computerproef in april '98.

St	CP	SE												
B01	7	11	B12	17	7	I01	11	8	I12	8	12	I23	17	14
B02	17	10	B13	17	13	I02	10	15	I13	13	10	I24	7	14
B02	5	6	B14	14	16	I02	16	16	I14	11	11	I25	13	14
B04	15	15	B15	11	16	I04	17	15	I15	15	14	I26	11	13
B05	6	12	B16	15	15	I05	18	16	I16	7	7	I27	10	9
B06	15	18	B17	18	16	I06	9	10	I17	15	16	I28	6	5
B07	14	16	B18	13	9	I07	15	13	I18	8	11	I29	14	13
B08	14	14	B19	12	11	I08	6	11	I19	13	14	I30	8	10
B09	10	9	B20	14	13	I09	12	12	I20	14	8	I31	14	13
B10	14	12	B21	16	13	I10	11	11	I21	4	6			
B11	10	11	B22	15	17	I11	15	13	I22	4	7			

Tabel 2.4: Geanonymiseerde resultaten van het schriftelijk examen (SE) statistiek en de computerproef (CP) in april '98 voor studenten Informatica en Biotech.

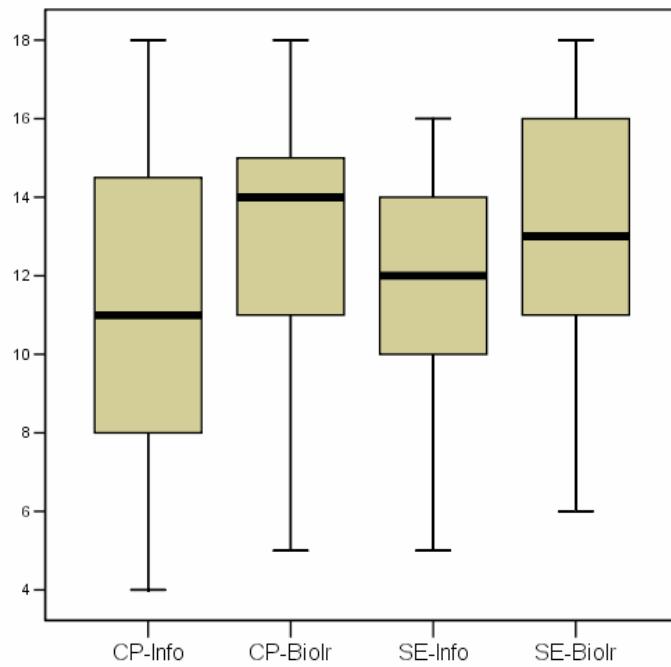
Deze data zijn tweedimensionaal omdat er voor iedere student twee cijfers zijn; bovendien betreft het twee groepen studenten. Later zullen we methoden behandelen om de twee cijferreeksen en de twee groepen met elkaar te kunnen vergelijken.

De kengetallen van deze dataset zijn:

	Info	Info	Biolt	Biolt	Tot	Tot
	CP	SE	CP	SE	CP	SE
Gemiddelde	11.4	11.6	13.1	12.7	12.1	12.1
Standaarddeviatie	3.91	3.05	3.62	3.27	3.86	3.16
Modus	11&15	13&14	14	16	14&15	13
Minimum	4	5	5	6	4	5
25%-percentiel	8	10	10.75	10.75	9.5	10
Mediaan	11	12	14	13	13	13
75%-percentiel	15	14	15.25	16	15	14.5
Maximum	18	16	18	18	18	18

Tabel 2.5: Kentallen voor bovenstaande data

De resultaten van de groepen Informatica en Biotech uit deze tabel kunnen we kwalitatief snel met elkaar vergelijken door er een boxplot van te maken:



Figuur 2.21: Boxplot van de cijfers van het schriftelijk examen (SE) statistiek in april '98 en de computerproef (CP) voor de studenten Informatica (Info) en Biotech (Bir).

Tot onze verbazing zien we dat de informaticastudenten juist de computerproef gemiddeld veel slechter deden dan de biotechstudenten en dat ook de resultaten van het schriftelijk examen iets lager lagen. Later zullen we technieken behandelen om te kunnen beslissen of de verschillen significant zijn, d.w.z. niet te wijten aan het toeval.

Tussen de cijfers die eenzelfde student behaalt voor het schriftelijk examen en de computerproef verwachten we een verband. In figuur 2.22 is voor iedere student het cijfer van het schriftelijk examen uitgezet tegen dat van de computerproef (de scatterplot). We zien grofweg een verband; studenten die goede cijfers behaalden voor het één, behaalden ook goede cijfers voor het ander.

Figuur 2.22: Cijfers van het schriftelijk examen statistiek in april'98 (verticaal) uitgezet tegen die van de computerproef (horizontaal) voor de studenten informatica (x) en Biotech (+); correlatiecoëfficiënt 0.60 .

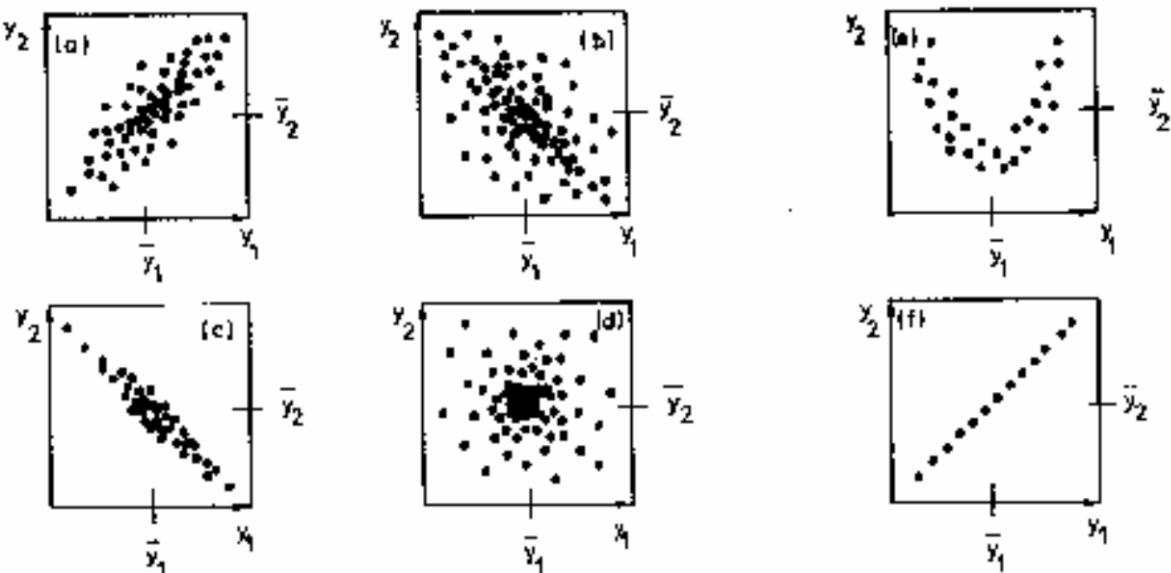
We kunnen hier (en bij meervoudige datasets in het algemeen) kengetallen definiëren die ons een idee geven van het (mogelijk lineair) verband tussen de grootheden. Numeriek kunnen we dit verband uitdrukken met behulp van de (empirische) covariantie en de correlatiecoëfficiënt (sample covariance en sample correlation).

Als $\mathbf{x} = \{x_i | i = 1, \dots, n\}$ en $\mathbf{y} = \{y_i | i = 1, \dots, n\}$ twee series van n metingen zijn (we kunnen ze beschouwen als vectoren in de n -dimensionale ruimte) met gemiddelden \bar{x} respectievelijk \bar{y} en standaarddeviaties s_x respectievelijk s_y , dan zijn:

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{en} \quad r(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{s_x s_y}$$

de empirische covariantie respectievelijk correlatie tussen \mathbf{x} en \mathbf{y} . Deze parameters beschrijven de sterkte van het lineair verband tussen de \mathbf{x} en \mathbf{y} -metingen.

Twee datasets heten ongecorreleerd als hun correlatie nul is; anders heten ze gecorreleerd. Men kan m.b.v. de ongelijkheid van Cauchy-Schwartz aantonen dat de correlatie in absolute waarde kleiner dan of gelijk aan 1 is. Indien ze een waarde heeft dicht bij 1 spreken we van een grote positieve correlatie en indien ze dicht bij -1 ligt over een grote negatieve correlatie.



Figuur 2.23: Scatterplots voor verschillende correlaties: a) $r=0.75$; b) $r=-0.32$; c) $r=-0.95$; d) $r=0$; e) $r=0$; f) $r=1$

Net zoals bij de variantie kunnen we de covariantie ook berekenen met de alternatieve formule:

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right)$$

Om dezelfde reden als tevoren is het gebruik hiervan af te raden.

Bemerk dat er in figuur 2.23(e) een duidelijk kwadratisch verband zichtbaar is tussen de variabelen x en y. Toch is de correlatie 0, omdat deze grootheid enkel een indicatie geeft over een lineair verband.

Meer algemeen kunnen we een situatie tegenkomen waarin we n onafhankelijke waarnemingen hebben van p grootheden. Als voorbeeld geven we hier de bekende historische dataset van Bumpus uit 1898 met metingen van lichaamskarakteristieken van een aantal (volwassen) mussen. De tabel hieronder bevat een deel van deze gegevens. Van iedere mus zijn vijf lichaamskarakteristieken gegeven.

nummer	tot. lengte	spanwijdte	kop+bek	humerus	sternum
1	155	240	31.4	18.0	20.7
2	156	240	31.5	18.2	20.6
3	160	242	32.6	18.8	21.7
4	152	232	30.3	17.2	19.8
5	160	250	31.7	18.8	22.5
6	155	237	31.0	18.5	20.0
7	157	245	32.2	19.5	21.4
8	165	245	33.1	19.8	22.7
9	153	231	30.1	17.3	19.8
10	162	239	30.3	18.0	23.1
11	162	243	31.6	18.8	21.3
12	159	245	31.8	18.5	21.7
13	159	247	30.8	18.1	19.0
14	155	243	30.9	18.5	21.3
15	162	252	31.9	19.1	22.2
16	152	230	30.4	17.3	18.6
17	159	242	30.8	18.2	20.5
18	155	238	31.2	17.9	19.3
19	163	249	33.4	19.5	22.8
20	163	242	31.0	18.1	10.7
21	156	237	31.7	18.2	20.3
22	159	238	31.5	18.4	20.3
23	161	245	32.1	19.1	20.8
24	155	235	30.7	17.7	19.6
25	162	247	31.9	19.1	20.4
26	153	237	30.6	18.6	20.4
27	162	245	32.5	18.5	21.1
28	164	248	32.3	18.8	20.9

Tabel 2.6

We noteren deze waarnemingen als een nxp-matrix; de datamatrix X (voor de Bumpus data een 28x5 matrix) met componenten x_{ij} ; $i=1,\dots,n$ en $j=1,\dots,p$. Dus de rij (x_{i1}, \dots, x_{ip}) bevat de p componenten van de i-de meting (voor $i=4$ zijn dit de gegevens van de 4^{de} mus) en de kolom (x_{1j}, \dots, x_{nj}) bevat de n (onafhankelijke) metingen van de j-de component (voor $j=4$ zijn dit al de humerusgegevens). Meestal zal $n > p$.

We berekenen het gemiddelde van elke kolom (dus voor elke variabele):

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

De empirische covariantiematrix definiëren we als de pxp-matrix S, waarvan het (j,k)-element de (empirische) covariantie is tussen de j-de en k-de kolom van de datamatrix X:

$$S_{jk} = \text{cov}(x_j, x_k) = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

We zien dat S een symmetrische matrix is ($S_{jk} = S_{kj}$) en dat het j-de diagonalelement S_{jj} van S de empirische variantie van de j-de kolom van X bevat:

$$S_{jj} = s_j^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

De empirische correlatiematrix R krijgen we door de elementen van S te herschalen met de standaarddeviaties van de j-de en k-de kolom van X:

$$R_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}}$$

	gemiddelen	Cov. mat					Cor. mat				
tot.lengte	158.4	15.07	17.19	2.24	1.75	1.24	1.000	0.776	0.674	0.682	0.143
spanwijdte	241.6	17.19	32.55	3.38	2.95	3.91	0.776	1.000	0.692	0.785	0.306
kop+bek	31.5	2.24	3.38	0.73	0.47	0.74	0.674	0.692	1.000	0.834	0.388
humerus	18.4	1.75	2.95	0.47	0.43	0.63	0.682	0.785	0.834	1.000	0.431
sternum	20.9	1.24	3.91	0.74	0.63	5.00	0.143	0.306	0.388	0.431	1.000

Tabel 2.7: Gemiddelden, covariantiematrix en correlatiematrix van Bumpus data

In het voorbeeld van Bumpus zien we dat de eerste vier variabelen sterk met elkaar gecorreleerd zijn en dat de correlatie van de afmetingen van het sternum met de andere afmetingen vrij klein is.

In de tabel merken we op dat alle getallen in de covariantiematrix boven de diagonaal onder de diagonaal terugkomen vanwege de symmetrie. Bovendien bestaat de diagonaal van de correlatiematrix enkel uit enen. In publicaties wordt daarom de ruimte in de covariantiematrix onder de diagonaal vaak gebruikt om de niet-triviale elementen van de correlatiematrix neer te schrijven:

	gemiddelen	Covariantie - Correlatie Matrix					
tot.lengte	158.4	15.07	17.19	2.24	1.75	1.24	
spanwijdte	241.6	0.776	32.55	3.38	2.95	3.91	
Kop+bek	31.5	0.674	0.692	0.73	0.47	0.74	
Humerus	18.4	0.682	0.785	0.834	0.43	0.63	
Sternum	20.9	0.143	0.306	0.388	0.431	5.00	

Tabel 2.8: Gemiddelden, covariantiematrix en correlatiematrix (cursief) van Bumpus data

Hoofdstuk 3: Kansrekening

“Geen enkele winnaar gelooft in toeval” Nietzsche

In dit hoofdstuk maken we kennis met enkele elementaire begrippen uit de kanstheorie. De meesten onder ons hebben reeds een intuïtief idee van het begrip “kans”: de weerman zegt dat de kans dat het morgen regent 25% is; de journalist zegt dat het Belgische elftal 40% kans heeft om tegen de Hollanders te winnen. Het is belangrijk dat je inzicht krijgt in de denkwereld van de kanstheorie en in de manier waarop toevalsfenomenen in modellen vertaald worden.

3.1 Universum en gebeurtenis

Bij het uitvoeren van experimenten of bij observatiestudies bemerken we dat in zeer veel gevallen de waarnemingsgegevens tot ons komen op een manier die vooraf niet met zekerheid te voorspellen is. Bij het herhaaldelijk uitvoeren van eenzelfde experiment zullen we constateren dat we verschillende uitkomsten observeren. Denk maar aan

- het gooien van een dobbelsteen
- het bepalen van een concentratie
- het tellen van het aantal weekendongevallen
- het bepalen van de dagelijkse hoeveelheid neerslag.

Alhoewel we de uitkomst van een experiment niet op voorhand kennen, is het meestal wel zo dat we ons een idee kunnen vormen van alle mogelijke uitkomsten, die kunnen bekomen worden bij de uitvoering van het experiment.

Definitie

De verzameling van alle mogelijke uitkomsten van een experiment noemen we het **universum** en we noteren het door Ω (hoofdletter omega).

We merken op dat Ω een verzameling uitkomsten is die, afhankelijk van het experiment, aftelbaar (eindig bij de dobbelsteen, oneindig bij de weekendongevallen) of niet aftelbaar (zoals bij het bepalen van een concentratie en de hoeveelheid neerslag) is.

Definitie

In de kanstheorie worden deelverzamelingen van het universum **gebeurtenissen** genoemd. Indien de deelverzameling slechts één element bevat, noemen we ze een **elementaire gebeurtenis**.

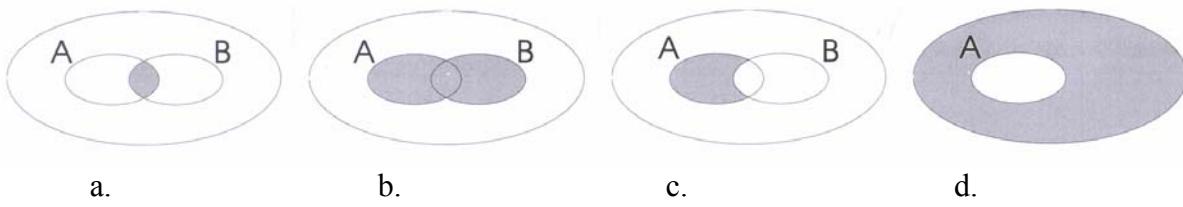
Aangezien het universum een verzameling is, kunnen we begrippen uit de verzamelingenleer, zoals Venndiagrammen, gebruiken om uitkomsten te beschrijven. Als voorbeeld bekijken we het experiment met de dobbelsteen. Er geldt $\Omega = \{1,2,3,4,5,6\}$.

De eigenschap: “gooi een even getal” correspondeert met de deelverzameling of gebeurtenis $A = \{2,4,6\} \subset \Omega$, en de eigenschap: “gooi een getal groter dan vier” met de gebeurtenis $B = \{5,6\} \subset \Omega$.

Eenvoudige rekenregels zijn als volgt:

- a. “Gooi een even getal dat tegelijk ook groter is dan 4” \rightarrow doorsnede = $A \cap B$
- b. “Gooi een getal dat even is of groter dan 4” \rightarrow unie = $A \cup B$
- c. “Gooi een getal dat even is en niet groter dan 4” \rightarrow verschil = $A \setminus B$

d. "Gooi geen even getal" \rightarrow complement van $A = A^c = \Omega \setminus A$



Figuur 3.1

Definitie

Twee gebeurtenissen A en B zijn **uitsluitend** of **disjunkt**, wanneer geen enkele uitkomst van het experiment tegelijkertijd tot A en B behoort. Dit betekent dat de doorsnede van de 2 verzamelingen A en B leeg is.

Voorbeeld: A = gooien een even getal; B = gooien een 3

3.2 Het begrip kans

Als we bij een experiment gebeurtenissen definiëren dan bemerken we dat bepaalde gebeurtenissen blijkbaar veel meer (of veel gemakkelijker) optreden dan andere. Bij het voorbeeld van de dobbelsteen zal men automatisch denken dat de gebeurtenis $A = \{2,4,6\}$ gemakkelijker te realiseren is dan gebeurtenis $B = \{5,6\}$. Een maat voor het begrip "gemakkelijk te realiseren" verkrijgen we door aan gebeurtenissen een getal te associëren dat we definiëren als de kans van de gebeurtenis. We noteren kans door de hoofdletter P (probability). De kans van de gebeurtenis A wordt dus geschreven als $P(A)$. Intuïtief is $P(A) = \frac{1}{2}$ en $P(B) = 1/3$ of:

$$P(E) = \frac{\text{aantal elementen in } E}{\text{aantal elementen in } \Omega} = \text{aantal gunstige gevallen / aantal mogelijke gevallen.}$$

Dit laatste is enkel juist in het geval van een eerlijke dobbelsteen of indien alle uitkomsten dezelfde waarschijnlijkheid hebben. (Indien we geen eerlijke dobbelsteen hebben zullen we de teerling "oneindig" maal moeten gooien en kijken naar de frequenties van het optreden van de gebeurtenis E . We krijgen dan een empirische kans. N.B.: empirisch = experimenteel bepaald)

We kunnen nu overgaan tot de mathematische definitie van het begrip kans.

Definitie

Een **kans** (waarschijnlijkheid, probabiliteit) is een functie waarbij met elke gebeurtenis A van een universum Ω een reëel getal $P(A)$ geassocieerd wordt. Dit getal moet aan de volgende regels voldoen:

1. $0 \leq P(A) \leq 1$
2. $P(\emptyset) = 0$ en $P(\Omega) = 1$
3. Als $A \cap B = \emptyset$ dan is $P(A \cup B) = P(A) + P(B)$

Definitie

Als $P(A) = 0$ noemt men A **stochastisch onmogelijk** en als $P(A) = 1$ noemt men A **stochastisch zeker**.

Gevolgen

1. Als $A \subset B$ dan is $P(A) \leq P(B)$

$$2. P(A^c) = 1 - P(A)$$

$$3. P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Bewijs dit zelf.

Toepassing: Geef de kansen van de vier gevallen, geschatst in Fig 3.1 (je mag ervan uitgaan dat je een eerlijke dobbelsteen hebt). Kloppen de bovenstaande regels met je intuïtieve idee?

3.3 Voorwaardelijke kans

In een groep van 50 studenten, waarvan er 15 biologie studeren en 35 scheikunde, zijn er 20 meisjes. Men kan zich nu de vraag stellen hoe groot de kans is dat een student uit de reeds vermelde groep van 50 biologie en scheikunde studenten een meisje is als ik al weet dat ze biologie studeert? Kennelijk moet ik mijn telwerk nu beperken tot de deelgroep van 15 biologiestudenten. Om de kans te weten moet ik het aantal vrouwelijke biologiestudenten delen door het totale aantal biologiestudenten; we noteren:

$$P(V|B) = \frac{P(V \cap B)}{P(B)} = \frac{\text{aantal vrouwelijke biologiestudenten}}{\text{totale aantal biologiestudenten}}$$

We noemen dit **de voorwaardelijke kans** op het optreden van gebeurtenis V als gebeurtenis B heeft plaatsgevonden (en $P(B) \neq 0$).

Bovenstaande regel $P(V|B) = \frac{P(V \cap B)}{P(B)}$ wordt ook wel de produktregel genoemd.

Bij een voorwaardelijke kans $P(V|B)$ beperken we de verzameling van gebeurtenissen in feite tot de deelverzameling B. Aangezien weer moet gelden $P(B|B) = 1$ moeten we alle kansen hernormaliseren door te delen door $P(B)$.

Opmerkingen

1. Omgekeerd heeft men: $P(B|V) = \frac{P(V \cap B)}{P(V)}$

2. Uitbreiding van de produktregel:

$$P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|B \cap C)P(B|C)P(C)$$

Voorbeeld: We werpen met twee dobbelstenen; wat is de kans op een even aantal ogen als één van beide dobbelstenen een 1 toont?

Antwoord: $P(\text{één van beide dobbelstenen toont een } 1) = 11/36$

$$P(\text{een steen toont een } 1 \text{ en de som is even}) = P(\{(1,1), (1,3), (3,1), (1,5), (5,1)\}) = 5/36$$

$$\text{Zodat } P(\text{aantal ogen even} | \text{een steen toont } 1) = 5/11$$

Voorbeeld

In een hoed stop ik drie identiek gevormde kaarten, waarvan de eerste aan beide zijden rood is, de tweede aan beide zijden wit en de derde aan een zijde rood en aan de andere wit is. Vervolgens trekken we er een willekeurige kaart uit en leggen deze op tafel. Als de bovenkant rood is, wat is dan de kans dat de onderkant ook rood is?

Antwoord 1: De kans op het trekken van de witte (ww), de wit-rode (wr) of de rode kaart (rr) is $1/3$. De kans dat rood boven ligt is $\frac{1}{2}$. Volgens de formule van de voorwaardelijke kans vinden we: $P(rr \mid \text{rood boven}) = P(\text{rr en rood boven})/P(\text{rood boven}) = 2/3$

Antwoord 2: Een alternatieve manier is de volgende beschouwingswijze: we trekken uit de hoed niet alleen een kaart maar ook een zijde die boven komt te liggen. Als we dus de voor- en achterzijde van iedere kaart nummeren met 1 en 2 moeten we willekeurig trekken uit de volgende verzameling:

Boven	r1	r2	r	w	w1	w2
Onder	r2	r1	w	r	w2	w1

Als er een rode zijde boven ligt, beperken we ons tot de eerste drie elementen en we zien dat er met kans $2/3$ ook rood onder ligt.

Opmerking: Een intuïtief acceptabele maar misleidende redenering is de volgende: omdat rood boven ligt, ligt de rode of de rood-witte kaart op tafel, ieder met kans $\frac{1}{2}$ en dus is de kans dat de achterzijde rood is, slechts $\frac{1}{2}$! Waar zit de fout?

Voorbeeld

Wat is de kans dat twee of meer personen in een groep van N dezelfde verjaardag hebben?

Antwoord: Draai de vraagstelling om en definieer p_n als de kans dat geen twee personen in een groep van n dezelfde verjaardag hebben. Kennelijk geldt $p_1 = 1$; de eerste heeft alle dagen van het jaar tot zijn beschikking voor zijn verjaardag. De tweede heeft alle dagen min één tot zijn beschikking en dus $p_2 = 364 / 365$. Voegen we een derde aan de groep toe, dan heeft deze alle dagen min twee tot zijn beschikking zodat $p_3 = 364/365 * 363/365$. Voegen we aan een groep van n personen, met onderling verschillende verjaardagen ($n < 365$), er één toe, dan zijn er nog $365-n$ dagen onbezet, zodat:

$$p_{n+1} = p_n \frac{365-n}{365} \quad \text{en dus} \quad p_{23} = \frac{364}{365} \frac{363}{365} \cdots \frac{343}{365} = 0.4927$$

De kans dat er in een groep van 23 personen minstens twee dezelfde verjaardag hebben is dus $1 - p_{23} = 0.5073$ en is groter dan een half!

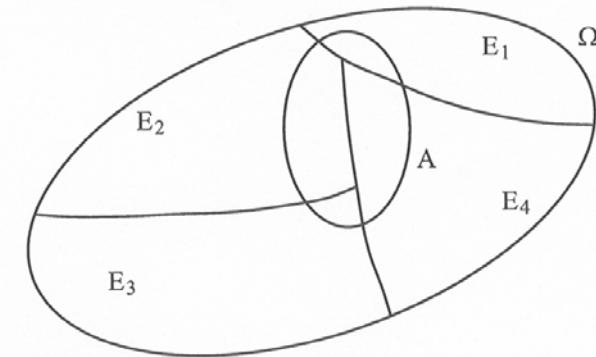
3.4 Totale kans en regel van Bayes

Het begrip voorwaardelijke kans samen met de elementaire rekenregels brengt ons tot twee eigenschappen die in de kanstheorie een speciale naam gekregen hebben.

3.4.1 Wet der totale kans

Onderstel dat E_1, E_2, \dots, E_n een partitie is van het universum Ω (wat wil zeggen dat $E_i \cap E_j = \emptyset$ voor $i \neq j$ en dat $\bigcup_{i=1}^n E_i = \Omega$). Neem nu een willekeurige gebeurtenis A, dan geldt:

$$P(A) = \sum_{i=1}^n P(A|E_i)P(E_i)$$



Figuur 3.2

Bewijs: gebaseerd op $A = A \cap [E_1 \cup E_2 \cup \dots \cup E_n] = [A \cap E_1] \cup \dots \cup [A \cap E_n]$

3.4.2 Regel van Bayes (Thomas Bayes 1763)

Onderstel dat E_1, E_2, \dots, E_n een partitie is van het universum Ω en zij A een willekeurige gebeurtenis met $P(A) > 0$. Dan geldt voor gelijk welke vaste k (met $1 \leq k \leq n$):

$$P(E_k | A) = \frac{P(A|E_k)P(E_k)}{\sum_{i=1}^n P(A|E_i)P(E_i)}$$

Bewijs: m.b.v. de produktregel en de wet der totale kans

Bemerk dat de regel van Bayes de voorwaardelijke kans als het ware omgedraaid uitrekent. Er wordt een vraag gesteld naar de kans van een gebeurtenis E_k , gegeven dat A zich heeft voorgedaan. En het antwoord hierop wordt geformuleerd in termen van kansen van de gebeurtenis A , gegeven dat de E_i 's zich hebben voorgedaan (men moet ook de kansen van de E_i 's zelf kennen). Dat hierbij de E_i 's een partitie van Ω vormen is een essentiële voorwaarde. De regel van Bayes kan je ook interpreteren als een procedure die je toelaat om, vanuit hetgene wat je ziet, een uitspraak te doen over een onderliggende “oorzaak”. Deze regel wordt veelvuldig gebruikt in alle takken van de wetenschap. We geven hier eerst een “hypothetisch” voorbeeld en nadien een voorbeeld uit de medische diagnostiek.

Voorbeeld 1

In een kamer bevinden zich 2 identieke vazen elk gevuld met 10 ballen. Langs de onderzijde is op de voet van de ene vaas de letter X en op de andere de letter Y geschreven. Iemand kiest een vaas, brengt ze naar u en vertelt u dat er in vaas X 6 witte en 4 zwarte ballen zitten en in vaas Y 2 witte en 8 zwarte. Als je op dit ogenblik verplicht bent om te raden welke vaas voor je staat, dan heb je alleen maar als informatie dat $P(X) = \frac{1}{2}$ en $P(Y) = \frac{1}{2}$, waarbij X betekent: “de gekozen vaas is die met letter X”, en analoog voor Y. Onderstel nu dat je bijkomende informatie mag inwinnen door lukraak (bv. geblinddoekt en na goed mengen) een bal uit de vaas te trekken. Nadien kijk je en je bemerkt dat de getrokken bal wit is. De bijkomende informatie zal er voor zorgen dat je nu op een andere manier gaat spreken over wat je niet gezien hebt. Immers als we door W (of Z) de gebeurtenis voorstellen dat de getrokken bal wit (of zwart) is dan hebben we:

$$\begin{aligned} P(W | X) &= 0.6 & P(Z | X) &= 0.4 \\ P(W | Y) &= 0.2 & P(Z | Y) &= 0.8 \end{aligned}$$

De regel van Bayes leert ons dat:

$$P(X|W) = \frac{P(W|X)P(X)}{P(W|X)P(X) + P(W|Y)P(Y)} = \frac{(0.6)(0.5)}{(0.6)(0.5) + (0.2)(0.5)} = 0.75$$

$$P(Y|W) = \frac{P(W|Y)P(Y)}{P(W|X)P(X) + P(W|Y)P(Y)} = \frac{(0.2)(0.5)}{(0.6)(0.5) + (0.2)(0.5)} = 0.25$$

Dus, gebaseerd op de voorwaardelijke kansen, gegeven dat de bal die je trok wit was, is er 75% kans dat de vaas met letter X voor je staat.

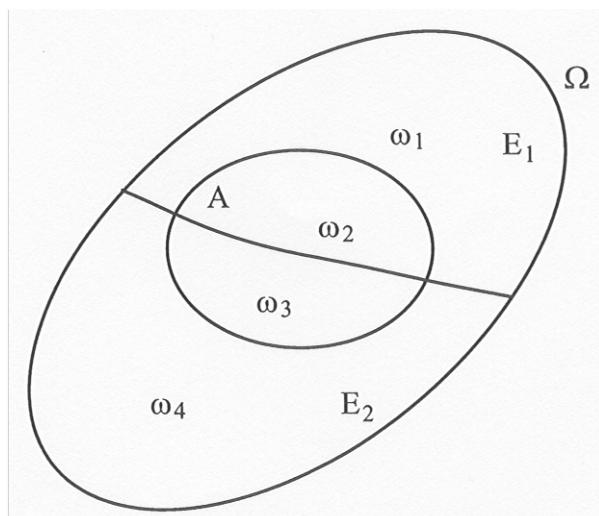
Onderstaande figuur verduidelijkt de situatie:

ω_1 = geselecteerde vaas is X en getrokken bal is zwart

ω_2 = geselecteerde vaas is X en getrokken bal is wit

ω_3 = geselecteerde vaas is Y en getrokken bal is wit

ω_4 = geselecteerde vaas is Y en getrokken bal is zwart



Figuur 3.3

De partitie van het universum $\{\omega_1, \omega_2, \omega_3, \omega_4\}$ is hier E_1 en E_2 met :

$E_1 = \{\omega_1, \omega_2\}$ = geselecteerde vaas is X

$E_2 = \{\omega_3, \omega_4\}$ = geselecteerde vaas is Y

En de "willekeurige" gebeurtenis uit de theorie komt hier overeen met

$A = \{\omega_2, \omega_3\}$ = de getrokken bal is wit

In de Bayesiaanse theorie worden eigen namen gegeven. $P(X)$ en $P(Y)$ heten de "kansen vooraf" (prior probabilities), terwijl de aangepaste kansen, voorwaardelijk op wat we gezien hebben, de naam "kansen achteraf" (posterior probabilities) krijgen. Voorbeelden zijn dus $P(X|W)$ en $P(Y|W)$.

Voorbeeld 2: Diagnostische testprocedures

In de geneeskunde maakt men gebruik van allerlei testen om zo nauwkeurig mogelijk een diagnose te stellen. Om bijvoorbeeld te ontdekken of een patiënt met HIV is besmet (seropositief is), kan men testen op de aanwezigheid van antilichamen in het serum van de patiënt door gebruik te maken van een ELISA techniek (enzyme linked immunosorbent

assay). Wanneer er verkleuring optreedt, wijst dit op de aanwezigheid van antilichamen en dus op besmetting met HIV. De ELISA test is natuurlijk niet 100% volmaakt: bij seropositiviteit zal de test slechts in 99 gevallen op 100 positief zijn. Onderstel nu dat je op HIV laat testen en dat de arts je nadien meldt dat de test positief was. Ben je nu seropositief? Het antwoord op bovenstaande vraag is niet zo simpel en je hebt zelfs nog niet alle informatie gekregen. Om wat structuur in dit experiment te brengen, voeren we volgende notaties in.

- T+ : test is positief,
- T-: test is negatief,
- Z+: je hebt de ziekte,
- Z-: je hebt de ziekte niet.

Er zijn 2 manieren waarop de test foutieve resultaten kan opleveren: een vals positief en een vals negatief:

- kans op een vals positief = $\alpha = P(T+|Z-)$
- kans op een vals negatief = $\beta = P(T-|Z+)$

Beide kansen moeten zo klein mogelijk zijn. Zij hangen enkel van de nauwkeurigheid van de test af en niet van de prevalentie van de ziekte ($P(Z+)$)

Wat nu uiteindelijk zeer belangrijk is, zijn de voorspellende waarden van een test: $P(Z+|T+)$ en $P(Z-|T-)$. Met de regel van Bayes zie je onmiddellijk dat:

$$P(Z+|T+) = \frac{P(T+|Z+)P(Z+)}{P(T+|Z+)P(Z+) + P(T+|Z-)P(Z-)}$$

$$P(Z-|T-) = \frac{P(T-|Z-)P(Z-)}{P(T-|Z+)P(Z+) + P(T-|Z-)P(Z-)}$$

Keren we nu terug naar ons voorbeeld en onderstel dat men schat dat er in België ongeveer 50.000 seropositieven zijn op een bevolking van 10 miljoen inwoners. Dit levert een prevalentie $P(Z+)=0.005$.

Een zeer nauwkeurige test met $\alpha = P(T+|Z-) = 0.02$ en $\beta = P(T-|Z+) = 0.01$ heeft dan de volgende voorspellende waarde:

$$P(Z+|T+) = \frac{(0.99)(0.005)}{(0.99)(0.005) + (0.02)(0.995)} = 0.20$$

$$P(Z-|T+) = 0.80$$

Het is dus allesbehalve zeker dat je seropositief bent als het testresultaat positief is. (Wanneer beleidsinstanties moeten beslissen of er miljoenen worden besteed aan massale screening of preventie, dan zou een elementaire kennis van de regel van Bayes wel eens nuttig kunnen zijn.)

3.5 Onafhankelijkheid

Gooi een witte en een zwarte dobbelsteen en definieer de volgende gebeurtenissen:

- A: het resultaat op de witte dobbelsteen is kleiner dan dat op de zwarte (kans: 15/36)
- B: de witte dobbelsteen valt op een 4 (kans: 6/36)
- C: de som van beide dobbelstenen is zeven (kans: 6/36)
- D: de zwarte dobbelsteen valt op een 3 of een 4 (kans: 12/36)
- E: het getal op de zwarte dobbelsteen is groter dan 2 keer het getal op de witte

(kans: 6/36)

F: we vinden minstens één 6 (kans: 11/36)

De kansen die tussen haakjes vermeld staan vind je gemakkelijk door te tellen. Tevens geldt:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{6/36} = \frac{12}{36}$$

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{3/36}{6/36} = \frac{18}{36}$$

$$P(A|D) = \frac{P(A \cap D)}{P(D)} = \frac{5/36}{12/36} = \frac{15}{36}$$

Herinner je dat $P(A)=15/36$. Het geven van extra informatie resulteert blijkbaar in een nieuwe voorwaardelijke kans van A, die zowel groter dan, gelijk aan of kleiner dan de oorspronkelijke, onvoorwaardelijke kans van A kan zijn. Met welke van de drie gevallen je te maken hebt kan je niet intuïtief te weten komen. De 3 situaties hebben eigen namen:

B draagt negatieve informatie over A want $P(A|B) < P(A)$

C draagt positieve informatie over A want $P(A|C) > P(A)$

D is onafhankelijk van A want $P(A|D) = P(A)$

Bemerkt dat A tevens onafhankelijk is van D.

$$P(D|A) = \frac{P(A \cap D)}{P(A)} = \frac{P(A|D)P(D)}{P(A)} = \frac{P(A)P(D)}{P(A)} = P(D)$$

Onafhankelijkheid is dus een symmetrisch begrip. We spreken daarom ook van 2 onafhankelijke gebeurtenissen. De definitie luidt als volgt:

Definitie

Twee gebeurtenissen A en B heten (stochastisch) **onafhankelijk** als het voor de kans op A niet uitmaakt of B al dan niet gebeurd is. $P(A) = P(A | B) = P(A | B^c)$

Gevolg

A en B onafhankelijk $\Leftrightarrow P(A \cap B) = P(A) P(B)$

We noemen bovenstaande regel de produktregel voor twee onafhankelijke gebeurtenissen. Hij wordt soms als definitie gebruikt.

Let wel dat afhankelijkheid in principe geen oorzakelijk verband impliceert: b.v. de kans dat een willekeurig gekozen getal uit $\{1,2,\dots,100\}$ deelbaar is door 4 is $\frac{1}{4}$; de kans dat het deelbaar is door 10 is $1/10$. De kans dat het deelbaar is door 4 en 10 is $1/20$; er is dus afhankelijkheid, waarom?

Onafhankelijkheid kan ook gedefinieerd worden voor meer dan twee gebeurtenissen.

Definitie

Men zegt dat $\{E_1, E_2, \dots, E_n\}$ een verzameling van n onafhankelijke gebeurtenissen is als en alleen als voor elke deelverzameling $\{E_{i1}, E_{i2}, \dots, E_{im}\}$ geldt:

$$P(E_{i1} \cap E_{i2} \cap \dots \cap E_{im}) = P(E_{i1}) P(E_{i2}) \dots P(E_{im})$$

Dit noemt men de produktregel voor onafhankelijke gebeurtenissen.

Onderstaand voorbeeld illustreert dat gebeurtenissen die twee aan twee onafhankelijk zijn nog niet noodzakelijk een verzameling van onafhankelijke gebeurtenissen vormen.

Voorbeeld

Men gooit gelijktijdig een zwarte en een witte dobbelsteen en definieert de volgende gebeurtenissen:

- A: het getal op de witte dobbelsteen is oneven
- B: het getal op de zwarte dobbelsteen is oneven
- C: de som is oneven

Zodat: $P(A)=1/2$; $P(B)=1/2$; $P(C)=1/2$

We hebben: $P(A \cap B) = 1/4 = P(A)P(B)$ dus A en B onafhankelijk

$P(A \cap C) = 1/4 = P(A)P(C)$ dus A en C onafhankelijk

$P(B \cap C) = 1/4 = P(B)P(C)$ dus B en C onafhankelijk

Nochtans is $\{A, B, C\}$ geen verzameling van onafhankelijke gebeurtenissen want $P(A)P(B)P(C) = 1/8$ en daarentegen $P(A \cap B \cap C) = 0$.

Hoofstuk 4: Stochastische variabelen en hun kansverdeling

4.1. Inleiding en definities

Tot nu toe hebben we experimenten beschreven door hun mogelijke uitkomsten explicet te benoemen en de kans van deze uitkomsten (beschouwd als elementaire gebeurtenissen) aan te geven. Het benoemen van elementaire uitkomsten, of van gebeurtenissen verbonden aan elementaire uitkomsten, moet bij elk verschillend experiment opnieuw gebeuren. Als we een dobbelsteen gooien dan gaat het over het “aantal ogen”, bij een muntstuk passen namen als “kop” en “munt”, enz...

Het is nu de bedoeling om van al deze verschillende namen af te raken en een uniform kader te creëren waarmee we voortaan systematisch kunnen werken. Uiteindelijk gaat het toch altijd over “uitkomsten” en hun “kansen”.

Een manier om naar een uniform kader over te stappen wordt gegeven door uitkomsten te identificeren met reële getallen. We hebben dan altijd **R** (of een deel ervan) als universum en kansen worden dan altijd met reële getallen geassocieerd.

Wiskundig gebeurt de overgang van het universum Ω naar **R** door een reële functie X te definiëren:

$$X : \Omega \rightarrow R : \omega \rightarrow X(\omega)$$

In de kanstheorie wordt zo’n functie X een stochastische veranderlijke (**stochastiek, stochastische variabele, stochast**) genoemd. We zullen deze namen in deze cursus door elkaar gebruiken, zodat jullie met de nomenclatuur vertrouwd raken.

De kansen projecteren gewoon mee: $P(X \leq a) = P(\{\omega \in \Omega \mid X(\omega) \leq a\})$

Meestal interesseren we ons meer voor de getalwaarde $X(\omega)$ dan voor de elementen ω van de onderliggende verzameling. Als ik schoenen wil verkopen in dit land, is de “verdeling” van de voetlengten (en breedten) het enige wat ik van de inwoners wil weten om de goede hoeveelheden van de verschillende maten te kunnen inkopen; ik wil dus iets weten over de getallen $X(\omega)$ voor iedere inwoner $\omega \in \Omega$.

Definitie

Als X een stochastische variabele is, dan heet de functie F_X :

$$F_X(a) = P(X \leq a)$$

de **verdelingsfunctie** van X (ook wel eens **cumulatieve delingsfunctie** genoemd).

Voorbeeld: De dobbelsteen

$$\Omega = \left\{ \begin{array}{c} \text{square} \\ \bullet \end{array}, \begin{array}{c} \text{square} \\ \bullet \\ \bullet \end{array}, \begin{array}{c} \text{square} \\ \bullet \\ \bullet \\ \bullet \end{array}, \begin{array}{c} \text{square} \\ \bullet \\ \bullet \\ \bullet \\ \bullet \end{array}, \begin{array}{c} \text{square} \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \end{array}, \begin{array}{c} \text{square} \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \end{array} \right\}$$

$$X \left(\begin{array}{c} \text{square} \\ \bullet \end{array} \right) = 1, X \left(\begin{array}{c} \bullet \\ \text{square} \\ \bullet \end{array} \right) = 2, \text{ etc...}$$

$$P \left(\begin{array}{c} \text{square} \\ \bullet \end{array} \right) = \frac{1}{6}, \text{ etc...}$$

Voor de kansen $P(X \leq a)$ vinden we:

$$P(X \leq 0) = 0 = P(X < 1)$$

$$P(X \leq 1) = 1/6 = P(X < 2)$$

$$P(X \leq 2) = 2/6 = P(X < 3) \text{ etc.}$$

We vinden een verdelingsfunctie zoals geschetst in Figuur 4.1. Dit is een trapfunctie die in de punten 1, 2, 3, 4, 5 en 6 een sprong van $1/6$ maakt. Deze verdeling is duidelijk “discreet”.

Voorbeeld: Lichaamslengtes

De verdeling van lichaamslengten van volwassen mannelijke inwoners van België is geschetst in Figuur 4.1. Neem een willekeurige volwassen mannelijke inwoner en lees de kans af dat deze kleiner is dan 190cm. In theorie is ook deze verdeling discreet, maar de groep mannen (en dus Ω) is zo groot dat we doen alsof deze “continu” is.

Figuur 4.1: de verdelingsfunctie van de dobbelsteen en van de lichaamslengte van de mannen

Alvorens we een formele definitie van de begrippen discrete en continue verdeling geven zullen we enkele eigenschappen van de verdelingsfunctie formuleren.

Eigenschappen

- i) $0 \leq F_X(a) \leq 1$
- ii) Vermits $a \leq b \rightarrow F_X(a) \leq F_X(b)$ geldt dat F_X een monotone functie is.
- iii) $P(a < X \leq b) = F_X(b) - F_X(a)$ en $P(X > a) = 1 - F_X(a)$

iv) F_X is rechtscontinu

$$v) \lim_{x \rightarrow \infty} F_X(x) = 1 \quad \text{en} \quad \lim_{x \rightarrow -\infty} F_X(x) = 0$$

vi) Als we een nieuwe stochastische variabele Y definiëren als een lineaire transformatie van een stochastische variabele X , namelijk $Y=aX+b$, dan transformeert de verdelingsfunctie als volgt:

$$\begin{aligned} a > 0: \quad F_Y(y) &= P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right) \\ a < 0: \quad F_Y(y) &= P(Y \leq y) = P(aX + b \leq y) = P\left(X \geq \frac{y-b}{a}\right) = 1 - P\left(X < \frac{y-b}{a}\right) \\ &= 1 - F_X\left(\frac{y-b}{a}\right) + P\left(X = \frac{y-b}{a}\right) \end{aligned}$$

Let op: Meestal is $P(X \leq a)$ niet gelijk aan $P(X < a)$.

4.2 Continue en discrete verdelingen

In de twee voorgaande voorbeelden hebben we duidelijk een verschil gezien: in het eerste voorbeeld kan de stochastische variabele X slechts een eindig aantal waarden aannemen; in het tweede voorbeeld daarentegen zijn er een niet aftelbaar aantal mogelijke waarden.

4.2.1. De discrete stochastiek

Definitie

We noemen een stochastiek X **discreet** als X slechts een eindig of aftelbaar aantal verschillende waarden kan aannemen. Dat wil zeggen dat er een verzameling (reële) getallen $\{x_i \mid i=1,2,\dots\}$ is, zodat

$$P(X = x_i) = p_i \quad \text{en} \quad \sum_{i=1}^{\infty} p_i = 1$$

Voorbeeld:

In een vaas bevinden zich 3 rode knikkers en 7 witte.

Er wordt 4 maal een knikker uit de vaas getrokken en het resultaat wordt genoteerd; na het trekken wordt de knikker weer in de vaas gedaan (*trekking met teruglegging*).

De stochast X wordt nu gedefinieerd als het totaal aantal getrokken rode knikkers. We schrijven wel: $X = \text{aantal rode knikkers}$. De verzameling van de mogelijke waarden van X is $\{0, 1, 2, 3, 4\}$. De bijbehorende kansen kunnen als volgt berekend worden:

$p_0 = P(X = 0) = 1 \times 7/10 \times 7/10 \times 7/10 \times 7/10 = 1 \times (0,3)^0 (0,7)^4 = 0,2401$	uitkomst: WWWWW
$p_1 = P(X = 1) = 4 \times 3/10 \times 7/10 \times 7/10 \times 7/10 = 4 \times (0,3)^1 (0,7)^3 = 0,4116$	uitkomsten: RWWW or WRWW or WWRW or WWWR
$p_2 = P(X = 2) = 6 \times 3/10 \times 3/10 \times 7/10 \times 7/10 = 6 \times (0,3)^2 (0,7)^2 = 0,2646$	uitkomsten: RRWW or RWRW or RWWR or WRRW of ...
$p_3 = P(X = 3) = 4 \times 3/10 \times 3/10 \times 3/10 \times 7/10 = 4 \times (0,3)^3 (0,7)^1 = 0,0756$	uitkomsten: RRRW or RRWR or RWRR of WRRR
$p_4 = P(X = 4) = 1 \times 3/10 \times 3/10 \times 3/10 \times 3/10 = 1 \times (0,3)^4 (0,7)^0 = 0,0081$	uitkomst: RRRR

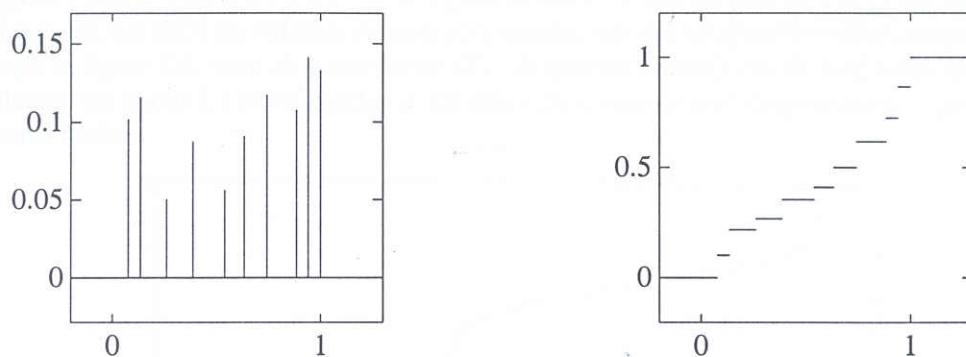
Samenvattend:

x_i	0	1	2	3	4
$P(X = x_i)$	0,2401	0,4116	0,2646	0,0756	0,0081

Merk hierbij op dat de som van de kansen gelijk is aan 1, zoals je verwacht.

Grafische weergave van een discrete verdeling

We kunnen de kansen p_i ook grafisch weergeven door een staafdiagram: op het punt x_i richten we een staafje op van lengte p_i . De verdelingsfunctie F_X is stuksgewijs constant met sprongen in de punten x_i ($i=0,1,2,\dots$) van grootte p_i . Hieronder een staafdiagram met rechts de bijbehorende verdelingsfunctie F_X . Maak zelf het staafdiagram en de verdeling voor het voorgaande voorbeeld.



Figuur 4.2: Een staafdiagram en de bijbehorende verdelingsfunctie

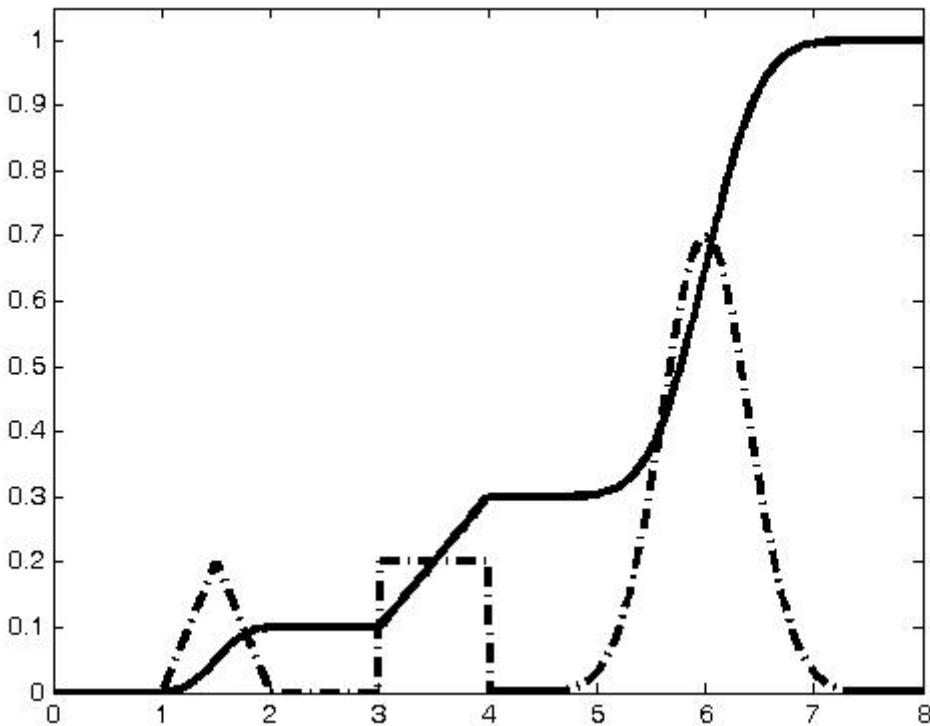
4.2.2. De continue stochastiek

Definitie

We noemen een stochastische variabele X **continu** als de verdelingsfunctie F_X overal continu is en bovendien overal differentieerbaar (behalve eventueel in een eindig aantal punten). Dit is een vrij zware eis, maar zij maakt het ons wel mogelijk om de **kansdichtheid (of dichtheidsfunctie)** f_X te definiëren als de afgeleide van F_X :

$$f_X = \frac{d}{dx} F_X(x) \quad \text{en dus ook} \quad F_X(x) = \int_{-\infty}^x f_X(t) dt$$

Voorbeeld van een continue verdeling



Figuur 4.3: Een hypothetische dichtheidsfunctie f (-.-) met bijbehorende verdelingsfunctie F (volle lijn)

Gevolg

1. Omdat F_X monotoon is, moet gelden dat $f_X(x) \geq 0 \quad \forall x$. Tevens moet de oppervlakte onder de staarten van f_X naar nul gaan.

$$2. P(a < X \leq b) = \int_a^b f(t) dt$$

Dus moet de oppervlakte onder f_X gelijk zijn aan 1 of $\int_{-\infty}^{+\infty} f(t) dt = 1$

3. Zoals je op figuur 4.3 ziet kan f_X een vrij wild gedrag hebben, maar door de eis van differentieerbaarheid (en dus continuïteit) van F_X voor continue verdelingen, sluiten we de sprongen in F_X expliciet uit, zodat $P(X \leq a) = P(X < a)$ voor alle beschouwde continue verdelingen.

N.B. In het discrete geval correspondeert het staafdiagram met de dichtheidsfunctie.

4.2.3. Functies van stochastieken en hun verdelingsfunctie

Als g een reële continue functie is en X een stochastiek (continu of discreet), dan is $g(X)$ de stochastiek met de verdelingsfunctie:

$$F_{g(X)}(a) = P(g(X) \leq a) = P(\{\omega \in \Omega | g(X(\omega)) \leq a\})$$

Ga zelf na wat de kansdichtheid van $g(X)$ is als X continu en g monotoon stijgend en differentieerbaar is.

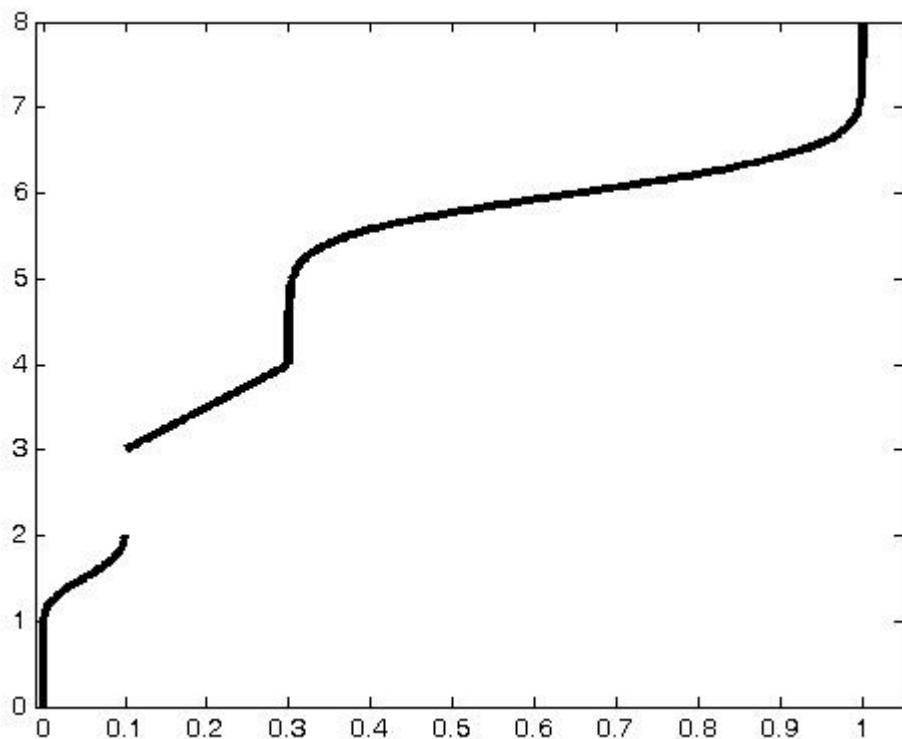
4.3 Percentielen

In de praktijk willen we voor een stochastiek X vaak een antwoord op de omgekeerde vraag: “voor welke waarde van x is 25% (of 50% of 90%) van de uitkomsten kleiner dan of gelijk aan x ?”. We gaan hier analoog te werk als in hoofdstuk 2, maar nu met de echte verdelingsfunctie i.p.v. de empirische.

De algemene vraag luidt dus: “gegeven een percentage α ($0 \leq \alpha \leq 100$) of een kans p ($=\alpha/100$), voor welke x geldt dat $F_X(x) = P(X \leq x) = p$?”. We zijn dus geïnteresseerd in de inverse functie van F_X .

Grafisch betekent dit dat we abcis en ordinaat (X- en Y-as) verwisselen, oftewel de figuur van F_X moeten spiegelen om de bissectrice van het eerste kwadrant ($y=x$). We hebben dit in figuur 4.4 gedaan voor de continue verdeling van figuur 4.3. De intervallen waar de verdeling constant is, geven een sprong in de inverse functie (dus in $x=0.1$ wordt een sprong van 2 naar 3 gemaakt). Hoe gaan we nu te werk om de percentielen af te lezen ?

- Als in een punt x met $p:=F_X(x)$ geldt $p>F_X(y)$ als $x>y$ en $p<F_X(y)$ als $x<y$, d.w.z. dat F_X strikt stijgend is in x , dan is x het enige punt met $F_X(x)=p$ en dan definiëren we x als het $100p\%$ -percentiel van X (dit is dus het p -de quantiel). Dit percentiel is dus gewoon de waarde van de inverse functie $F_X^{-1}(p)$ (Figuur 4.4: het 20%-percentiel is 3.5).
- Als er geen punt x is waarvoor $F_X(x)=p$, dan is F_X discontinu en maakt deze ergens een sprong van een waarde kleiner dan p naar een waarde groter dan p . Het $100p\%$ -percentiel van X is dan het punt waarin F_X deze sprong maakt (Figuur 4.4 heeft zo geen punt want F is continu).



Figuur 4.4: Kwantielen van de verdeling van figuur 4.3

- Als $F_X(x)=p$ constant is voor alle x in een interval $[a,b]$, dan zouden we ieder element van dat interval het p -de percentiel van X kunnen noemen. Voor een eenduidige definitie maken we dan de volgende afspraak:

- i) Als $F_X(x)=0$ voor alle $x \leq a$ en $F_X(x)>0$ voor alle $x > a$, dan heet a het 0%-percentiel van X ; a is dus het punt vanaf waar F_X niet triviaal is (Figuur 4.4: het 0%-percentiel is 1).
- ii) Als $F_X(x)=1$ voor alle $x \geq b$ en $F_X(x)<1$ voor alle $x < b$, dan heet b het 100%-percentiel van X ; b is dus het punt vanaf waar F_X weer triviaal is (Figuur 4.4: het 100%-percentiel is 7.1).
- iii) Als $F_X(x)=p$ voor alle x in het interval $[c,d]$, $F_X(x)<p$ voor alle $x < c$ en $F_X(x)>p$ voor alle $x > d$, dan kiezen we het midden $\frac{1}{2}(c+d)$ als het $100p\%$ -percentiel van X (Figuur 4.4: het 10%-percentiel is 2.5).

Deze definitie lijkt niet consistent met die van empirische percentielen uit hoofdstuk 2. Dit is echter maar schijn, omdat de empirische percentielen gebaseerd zijn op slechts eindig veel waarnemingen. Op grond van de wet van de grote getallen (zie later) kunnen we het volgende laten zien: als $\{x_1, x_2, \dots, x_n\}$ onafhankelijke waarnemingen zijn van een stochastiek X , dan convergeert de empirische verdelingsfunctie van deze waarnemingen naar F_X voor $n \rightarrow \infty$ en convergeren de empirische percentielen naar de hierboven gedefinieerde percentielen van X .

4.4 Meerdimensionale stochastieken

4.4.1 Discrete en continue stochastieken

In vele gevallen kan men aan een element van een steekproefruimte meer dan één reële getalwaarde toekennen.

Bijvoorbeeld, bij een onderzoek van de inwoners van België is men zowel geïnteresseerd in de lengte als het gewicht van elke inwoner. Bij de analyse van samples is men geïnteresseerd in de samenstelling: de concentratie van molecule 1, 2 en 3.

Als we n eigenschappen gelijktijdig beschouwen, hebben we een vectorfunctie $Z: \Omega \rightarrow \mathbf{R}^n$.

Definitie

Indien de componenten X_1, X_2, \dots, X_n van een vectorfunctie $Z: \Omega \rightarrow \mathbf{R}^n$ stochastische variabelen zijn, dan noemen we Z een **n -dimensionale stochastische variabele of kansvector**.

Voor de eenvoud zullen we ons in wat volgt beperken tot het geval $n=2$.

Definitie

De **verdelingsfunctie** $F_Z: \mathbf{R}^2 \rightarrow [0,1]$ van een tweedimensionale kansvector $Z = (X, Y)$ wordt als volgt gedefinieerd:

$$F_Z(a,b) = P(X \leq a \text{ en } Y \leq b) = P(\{\omega \in \Omega : X(\omega) \leq a \text{ en } Y(\omega) \leq b\})$$

Net zoals in het voorgaande hoofdstuk zullen we onderscheid maken tussen **continue** en **discrete** kansvectoren.

We noemen **Z discreet**, indien er een eindig of aftelbaar aantal punten $z_1=(x_1,y_1), z_2=(x_2,y_2), z_3=(x_3,y_3), \dots$ bestaan, zodat $P(Z=z_i) = p_i$, met $p_i \in [0,1]$ en $\sum_i p_i = 1$. $P(Z=z)=0$ voor alle andere punten van \mathbf{R}^2 .

Definitie: De dichtheidsfunctie voor continue kansvectoren

Voor **continue** kansvectoren definiëren we de dichtheidsfunctie f_Z als de volgende 2-de orde partiële afgeleide:

$$f_Z = \frac{\partial^2 F_Z}{\partial x \partial y}$$

(We zullen in deze syllabus steeds de extra eis opleggen dat alle tweede (n-de in n dimensies) orde gemengde partiële afgeleiden van de verdelingsfunctie continu moeten zijn)

Gevolg

Indien f_Z gekend is, kunnen we de verdelingsfunctie F_Z terugvinden door integratie:

$$F_Z(x, y) = \int_{-\infty}^x \left(\int_{-\infty}^y f_Z(u, v) dv \right) du$$

Voor iedere (meetbare) deelverzameling A van \mathbb{R}^2 kunnen we de kans bepalen, dat $Z \in A$:

$$P(Z \in A) = \iint_A f_Z(x, y) dx dy$$

Definitie: Marginale verdeling

Veronderstel dat $\mathbf{Z} = (X, Y)$ een kansvector is, en dat de verdelingsfunctie F_Z bekend is. Dan kunnen we hieruit de verdelingsfunctie voor de stochastische variabelen X en Y berekenen:

$$F_X(a) = P(X \leq a) = P(X \leq a \text{ en } Y < +\infty) = \lim_{y \rightarrow +\infty} F_Z(a, y)$$

De verkregen verdeling heet de **marginale kansverdeling** van X. Analoog vinden we de marginale kansverdeling voor Y (doe zelf).

Gevolg

Voor een continue verdeling kunnen we de dichtheidsfuncties van de marginale verdeling gemakkelijk terugvinden:

$$f_X(x) = \lim_{y \rightarrow +\infty} F_Z(x, y) = \int_{-\infty}^x \left(\int_{-\infty}^{+\infty} f_Z(u, v) dv \right) du$$

en dus is de marginale kansdichtheid

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{+\infty} f_Z(x, v) dv$$

Voorbeeld 1

We werpen met twee dobbelstenen (of we werpen 2 maal met dezelfde dobbelsteen) en beschouwen de volgende stochastische variabelen:

X: het aantal enen dat gegooid wordt;

Y: het aantal zessen dat gegooid wordt.

$\mathbf{Z} = (X, Y)$ is dan een kansvector, die enkel paren gehele waarden (i, j) met $0 \leq i+j \leq 2$ kan aannemen. Verifieer dat

$$P\{(0,0)\} = \frac{4^2}{6^2} = \frac{16}{36}$$

$$P\{(1,0)\} = P\{(0,1)\} = 2 \cdot \frac{1}{6} \cdot \frac{4}{6} = \frac{8}{36}$$

$$P\{(1,1)\} = \frac{2}{36}$$

$$P\{(2,0)\} = P\{(0,2)\} = \frac{1}{36}$$

De grafiek van de verdelingsfunctie wordt gegeven in onderstaande figuur.

Figuur 4.5:Driedimensionale tekening en hoogtelijnen van de kansverdeling

Voor het maken van deze figuur bedenk dat voor:

- i) $x < 0$ en/of $y < 0 \rightarrow f_Z(x,y)=0$
- ii) $0 \leq x, y < 1 \rightarrow f_Z(x,y)=P\{(0,0)\}=16/36$
- iii) $0 \leq x < 1$ en $1 \leq y < 2 \rightarrow f_Z(x,y)=P\{(0,0),(0,1)\}=24/36$
- iv) $0 \leq x < 1$ en $2 \leq y \rightarrow f_Z(x,y)=P\{(0,0),(0,1),(0,2)\}=25/36$
- v) $1 \leq x < 2$ en $1 \leq y < 2 \rightarrow f_Z(x,y)=P\{(0,0),(0,1),(1,0),(1,1)\}=34/36$
- vi) $1 \leq x < 2$ en $2 \leq y \rightarrow f_Z(x,y)=P\{(0,0),(0,1),(1,0),(1,1),(0,2)\}=35/36$
- vii) $2 \leq x, y \rightarrow f_Z(x,y)=P\{(0,0), (0,1), (1,0), (1,1), (0,2), (2,0)\}=1$

Tevens is $f_Z(x,y)=f_Z(y,x)$

Voorbeeld 2

We hernemen voorbeeld 1, maar we werpen nu drie dobbelstenen in plaats van twee.

$Z = (X,Y)$ is nu een kansvector, die enkel paren gehele waarden (i,j) met $0 \leq i+j \leq 3$ kan aannemen. Verifieer dat

$$P\{(0,0)\} = \frac{4^3}{6^3}$$

$$P\{(1,0)\} = P\{(0,1)\} = \binom{3}{1} \frac{1}{6} \frac{4}{6} \frac{4}{6} = \frac{48}{6^3}$$

$$P\{(1,1)\} = 2 \binom{3}{2} \frac{1}{6} \frac{1}{6} \frac{4}{6} = \frac{24}{6^3}$$

$$P\{(2,0)\} = P\{(0,2)\} = 3 \binom{3}{1} \frac{1}{6} \frac{1}{6} \frac{4}{6} = \frac{12}{6^3}$$

$$P\{(2,1)\} = P\{(1,2)\} = 3 \frac{1}{6} \frac{1}{6} \frac{1}{6} = \frac{3}{6^3}$$

$$P\{(3,0)\} = P\{(0,3)\} = \frac{1}{6^3}$$

Aanwijzing: het aantal mogelijke gevallen is steeds 6^3 ; zoek met behulp van combinatieleer steeds het aantal gunstige gevallen.

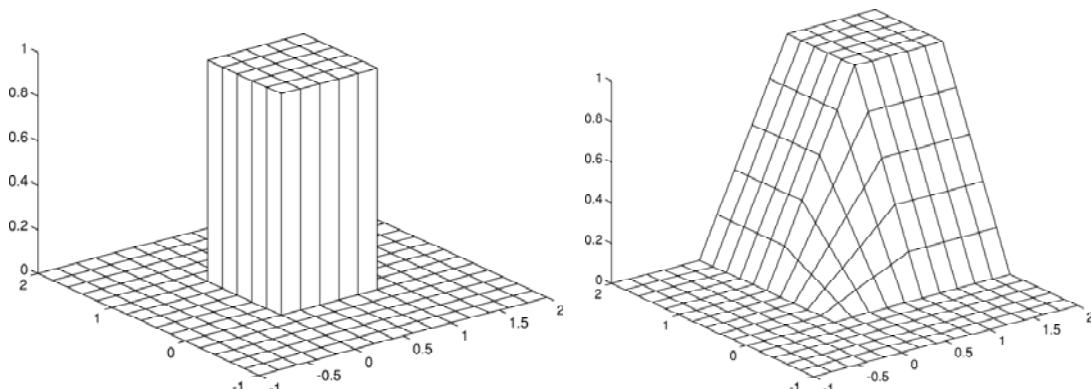
Voorbeeld 3

Men kiest willekeurig en onafhankelijk van elkaar twee getallen tussen 0 en 1. Laat X het eerste getal, Y het tweede en $Z = (X, Y)$ zijn. Dan is het duidelijk dat:

$$f_Z(x, y) = \begin{cases} 1 & \text{als } 0 \leq x, y \leq 1 \\ 0 & \text{anders} \end{cases}$$

De verdelingsfunctie wordt gegeven door de formules:

$$F_Z(x, y) = \begin{cases} 0 & \text{als } x \leq 0 \text{ of } y \leq 0 \\ xy & \text{als } 0 \leq x, y \leq 1 \\ x & \text{als } 0 \leq x \leq 1 \text{ en } y \geq 1 \\ y & \text{als } 0 \leq y \leq 1 \text{ en } x \geq 1 \\ 1 & \text{als } x \geq 1 \text{ en } y \geq 1 \end{cases}$$



Figuur 4.6: Dichtheidsfunctie en verdelingsfunctie van de uniforme verdeling

4.5 Onafhankelijke stochastische variabelen

In bovenstaand voorbeeld lieten we het woord “onafhankelijk” vallen. In het vorige hoofdstuk definieerden we onafhankelijkheid van gebeurtenissen. Wat betekent het nu dat twee stochastische variabelen onafhankelijk zijn?

Definitie

Twee stochastische variabelen X en Y heten onafhankelijk als de gebeurtenissen $\{a_1 < X \leq b_1\}$ en $\{a_2 < Y \leq b_2\}$ onafhankelijk zijn voor alle a_i, b_i in \mathbf{R} ($i=1,2$), of, equivalent, als:

$$P(\{a_1 < X \leq b_1\} \cap \{a_2 < Y \leq b_2\}) = P(a_1 < X \leq b_1) P(a_2 < Y \leq b_2)$$

Voorbeeld

De stochastische variabelen uit het derde voorbeeld zijn onafhankelijk, maar die uit het tweede voorbeeld niet! Immers,

$$P(X=1) = \frac{75}{6^3}, P(Y=2) = \frac{15}{6^3} \text{ maar } P(X=1 \text{ en } Y=2) = \frac{3}{6^3} \neq \frac{75 \times 15}{6^6}$$

Gevolg 1

De componenten van een tweedimensionale kansvector $\mathbf{Z} = (X, Y)$ zijn onafhankelijk als en slechts als de verdelingsfunctie van \mathbf{Z} het product is van de marginale verdelingsfuncties:

$$F_Z(x, y) = F_X(x)F_Y(y)$$

Gevolg 2

De componenten van een tweedimensionale continue kansvector $\mathbf{Z} = (X, Y)$ zijn onafhankelijk als en slechts als de dichtheidsfunctie van \mathbf{Z} het product is van de marginale dichtheidsfuncties: $f_Z(x, y) = f_X(x)f_Y(y)$

Voor de fijnproevers: De som van twee onafhankelijke stochastische variabelen

Bij een halte passeert om de tien minuten een tram. Je neemt elke dag deze tram op een willekeurig tijdstip. De wachttijd T op de eerstvolgende tram bezit dan de volgende dichtheidsfunctie:

$$f_T(t) = \begin{cases} \frac{1}{10} & \text{als } 0 \leq t \leq 10 \\ 0 & \text{anders} \end{cases}$$

We noemen zo'n T uniform verdeeld over $[0, 10]$ (zie verder). Als je nu tweemaal de tram neemt, hoelang moet je dan in totaal wachten; m.a.w. wat is de dichtheidsfunctie $f_{T_1+T_2}$ van de som T_1+T_2 , als T_1 en T_2 de eerste resp. tweede wachttijd aan de halte zijn.

Dit probleem is een speciaal geval van het volgende: veronderstel dat X en Y twee onafhankelijke continue stochastische variabelen zijn, met dichtheidsfuncties f_X en f_Y . Hoe vinden we f_{X+Y} ? Dit gebeurt als volgt: we bepalen eerst de verdelingsfunctie F_{X+Y} van de som

$$F_{X+Y}(x) = P(X + Y \leq x) = \iint_{u+v \leq x} f_Z(u, v) du dv = \int_{-\infty}^{+\infty} \int_{-\infty}^{x-u} f_Z(u, v) dv du = \int_{-\infty}^{+\infty} f_X(u) \int_{-\infty}^{x-u} f_Y(v) dv du$$

(maak een tekening van het integratiegebied).

Afleiden naar x geeft (in de veronderstelling dat we differentiatie en integratie mogen verwisselen):

$$f_{X+Y}(x) = \frac{d}{dx} \left(\int_{-\infty}^{+\infty} f_X(u) \int_{-\infty}^{x-u} f_Y(v) dv du \right) = \int_{-\infty}^{+\infty} f_X(u) f_Y(x-u) du$$

We verkrijgen wat men in de wiskunde het convolutieprodukt van de dichtheseden van X en Y noemt en noteert als $(f_X * f_Y)(x)$

Besluit:

Als X en Y twee onafhankelijke stochastische variabelen zijn met continue verdeling, dan is de dichtheidsfunctie van $X+Y$ de convolutie van de dichtheseden van X en Y :

$$f_{X+Y}(x) = (f_X * f_Y)(x)$$

We keren nu terug naar de toepassing hierboven. Met behulp van het bovenstaande kunnen we de dichtheidsfunctie van T_1+T_2 bepalen:

$$f_{T_1+T_2}(t) = \int_{-\infty}^{+\infty} f_{T_1}(u)f_{T_2}(t-u)du = \frac{1}{10} \int_0^{10} f_{T_2}(t-u)du$$

We onderscheiden nu vier gevallen:

i) $t < 0$. Voor $0 \leq u \leq 10$ geldt dan dat $f_{T_2}(t-u) = 0$, zodat: $f_{T_1+T_2}(t) = 0$

Dit is uiteraard wat we verwachten: een negatieve wachttijd kan nooit optreden.

ii) $0 \leq t \leq 10$. Dan is $t-u \leq 10$. Voor u gelegen tussen 0 en t hebben we bovendien dat $0 \leq t-u$,

$$\text{zodat: } f_{T_1+T_2}(t) = \frac{1}{10} \int_0^t \frac{1}{10} du = \frac{t}{100}$$

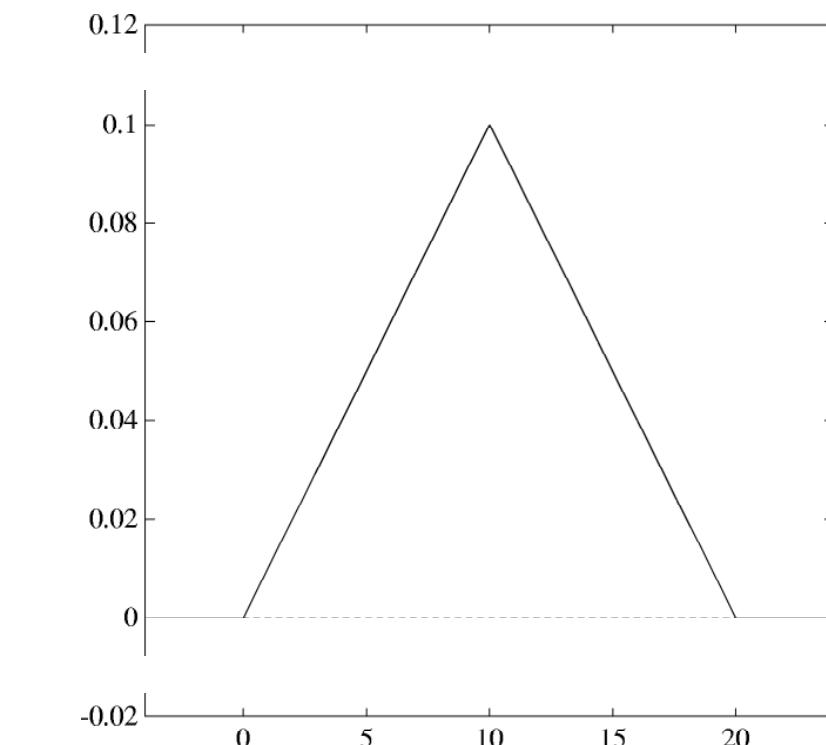
iii) $10 \leq t \leq 20$. Nu is $0 \leq t-u$ en bovendien geldt dat voor u gelegen tussen $t-10$ en 10 dat $t-u \leq 10$, zodat:

$$f_{T_1+T_2}(t) = \frac{1}{10} \int_{t-10}^{10} \frac{1}{10} du = \frac{20-t}{100}$$

iv) $t > 20$. Dan is $t-u > 10$, zodat $f_{T_2}(t-u) = 0$ voor u tussen 0 en 10 zodat, net als in het eerste geval geldt: $f_{T_1+T_2}(t) = 0$

Inderdaad is het onmogelijk dat we een totale wachttijd hebben die langer duurt dan 20 minuten.

De grafiek van $f_{T_1+T_2}(t)$ wordt gegeven in figuur 4.7. Bepaal zelf de verdelingsfunctie $F_{T_1+T_2}$ en teken de grafiek.



Figuur 4.7: Kansdichtheid voor de som van de wachttijden van twee tramritten.

4.6 Verwachtingswaarde, variantie en standaardafwijking

Bij een loterij zijn er 1000 loten van 1 Euro. Het winnende nummer is goed voor 400 Euro en er zijn 5 troostprijzen van 20 Euro. Wat is de waarde die je aan zo'n lot kunt toekennen?

Op voorhand weten we natuurlijk niet op welk lot de hoofdprijs gaat vallen en dus welk lot een grote waarde heeft. We kunnen wel een soort “gemiddelde” waarde van een lot bepalen. Stel, dat we alle loten zouden kopen, dan zijn we 1000 Euro kwijt en we winnen 500 Euro aan prijzen; het verlies is dus gemiddeld 0.50 Euro per lot. Aan ieder lot kunnen we dus een “waarde” toekennen van -0.50 Euro. We noemen dit de verwachtingswaarde van een lot uit de betreffende loterij. Dit voorbeeld suggerert de definitie:

Definitie

Voor een gegeven stochastische variabele X definiëren we de verwachtingswaarde $E[X]$ (Eng.: expectation) door:

$$E[X] = \begin{cases} \sum_j x_j p_j & \text{indien } X \text{ discreet verdeeld} \\ \int_{-\infty}^{+\infty} xf_X(x)dx & \text{indien } X \text{ continu verdeeld} \end{cases}$$

Merk op dat $E[X]$ niet altijd bestaat; het is inderdaad mogelijk dat de reeks of oneigenlijke integraal divergeert.

Voorbeeld

Bij de boven vermelde loterij is Ω de verzameling van de 1000 loten en $X(\omega)$ is de winst die je maakt bij het kopen van één ervan:

$X(\omega)$ is dus 400-1 voor het winnende lot; 20-1 voor de troostprijzen; -1 voor de andere loten.

Bijgevolg is

$$E[X] = (400 - 1)/1000 + (20 - 1)/1000 + (-1) 994/1000 = -0.5$$

We vinden dus een negatieve verwachtingswaarde. Gemiddeld gezien zullen we dus verliezen.

Voorbeeld

Men werpt een dobbelsteen. X is het aantal ogen dat bovenaan komt te liggen. Dan is

$$E[X] = \sum_{j=1}^6 \frac{j}{6} = \frac{21}{6} = 3.5$$

Definitie: Verwachtingswaarde van $g(X)$

Neem nu een (continue) functie $g: \mathbf{R} \rightarrow \mathbf{R}$, dan kunnen we, zoals reeds eerder beschreven, een nieuwe stochastische variabele $g(X)$ definiëren voor een gegeven stochastiek X als de stochastiek met de verdelingsfunctie $F_{g(X)}(z) := P(g(X) \leq z)$.

Voor X discreet vindt men gemakkelijk dat

$$E[g(X)] = \sum_j g(x_j) p_j$$

Voor een continu verdeelde stochastische variabele X heeft men, op analoge manier

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$$

In het meerdimensionale geval definiëren we analoog:

Discreet: $E[g(X, Y)] = \sum_i \sum_j g(x_i, y_j) P(X = x_i, Y = y_j)$

Continu: $E[g(X, Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f_{(X, Y)}(x, y) dx dy$

Gevolg

Met deze formule kunnen we eenvoudig laten zien dat de verwachtingswaarde van de som van twee stochastieken X en Y altijd de som van de verwachtingswaarden is:

$$E[X+Y] = E[X] + E[Y]$$

Bewijs: $E[X+Y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x+y) f_{(X, Y)}(x, y) dx dy$

$$E[X] = \int_{-\infty}^{+\infty} xf_X(x) dx = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf_{(X, Y)}(x, y) dx dy$$

$$E[Y] = \int_{-\infty}^{+\infty} yf_Y(y) dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yf_{(X, Y)}(x, y) dx dy$$

Optellen van de laatste twee vergelijkingen geeft de eerste.

Analoog kun je bewijzen dat: $E[X-Y] = E[X] - E[Y]$

Gevolg

We kunnen dit laatste gemakkelijk uitbreiden tot een som van meer dan twee stochasten:

$$E[X_1 + X_2 + \dots + X_n] = E[X_1] + E[X_2] + \dots + E[X_n]$$

Definitie

De variantie van een stochastische variabele X is de verwachtingswaarde van het kwadraat van de afwijking t.o.v. de verwachtingswaarde E[X]:

$$\text{Var}[X] = E[(X - E[X])^2]$$

Zij bestaat alleen als de bijbehorende integraal of oneindige som niet divergeert, dus indien:

$$\sum_j (x_j - E[X])^2 p_j < \infty \quad \text{of} \quad \int_{-\infty}^{+\infty} (x - E[X])^2 f_X(x) dx < \infty$$

Definitie

De standaardafwijking van X is de vierkantswortel van de variantie:

$$\sigma_X = \sqrt{\text{Var}[X]}$$

De variantie geeft aan hoe X varieert rond zijn verwachtingswaarde. Hoe groter de kans is dat $X(\omega)$ dicht bij $E[X]$ ligt, hoe kleiner $\text{Var}[X]$ is.

Merk ook op dat σ_X en X dezelfde dimensies hebben.

Voorbeeld

In het voorbeeld van de loterij hebben we:

$$\text{Var}[X] = (1 \times 399.5^2 + 5 \times 19.5^2 + 994 \times 0.5^2) / 1000 = 161.75 \quad \text{en} \quad \sigma_X = 12.72$$

We kunnen de volgende stelling voor $\text{Var}[X]$ bewijzen (de daarin genoemde eigenschap is eenvoudiger te gebruiken dan de definitie):

Stelling

$$\text{Var}[X] = E[X^2] - E[X]^2$$

Bewijs: zelf

N.B.: We wijzen er wel op dat deze formule niet zo nauwkeurig is indien we de getallen afronden.

Voor het laatste voorbeeld, $E[X^2] = (1 \times 399^2 + 5 \times 19^2 + 994 \times 1^2)/1000 = 162$; en dus $\text{Var}[X] = 162 - (0.5)^2 = 161.75$

Eigenschappen van verwachtingswaarde en variantie

Voor twee willekeurige getallen a en b met a verschillend van 0 geldt:

- i) $E[aX + b] = a \cdot E[X] + b$
- ii) $|E[X]| \leq E[|X|]$
- iii) $\text{Var}[aX + b] = a^2 \text{Var}[X]$ en $\sigma_{aX+b} = |a| \sigma_X$

Stelling

Voor onafhankelijke veranderlijken X en Y geldt:

- i) $E[XY] = E[X]E[Y]$
- ii) $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]$
- iii) $\text{Var}[XY] = \text{Var}[X] \text{Var}[Y] + E[X]^2 \text{Var}[Y] + \text{Var}[X]E[Y]^2$

Bewijs in de les

Gevolg

Voor onafhankelijke veranderlijken X en Y geldt:

$$\text{Var}[X-Y] = \text{Var}[X] + \text{Var}[Y]$$

Bewijs dit zelf. Maak nooit de fout beide varianties af te trekken !

Voorbeelden

- 1) Werp met een dobbelsteen en beschouw volgende stochastische variabelen

X = aantal ogen dat geworpen wordt,

$Y = 1$ als het aantal ogen even is en 0 als het aantal ogen oneven is.

Dan zijn X en Y afhankelijk:

$$P(X = 3 \text{ en } Y=1) = 0 \text{ terwijl } P(X = 3) P(Y=1) = 1/6 \times 1/2 = 1/12$$

Voor de verwachtingswaarde van de som geldt inderdaad

$$E[X+Y] = E[X] + E[Y] = 4 \text{ (verifieer door beide leden uit te rekenen).}$$

Voor de verwachtingswaarde van het product en voor de variantie van de som hebben we echter: $E[XY] = 2$ terwijl $E[X] E[Y] = 3.5 \times 0.5$ en

$$\text{Var}[X+Y] = 22/6 \text{ terwijl } \text{Var}[X] + \text{Var}[Y] = 35/12 + 1/4 = 19/6$$

- 2) We werpen met twee dobbelstenen en kiezen de stochastieken X en Y als volgt:

X = aantal ogen van de eerste dobbelsteen

$Y = 1$ als het aantal ogen van de tweede dobbelsteen even is en 0 als dit aantal ogen oneven is
 X en Y zijn nu onafhankelijk. Verifieer dat

$$E[X+Y] = E[X]+E[Y] = 4$$

$$E[XY] = E[X] E[Y] = 1.75$$

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] = 38/12 = 19/6$$

4.7 Momenten

In de mechanica worden bij een gegeven object bestaande uit massapunten x_i met gewichten p_i de begrippen totale massa, zwaartepunt en traagheidsmoment (t.o.v. het zwaartepunt) gedefinieerd als : $m = \sum_i p_i$, $g = \sum_i p_i x_i$ en $\sum_i (x_i - g)^2 p_i$

Deze begrippen zijn volledig analoog met de definities van totale kans (=1), verwachtingswaarde en variantie in de waarschijnlijkheidsrekening.

Algemeen kunnen we (zoals in de mechanica) het ruwe moment en het centrale moment van orde k definiëren als:

Definitie

Voor elk natuurlijk getal k definiëren we het ruwe moment α_k en het centrale moment μ_k van orde k door

$$\alpha_k(X) = E[X^k] \text{ en } \mu_k(X) = E[(X - E[X])^k]$$

Bewijs zelf de volgende eigenschappen:

- i) $\alpha_1(X) = E[X]$
- ii) $\mu_1(X) = 0$
- iii) $\mu_2(X) = \text{Var}[X] = \alpha_2(X) - \alpha_1(X)^2$
- iv) $\mu_3(X) = \alpha_3(X) - 3\alpha_1(X)\alpha_2(X) + 2\alpha_1(X)^3$

De momenten van orde drie en vier worden verder behandeld. Zij hebben te maken met de vorm van de verdeling.

De variantie geeft een maat voor de grootte van het gebied, waar we het grootste deel van de "kansmassa" kunnen verwachten.

Dit wordt geïllustreerd door het volgende belangrijke resultaat, dat we ook reeds in de beschrijvende statistiek vermeldden:

Stelling: formule van Chebyshev

Als X een stochastische variabele is met verwachtingswaarde $\alpha_1(X)$ en variantie $\mu_2(X)$, dan geldt voor elke $\lambda > 0$:

$$P(|X - \alpha_1(X)| \geq \lambda) \leq \frac{\mu_2(X)}{\lambda^2}$$

Bewijs: in de les enkel in het geval waar X continu verdeeld is. Het geval waar X discreet verdeeld is laten we als oefening.

4.8 Kengetallen van locatie, schaal en vorm van een stochastiek

4.8.1 Inleiding

We definieerden reeds empirische kengetallen van locatie, schaal en vorm in hoofdstuk 2. Deze grootheden werden berekend met de steekproefgegevens. Nu kunnen we soortgelijke getallen definiëren voor een stochastische variabele m.b.v. zijn theoretische kansverdeling. Zoals reeds eerder gezegd zullen we de laatste noteren met griekse letters.

We weten reeds dat de verwachtingswaarde van een stochastische variabele ons informatie geeft over de locatie of ligging van de kansverdeling; de variantie vertelt ons iets over de spreiding, of schaal van de verdeling. In dit hoofdstuk beperken we ons tot: de mediaan, de modus, het interkwartiel, de mediane absolute afwijking (MAD of median absolute deviation), de scheefheidscoëfficiënt en de kurtosis. Deze parameters geven ons informatie over de locatie, de schaal en de vorm van de kansverdeling van de stochastiek.

4.8.2 Kengetallen van locatie

a. Het rekenkundig gemiddelde

Het rekenkundig gemiddelde wordt gedefinieerd als $E[X]$.

Deze grootheid bestaat niet altijd (bv Cauchy-verdeling: $F_X(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arctg}(x)$) en wordt sterk beïnvloed door uitschieters (is niet robuust).

b. De mediaan

De mediaan is het 50%-percentiel: $\operatorname{med}(X) = F_X^{-1}\left(\frac{1}{2}\right)$

Onderstaande figuur illustreert het berekenen van de mediaan in gevallen waar $F_X^{-1}\left(\frac{1}{2}\right)$ niet bestaat of niet eenduidig bepaald is.

Figuur 4.7: Het berekenen van de mediaan in specifieke gevallen

Het berekenen van de mediaan kost meer moeite dan het berekenen van het gemiddelde, maar het is het een meer robuuste parameter. Voor symmetrische verdelingen geldt: $\operatorname{med}(X) = E[X]$

c. De modus

X discreet \rightarrow modus = x_j waarvoor $p_j = P(X=x_j)$ maximaal is.

X continu \rightarrow modus = x waarvoor f_x zijn absolute maximum bereikt.

Er kunnen meerdere modi bestaan. Indien er slechts één modus is noemen we de verdeling unimodaal.

4.8.3 Kengetallen van schaal

a. De variantie en de standaardafwijking

Deze grootheden werden reeds gedefinieerd in paragraaf 4.6.

De variantie: $\operatorname{Var}[X] = E[(X - E[X])^2]$

De standaardafwijking van X is de vierkantswortel van de variantie: $\sigma_X = \sqrt{\operatorname{Var}[X]}$

b. Het interkwartiel (IQR)

Dit is het verschil tussen het 75%- en het 25%-percentiel:

$$\text{Interkwartiel} = F_X^{-1}\left(\frac{3}{4}\right) - F_X^{-1}\left(\frac{1}{4}\right)$$

c. De mediane absolute afwijking (MAD)

$$\text{MAD}(X) = \text{med}|X - \text{med}(X)|$$

50% van de kansmassa bevindt zich tussen med-MAD en med+MAD

4.8.4 Kengetallen van vorm

a. De scheefheidscoëfficiënt

De scheefheidscoëfficiënt wordt gedefinieerd mbv momenten en is gelijk aan:

$$\gamma_1(X) = \frac{\mu_3(X)}{\sigma_X^3}$$

Indien X symmetrisch is zal het derde centrale moment nul zijn (elk moment van oneven orde is in dit geval gelijk aan nul). Indien de dichtheidsfunctie een brede lange staart naar rechts heeft, dan zullen de positieve afwijkingen in het centrale derde moment overwegen en zal dit moment positief zijn en dus ook de scheefheidscoëfficiënt. Analoog zal voor een dichtheidsfunctie met brede lange staart naar links de scheefheidscoëfficiënt negatief zijn.

Voorbeeld: In het voorbeeld van de loterij zal de scheefheidscoëfficiënt positief zijn (=31). (Ga na)

b. De kurtosis

De coëfficiënt van kurtosis wordt analoog aan de scheefheidscoëfficiënt gedefinieerd:

$$\gamma_2(X) = \frac{\mu_4(X)}{\sigma_X^4} - 3$$

In de vierde centrale momenten is de bijdrage van de staarten veel groter dan in de lagere momenten zoals de variantie. Als de staart “dik” is zal dit vierde moment grotter zijn dan bij een “dunne” staart. Een mate voor de dikte van de staart is dus:

$$b_2(X) = \frac{\mu_4(X)}{\sigma_X^4}$$

Deze coëfficiënt is steeds positief. Voor een normale verdeling is $b_2(X) = 3$. De coëfficiënt van kurtosis neemt dus de normale verdeling als referentie. Indien deze coëfficiënt negatief is (leptocurtic) heeft de verdeling “dunnere” staarten dan de normale verdeling; indien positief (platycurtic) “dikkere”. Deze coëfficient is nul (mesocurtic) voor de normale verdeling.

4.8.5 Covariantie en correlatiecoëfficiënt

Neem 2 stochastische variabelen X en Y. Als X en Y onafhankelijk zijn, dan geldt, zoals we gezien hebben, dat $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$. In het algemeen (voor afhankelijke stochastieken) hebben we:

$$\begin{aligned} \sigma_{X+Y}^2 &= E[(X+Y - E[X+Y])^2] \\ &= E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] \\ &= \sigma_X^2 + \sigma_Y^2 + 2E[(X - E[X])(Y - E[Y])] \end{aligned}$$

De term $2E[(X - E[X])(Y - E[Y])]$ geeft dus een idee van de mate van onderlinge afhankelijkheid van X en Y. Dit leidt tot de begrippen covariantie en correlatie.

Definitie

De covariantie van twee stochastische variabelen wordt gegeven door:

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Definitie

De correlatiecoëfficiënt van twee stochastische variabelen wordt gegeven door:

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Gevolg

De correlatiecoëfficiënt is begrensd: $1 \geq \rho \geq -1$

NB: Analoog aan de alternatieve formule voor variantie kunnen we de covariantie ook op een alternatieve manier berekenen door de formule (numeriek niet zo stabiel):

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Definitie

Twee stochastische variabelen X en Y heten ongecorreleerd als:

$$Cov(X, Y) = 0 \text{ of } E[XY] = E[X]E[Y] \text{ of } \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

Gevolg

Twee onafhankelijke stochastische variabelen zijn ongecorreleerd. Het omgekeerde geldt echter niet zoals blijkt uit het volgende voorbeeld.

Voorbeeld

$Z = (X, Y)$ met dichtheidsfunctie: $f_Z(x, y) = \begin{cases} 1/\pi & \text{als } x^2 + y^2 \leq 1 \\ 0 & \text{als } x^2 + y^2 > 1 \end{cases}$

X en Y zijn niet gecorreleerd want $E[XY] = 0$ en $E[X] = 0 = E[Y]$

X en Y zijn niet onafhankelijk want $P(X > \sqrt{2}/2, Y > \sqrt{2}/2) = 0$ en

$P(X > \sqrt{2}/2) P(Y > \sqrt{2}/2) \neq 0$

Hoofstuk 5: Verdelingen

In dit hoofdstuk geven we een overzicht van enkele veel gebruikte verdelingen. Herinner je dat een verdeling een stochastische veranderlijke karakteriseert en aldus een theoretisch model vastlegt.

De discrete stochastische veranderlijken karakteriseren we met discrete kansverdelingen en de continue beschrijven we met hun dichtheidsfunctie.

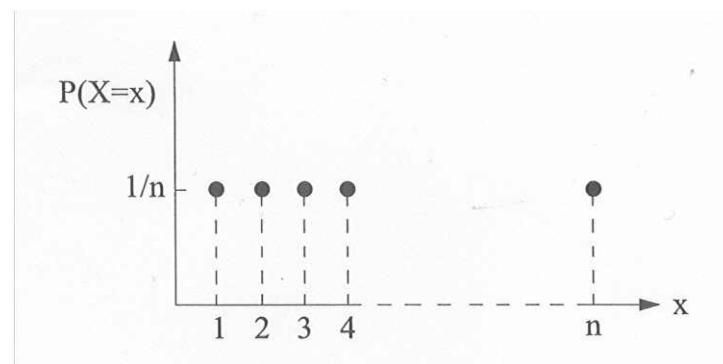
5.1 Discrete verdelingen

5.1.1 Discreet uniforme verdeling

Een uniforme verdeling beschrijft de eigenschap dat elke uitkomst dezelfde kans heeft. Als een discrete stochastische veranderlijke X juist n verschillende waarden x_1, \dots, x_n kan aannemen en men zegt ons dat X uniform verdeeld is, dan betekent dit dat $P(X=x_i) = 1/n$ voor $i = 1, \dots, n$.

Voor het speciale geval dat x_1, x_2, \dots, x_n gelijk is aan $1, 2, \dots, n$ kan je berekenen dat:

$$E[X] = \frac{n+1}{2} \quad \text{en} \quad Var[X] = \frac{n^2 - 1}{12}$$



Figuur 5.1 : Staafdiagram van een discrete uniforme verdeling op n punten

Figuur 5.2 : Kansverdeling en staafdiagram van een discrete uniforme verdeling op 9 punten

Voorbeeld

Als we met een eerlijke dobbelsteen gooien en X is het aantal ogen, dan is X uniform verdeeld op $\{1,2,\dots,6\}$. We kunnen dan kansen uitrekenen zoals:

$$P(X \leq 5) = P(X=1) + \dots + P(X=5) = 5/6$$

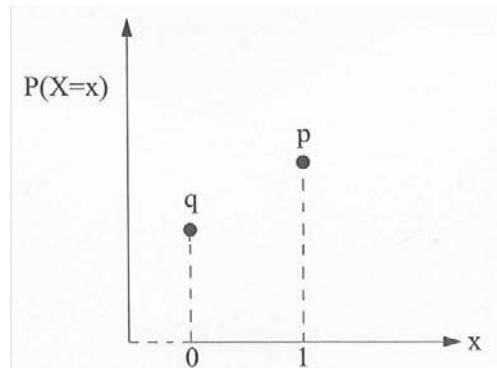
wat je ook als volgt kunt berekenen:

$$P(X \leq 5) = 1 - P(X > 5) = 1 - P(X=6) = 1 - 1/6 = 5/6$$

5.1.2 De Bernoulli verdeling

Een experiment dat 2 verschillende uitkomsten kan hebben (succes en mislukking) noemen we een Bernoulli experiment. De kans op succes noteren we door p en de kans op mislukking door q . Bemerk dat $q = 1-p$. Als we nu succes associëren met 1 en mislukking met 0 dan spreken we over een discrete stochastische variabele X die Bernoulli verdeeld is met parameter p ; $X \sim B(1;p)$. Er geldt dan:

$$E[X] = p \quad \text{en} \quad Var[X] = pq$$



Figuur 5.3 : Staafdiagram van een Bernoulli verdeling

Figuur 5.4 : De Bernoulli verdeling met $p=2/3$ en bijbehorend staafdiagram

Voorbeeld

Indien je een muntstuk opgooit en met kop 1 associeert en met munt 0, dan heb je een discrete stochastische variabele X die Bernoulli verdeeld is met parameter $\frac{1}{2}$.

5.1.3 De binomiale verdeling

Herhaal eenzelfde Bernoulli experiment n keer na elkaar op zo'n manier dat de uitkomsten van de verschillende herhalingen onafhankelijk van elkaar zijn. Een typische realisatie van dergelijk experiment ziet eruit als een geordend n -tal van successen en mislukkingen,

bijvoorbeeld $\{S,S,M,S,M,\dots,M,M\}$. Onderstel dat we nu enkel geïnteresseerd zijn in het aantal successen bij n herhalingen; we noteren dit aantal door X . Er geldt: $X = X_1 + X_2 + \dots + X_n$ met $X_i \sim B(1; p)$ voor $1 \leq i \leq n$. We zeggen dan dat X binomiaal verdeeld is met de parameters n en p en schrijven $X \sim B(n; p)$. Bemerk dat X de waarden $0, 1, 2, \dots, n$ kan aannemen. Door gebruik te maken van de combinatieleer (om het aantal disjuncte gebeurtenissen te tellen, die allemaal eenzelfde aantal successen opleveren) en van de produktregel voor onafhankelijke gebeurtenissen (om de kans van een typische realisatie te bepalen) vinden we dat de kansverdeling van X gegeven wordt door:

$$P(X=x) = \binom{n}{x} p^x q^{n-x} \quad \text{voor } x=0,1,2,\dots,n \text{ en met } 0 < p < 1 \text{ en } q = 1-p.$$

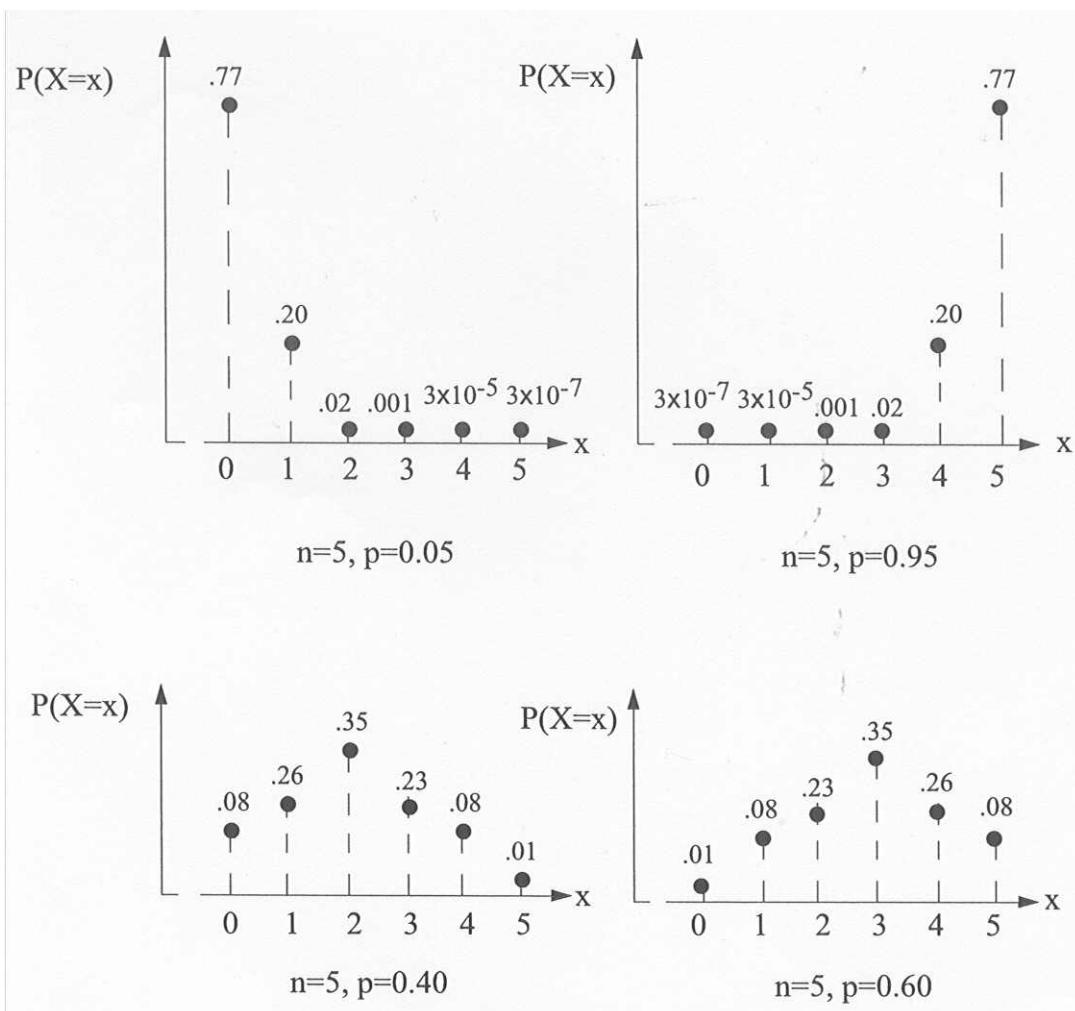
N.B.: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is de binomiaalcoëfficiënt

Met een beetje rekenwerk (gebruik het binomium van Newton) kan je aantonen dat voldaan is aan:

$$\sum_{x=0}^n \binom{n}{x} p^x q^{n-x} = 1$$

Tevens: $E[X] = np$ en $Var[X] = npq$

Om dit te bewijzen kun je gebruik maken van het feit dat X de som is van n onafhankelijke Bernoulli experimenten.



Figuur 5.5 : Staafdiagram van de binomiaalverdeling voor verschillende waarden van de parameters n en p

Figuur 5.6 : Staafdiagram en verdelingsfunctie van B(10;0.5)

Voor n klein en voor geselecteerde waarden van p is de verdelingsfunctie van de binomiale verdeling getabellleerd. In vele boeken vind je ook tabellen met de kansen $P(X=x)$. Let dus goed op bij het gebruik van een tabel.

Tevens wordt gebruik gemaakt van het feit dat, bij een vast aantal herhalingen n, het totaal aantal successen samen met het totaal aantal mislukkingen juist n oplevert. Dus als $X \sim B(n;p)$, dan is:

$$\begin{aligned} P(X=x) &= P(\text{aantal successen} = x) \\ &= P(\text{aantal mislukkingen} = n-x) \\ &= P(Y=n-x) \text{ met } Y \sim B(n;q) \text{ en } q=1-p \end{aligned}$$

Voorbeeld

Als $X \sim B(7;0.65)$ dan is: $P(X=3) = P(Y=7-3) = P(Y=4)$ met $Y \sim B(7;0.35)$.

Uit de tabel voor de cumulatieve kansen kun je nu aflezen dat $P(Y=4) = P(Y \leq 4) - P(Y \leq 3) = 0.9444 - 0.8002 = 0.1442$

$$\text{En dus } P(X=3) = \binom{7}{3} 0.65^3 0.35^4 = 0.1442$$

Voorbeeld

In België bezoekt 30% van de jeugd tussen 14 en 19 jaar wekelijks een disco. Bij een onderzoek naar het uitgangsgedrag onder jongeren worden 15 personen tussen de 14 en 19 jaar ondervraagd. Bereken tot 3 decimalen nauwkeurig de kans dat van die 15 personen er

a. 5 wekelijks een discotheek bezoeken

b. hoogstens twee wekelijks een discotheek bezoeken.

Antwoord: Noem N het aantal jongeren uit de groep van 15 dat wekelijks een discotheek bezoekt. We zijn dus op zoek naar a) de kans $P(N=5)$ en b) de kans $P(N \leq 2)$. N telt eigenlijk het aantal successen in 15 onafhankelijke Bernoulli-experimenten, namelijk: je neemt een jongere, die is met 30% kans een uitgangstype en met 70% kans geen uitgangstype. Dat herhaal je 15 keer. De stochastische veranderlijke N is dan binomiaal verdeeld met parameters 15 en 0.3. $N \sim B(15;0.3)$. Je kunt nu weer de tabel met de cumulatieve kansen gebruiken. Dit levert:

$$a. P(N=5) = P(N \leq 5) - P(N \leq 4) = 0.7216 - 0.5155 = 0.206$$

$$b. P(N \leq 2) = 0.127$$

Je kan natuurlijk ook de formule gebruiken: $P(N = n) = \binom{15}{n} 0.3^n 0.7^{15-n}$

Voorbeeld

Als je een eerlijk muntstuk 8 keer gooit dan heb je kans 0.5 dat je 4 keer kop en 4 keer munt vindt. Of niet soms?

Noem kruis gooien succes, dan kan je bovenstaand experiment beschrijven door een binomiale verdeling met $n=8$ en $p=0.5$. Stel door X het totaal aantal successen bij 8 herhalingen voor, dan is $X \sim B(8; 0.5)$. Juist 4 keer kruis treedt op met kans: $P(X=4) = 0.6367 - 0.3633 = 0.2734$

5.1.4 De geometrische verdeling

Onderstel dat je eenzelfde Bernoulli experiment herhaaldelijk uitvoert (onafhankelijke herhalingen) en dat je dit zolang volhoudt tot je voor de eerste keer succes hebt. Noem X het aantal mislukkingen voor het eerste succes. Dan is X een stochastische variabele die de waarden $0, 1, 2, \dots$ kan aannemen en de kansverdeling is gegeven door:

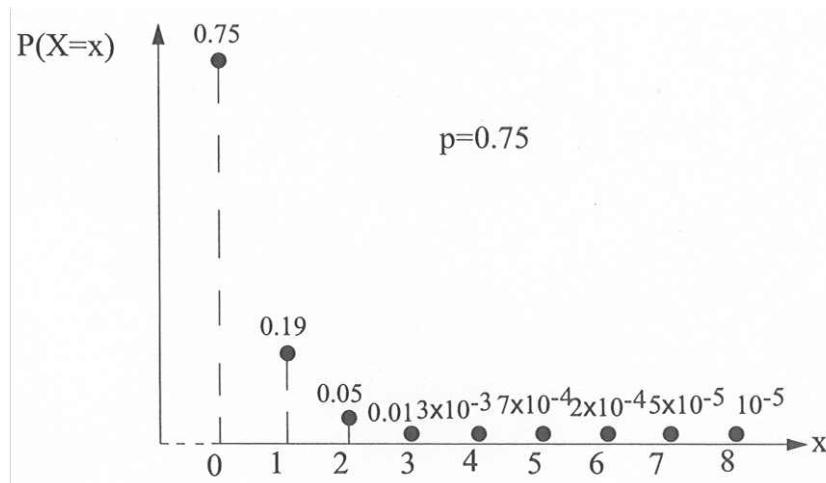
$$P(X=x) = q^x p \text{ voor } x = 0, 1, 2, \dots \text{ en } 0 < p < 1 ; q = 1-p$$

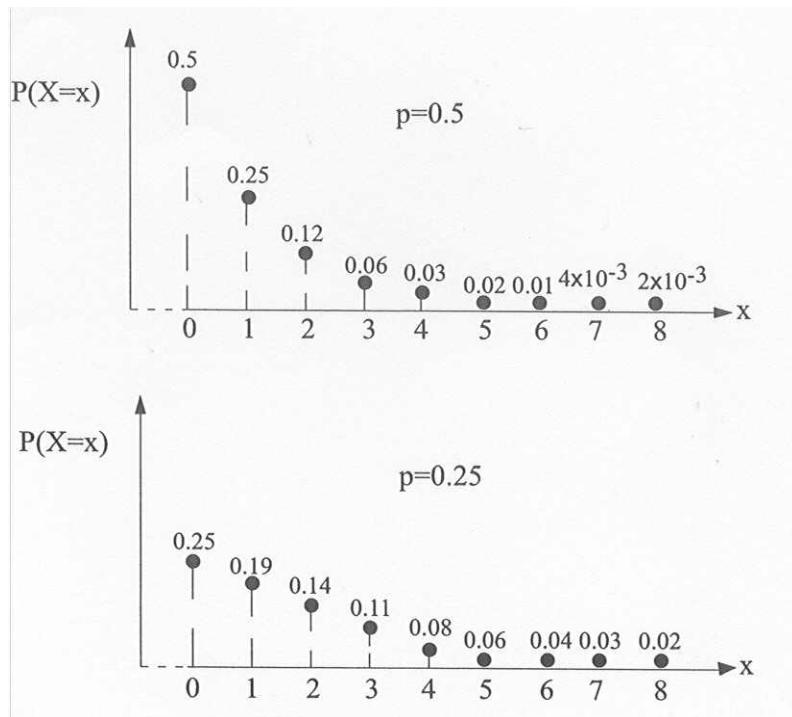
Bewijs dat $\sum_{i=0}^{\infty} q^i p = 1$

We zeggen dat X geometrisch verdeeld is met parameter p . De geometrische verdeling wordt ook wel eens Pascal verdeling genoemd. De naam "geometrisch" komt van de bemerking dat de waarden die de kansverdeling aanneemt de termen van een geometrische (meetkundige) reeks zijn met rede q .

Voor het gemiddelde en de variantie geldt:

$$E[X] = \frac{q}{p} \quad \text{en} \quad Var[X] = \frac{q}{p^2}$$





Figuur 5.7 : Staafdiagram van de geometrische verdeling voor verschillende waarden van p

Figuur 5.8 : Kansverdeling en staafdiagram van de geometrische verdeling met $p=0.3$

Gevolg

$$\text{i)} \quad P(X \geq k) = \sum_{x=k}^{\infty} q^x p = \frac{p q^k}{1-q} = q^k$$

$$\text{ii)} \quad P(X \geq i+j | X \geq i) = \frac{P(X \geq i+j)}{P(X \geq i)} = \frac{q^{i+j}}{q^i} = q^j = P(X \geq j)$$

Als je bij het gooien van een muntstuk succes zegt bij kruis en mislukking bij munt, dan volgt uit bovenstaande regel dat de kans om nog j keer na elkaar munt te gooien en dan kruis, gegeven dat je al i keer na elkaar munt hebt gegooid, juist dezelfde is als de kans om j keer na elkaar munt te gooien en dan kruis wanneer je pas aan het experiment begint. Een geometrische stochastische variabele heeft dus geen geheugen en lijdt dus aan geheugenverlies.

Voorbeeld

Op een atol in de Stille Oceaan is de leefruimte en de hoeveelheid voedsel beperkt. Om overbevolking te voorkomen wil de stammoeder het aantal kinderen beperken. Om de

vrouwen, die allen graag dochters zouden hebben, niet de mogelijkheid te ontnemen om een dochter te baren, bepaalt zij, dat een vrouw na het baren van een dochter niet meer zwanger mag worden. Zal zij in haar opzet slagen ?

Laat ons veronderstellen dat de kans op het baren van een zoon of een dochter even groot is en laten we de kindersterfte e.d. verwaarlozen. $P(k+1^{\text{ste}} \text{ kind is een dochter en de eerste } k \text{ kinderen zijn zonen}) = (\frac{1}{2})^{k+1}$. We hebben hier dus een geometrische verdeling met $p=1/2$. Het gemiddelde aantal kinderen per vrouw komt dus uit op 1 (de dochter) plus de verwachtingswaarde van deze verdeling ($p/q=1$). De maatregel werkt dus perfect.

5.1.5 De hypergeometrische verdeling

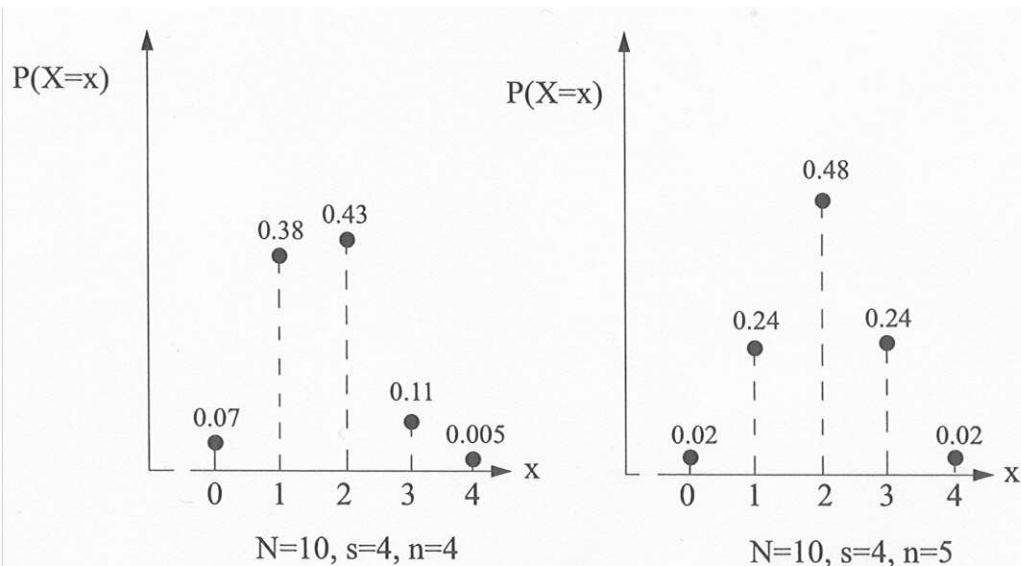
Onderstel dat een vaas N ballen bevat waarvan er juist s zijn van een bepaalde soort (succes) en de overige N-s van een andere soort. Trek nu lukraak uit deze vaas, en zonder terugleggen, n ballen. Noem X het aantal successen bij deze n trekkingen. Dan is X hypergeometrisch verdeeld met parameters N, s en n; $X \sim H(N; s; n)$. De kansverdeling krijgen we door het aantal gunstige gevallen met het aantal mogelijke te vergelijken:

$$P(X=x) = \frac{\binom{s}{x} \binom{N-s}{n-x}}{\binom{N}{n}}$$

waarbij x een natuurlijk getal en $\max(0, n-(N-s)) \leq x \leq \min(s, n)$.

Tevens als $X \sim H(N; s; n)$ dan is:

$$E[X] = n \frac{s}{N} \quad \text{en} \quad \text{Var}[X] = \frac{N-n}{N-1} n \frac{s}{N} \left(1 - \frac{s}{N}\right)$$



Figuur 5.9 : Staafdiagram van de hypergeometrische verdeling voor verschillende waarden van N, s, n

Voor geselecteerde waarden van de parameters bestaan er tabellen voor de hypergeometrische kansen. Ook kan men gebruik maken van tabellen voor binomiaalcoëfficiënten (driehoek van Pascal). Dikwijls is het nuttig de formule eerst te vereenvoudigen vooraleer aan de berekening te beginnen.

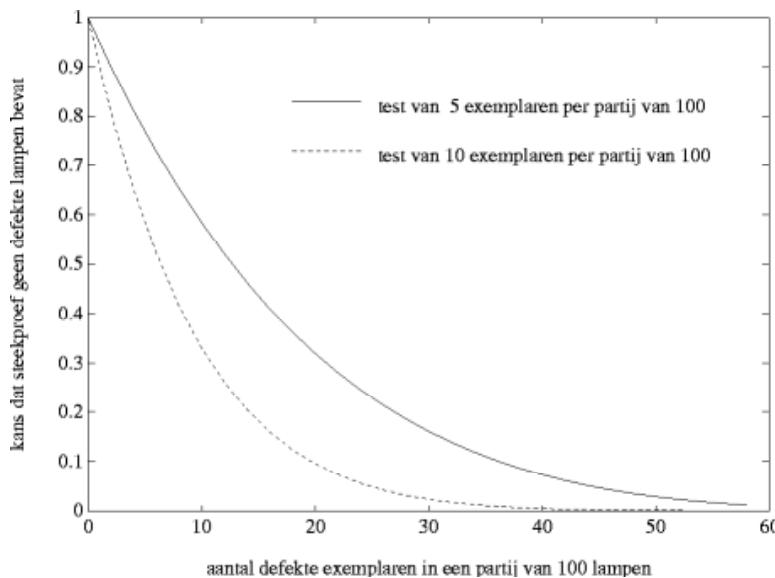
Deze verdeling vindt zijn toepassing bij kwaliteitscontrole van massaproducten.

Voorbeeld

Een lampenwinkel koopt bij de groothandel dozen van 100 lampen. Bij levering worden uit iedere doos 5 willekeurig gekozen lampen getest. Stel dat er in een gegeven doos 5 kapotte exemplaren zitten. Wat is dan de kans dat er minstens één gevonden wordt?

Noem Y het aantal kapotte lampen onder de gekozen lampen. $\rightarrow Y \sim H(100; 5; 5)$ en $P(Y \geq 1)$

$$= 1 - P(Y < 1) = 1 - P(Y = 0) = 1 - \frac{91 \cdot 92 \cdot 93 \cdot 94 \cdot 95}{96 \cdot 97 \cdot 98 \cdot 99 \cdot 100} = 0.23$$



Figuur 5.10 : Percentage defecte lampen in een doos vs. de kans op geen defecte lampen in de steekproef

Voorbeeld

Tijdens je exotische vakantie moet je dringend een operatie ondergaan. Gedurende en na de operatie werden in totaal 15 eenheden plasma toegediend. Het lokale ziekenhuis, waarin je opgenomen was, had in totaal 100 eenheden plasma in voorraad, waarvan er 95 goed en 5 besmet waren. De 15 eenheden waarmee jij bent behandeld werden lukraak uit de voorraad genomen. Wat is de kans dat je besmet bent?

Bovenstaand verhaal is te herleiden tot het hypergeometrische model waarbij 15 keer zonder terugleggen getrokken wordt uit een vaas met 100 ballen waarvan er 95 succes zijn.

Als we het aantal successen X noteren dan is $X \sim H(100; 95; 15)$ zodat:

$$P(X = 15) = \frac{\binom{95}{15} \binom{5}{0}}{\binom{100}{15}} = \frac{95! / 80!}{100! / 95!} = \frac{85 \cdot 84 \cdot 83 \cdot 82 \cdot 81}{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96} = 0.44$$

Dus de kans dat alle 15 plasma-eenheden goed waren is 0.44. De kans op besmetting is de kans op het complement (minstens één van de plasma-eenheden is besmet) en is dus gelijk aan $1 - 0.44 = 0.56$.

Voorbeeld

Een ecoloog wordt nogal eens geconfronteerd met het probleem, het aantal dieren te schatten in een habitat; b.v. het aantal vissen in een vijver. Hierbij kan hij gebruik maken van de

zogenaamde “capture/recapture” techniek. Laat de vijver N vissen bevatten. (N is dus onbekend.) We vangen n vissen, merken ze en zetten ze weer uit in de vijver. Enkele dagen later vangen we m vissen waarvan we zien dat er k gemerkt zijn. Met deze kennis kunnen we nu een idee krijgen van N. Op voorwaarde dat de tweede vangst willekeurig is, is het aantal gemerkte vissen bij de tweede vangst X hypergeometrisch verdeeld: $X \sim H(N; n; m)$. Er geldt dus dat: $k \approx E[X] = m n/N$ en dus is mn/k een goede schatting voor het aantal vissen N.

5.1.6 De Poissonverdeling

Bij veel stochastische fenomenen is men geïnteresseerd in hoeveel keer een gebeurtenis optreedt in een bepaald tijdsinterval (of in een bepaald gebied). Voorbeelden hiervan zijn:

- het aantal dodelijke ongevallen per week in Vlaanderen
- het aantal radioactieve deeltjes dat wordt uitgestraald per tijdseenheid
- het aantal telefoonoproepen dat per uur in een centrale binnenkomt
- het aantal meteorieten dat inslaat op een satelliet bij 1 omwenteling
- het aantal organismen per milliliter
- het aantal lekken per 100km in een pijpleiding
- het aantal auto's dat per uur aan een benzinestation komt tanken
- het aantal pinten dat per kwartier getapt wordt in een studentencafé

Niet alles wat men kan tellen volgt een Poissonverdeling. De manier waarop de gebeurtenissen, die men telt, geschieden moet voldoen aan 3 eisen (het woord “tijd” kan ook vervangen worden door “lengte”, “gebied”,...):

1) de kans dat juist één gebeurtenis optreedt in een klein tijdsinterval h is evenredig met de lengte van dit interval:

$$\lim_{h \rightarrow 0} \frac{P(\text{juist één gebeurtenis in interval van lengte } h)}{h} = \lambda$$

2) de kans dat in een zeer klein tijdsinterval meer dan één gebeurtenis plaatsvindt is verwaarloosbaar klein

3) het aantal gebeurtenissen dat optreedt in disjuncte tijdsintervallen is onafhankelijk van elkaar.

Tellingen van fenomenen die aan bovenstaande criteria voldoen worden gemodelleerd door de Poisson verdeling. De parameter λ kan geïnterpreteerd worden als het gemiddelde aantal per tijdseenheid of de intensiteit waarmee de gebeurtenissen zich voordoen.

We zeggen dat X Poissonverdeeld is met parameter λ ; $X \sim \mathcal{P}(\lambda)$ als X de volgende kansverdeling heeft:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{voor } x = 0, 1, 2, \dots$$

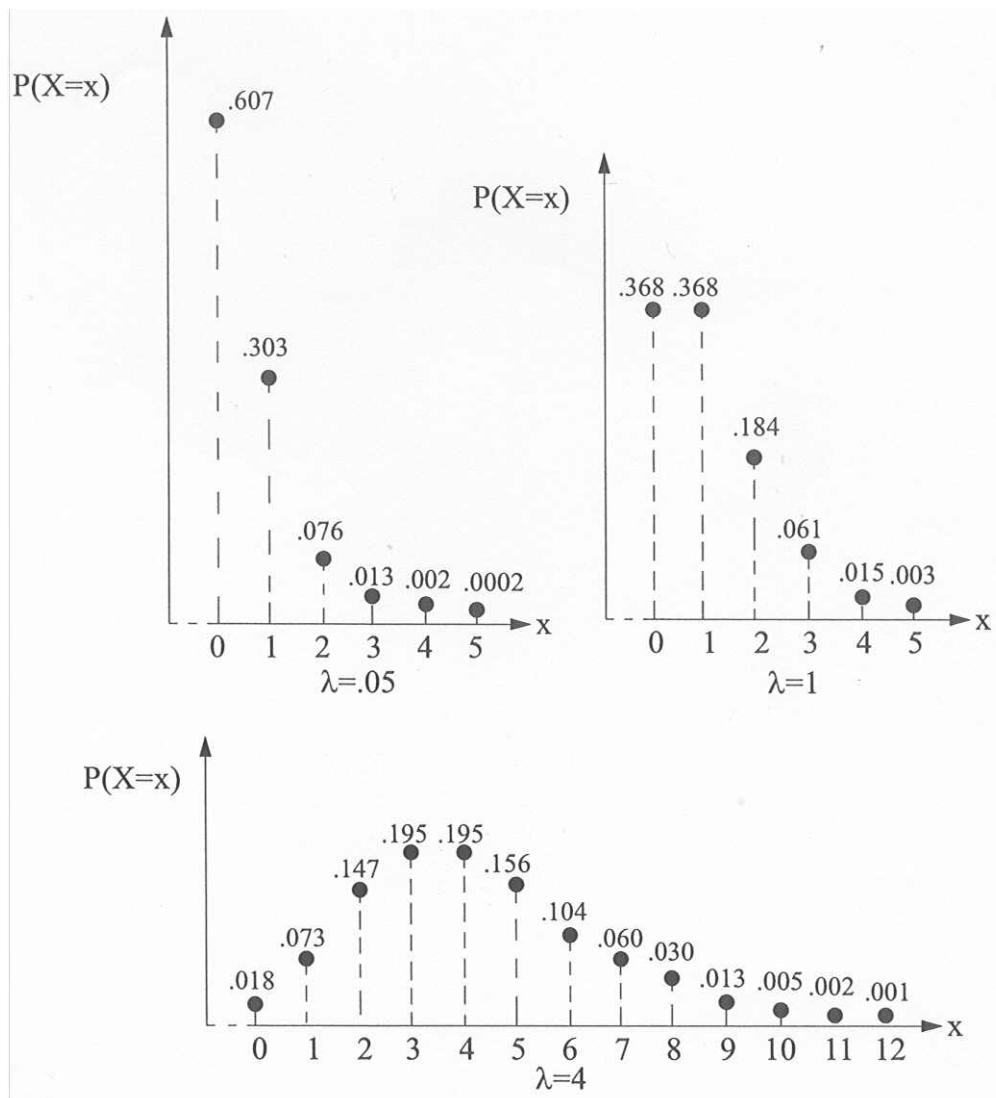
Hierbij telt X het aantal gebeurtenissen per tijdseenheid en λ is het gemiddelde aantal gebeurtenissen per tijdseenheid.

Er geldt:

$$E[X] = \lambda \qquad \text{en} \qquad Var[X] = \lambda$$

Eigenschap

$$\mathcal{P}(\lambda + \mu) = \mathcal{P}(\lambda) + \mathcal{P}(\mu)$$



Figuur 5.11 : Staafdiagram van de Poissonverdeling voor verschillende waarden van λ

Figuur 5.12 : Kansverdeling en staafdiagram van de Poissonverdeling met $\lambda=2.5$

Voor geselecteerde waarden bestaan er tabellen met uitgerekende Poisson kansen of cumulatieve kansen (zie appendix).

Voorbeeld

In een car-wash komen gemiddeld 4 auto's per uur. Onderstel dat aan de voorwaarden van Poisson voldaan is, wat is dan de kans dat gedurende het eerstvolgende uur juist 6 auto's toekomen ?

Daar $X \sim \mathcal{P}(4)$ geldt: $P(X=6) = P(X \leq 6) - P(X \leq 5) = 0.889 - 0.785 = 0.104$ (met tabel)

$$= e^{-4} \frac{4^6}{6!} = 0.104 \quad (\text{met formule})$$

5.2 Continue verdelingen

5.2.1 Continue uniforme verdeling

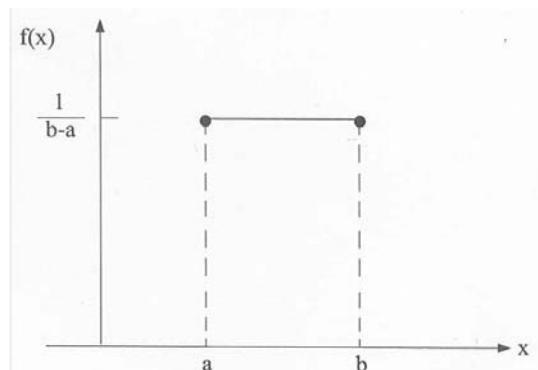
Het continue analoge van een discreet uniforme verdeling beschrijft het lukraak trekken uit een interval $[a,b]$.

X is uniform verdeeld op het interval $[a,b]$, $X \sim U(a,b)$, als X de volgende dichtheid heeft:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{voor } a \leq x \leq b \\ 0 & \text{anders} \end{cases} \quad F_X(x) = \begin{cases} \frac{x-a}{b-a} & \text{voor } a \leq x \leq b \\ 0 & \text{anders} \end{cases}$$

Soms wordt de continue uniforme verdeling ook rechthoekige verdeling genoemd. Voor het gemiddelde en de variantie kan je gemakkelijk zelf uitrekenen dat:

$$E[X] = \frac{a+b}{2} \quad \text{en} \quad Var[X] = \frac{(b-a)^2}{12}$$



Figuur 5.13 : De kansdichtheid van de uniforme verdeling op $[a,b]$

Figuur 5.14 : Kansverdeling en kansdichtheid van een uniforme verdeling op het interval $[0,4]$

5.2.2 De exponentiële verdeling

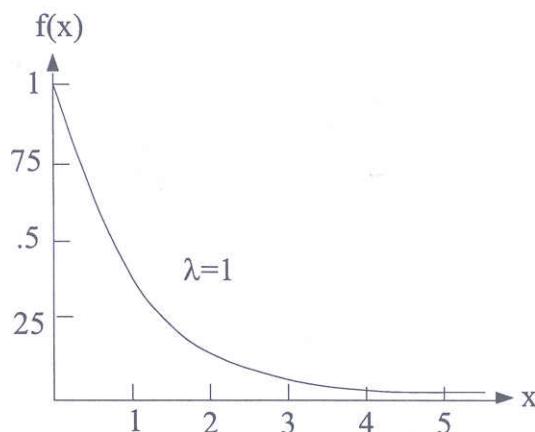
Het continue analoge van een geometrische verdeelde stochastiek is een exponentieel verdeelde stochastiek. Bemerk dat dit een positieve stochastiek is (die waarden aanneemt op $[0, \infty]$). De exponentiële verdeling wordt onder meer gebruikt bij de studie van overlevingstijden, of in de kwaliteitscontrole als men wil nagaan hoe lang het duurt vooraleer een machine defect geraakt.

X is exponentieel verdeeld met parameter λ ; $X \sim \mathcal{E}(\lambda)$, als X bepaald wordt door volgende dichtheid:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{voor } x \geq 0 \text{ en } \lambda > 0 \\ 0 & \text{anders} \end{cases} \quad F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{voor } x \geq 0 \text{ en } \lambda > 0 \\ 0 & \text{anders} \end{cases}$$

Reken na dat:

$$E[X] = \frac{1}{\lambda} \quad \text{en} \quad Var[X] = \frac{1}{\lambda^2}$$



Figuur 5.15 : Kansdichtheid van de exponentiële verdeling

Gevolg

i) $P(X \geq x) = e^{-\lambda x}$

ii) $P(X \geq x_0 + x | X \geq x_0) = \frac{P(X \geq x_0 + x)}{P(X \geq x_0)} = \frac{e^{-\lambda(x_0+x)}}{e^{-\lambda x_0}} = e^{-\lambda x} = P(X \geq x)$

Zoals bij de geometrische verdeling heeft ook de exponentiële geen geheugen. Denk maar aan radio-actief verval. De leeftijdsverdeling van een machine die reeds x_0 jaren werkt is dezelfde als wanneer deze machine splinternieuw is. De exponentiële verdeling beschrijft dus onder meer situaties die niet aan slijtage onderhevig zijn.

5.2.3 De normale verdeling (of Gauss verdeling)

De normale dichtheidsfunctie, met zijn typische klokvormige curve, werd rond 1720 "gecreëerd" door Abraham de Moivre met de bedoeling problemen die verband houden met kansspelen op te lossen. Rond 1870 had (de Belgische wiskundige) Adolph Quetelet het idee om de normale dichtheid te beschouwen als een "ideaal" histogram. Sindsdien is de normale verdeling in de statistiek het meest gebruikte model geworden. De normale kansdichtheid bevat twee parameters μ en σ met $-\infty < \mu < \infty$ en $\sigma > 0$ zodat we eigenlijk te maken hebben met een hele familie dichthesen. De notatie van een normaal verdeelde stochastische veranderlijke is $X \sim \mathcal{N}(\mu; \sigma^2)$.

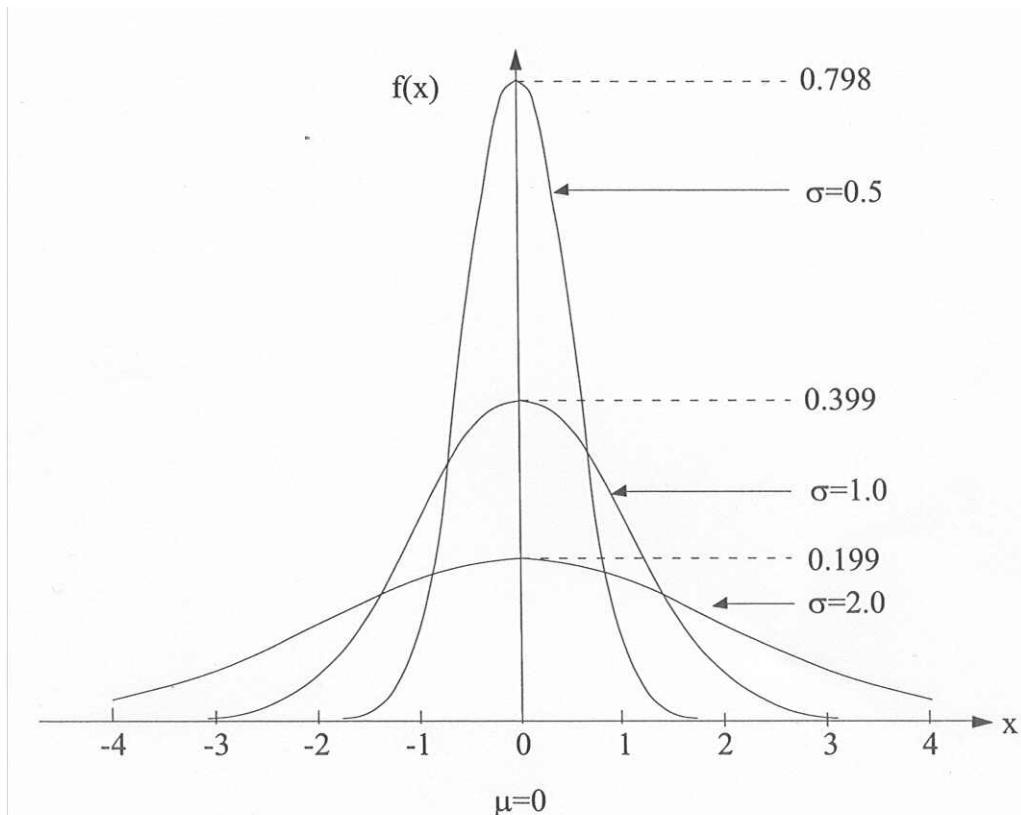
Bemerk dat als tweede parameter σ^2 staat ! De dichtheidsfunctie is :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{voor } -\infty < x < \infty \text{ en met } -\infty < \mu < \infty \text{ en } \sigma > 0.$$

De karakteristieken van de normale verdeling zijn:

$$E[X] = \mu \quad \text{en} \quad \text{Var}[X] = \sigma^2$$

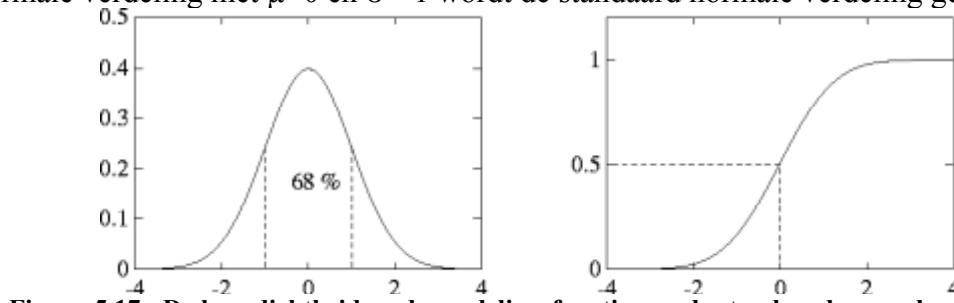
zodat we ook spreken van een normale verdeling met gemiddelde μ en variantie σ^2 . Deze twee karakteristieken leggen de normale verdeling vast door explicet de parameters in de dichtheidsfunctie te bepalen.



Figuur 5.16 : De kansdichtheid van de normale verdeling voor verschillende waarden van σ

De standaard normale verdeling

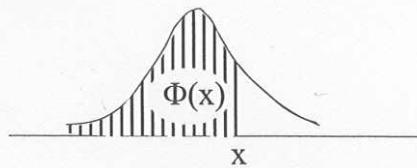
De normale verdeling met $\mu=0$ en $\sigma^2=1$ wordt de standaard normale verdeling genoemd.



Figuur 5.17 : De kansdichtheid en de verdelingsfunctie van de standaard normale verdeling

Voor geselecteerde waarden van x is de waarde van de cumulatieve verdelingsfunctie van de standaard normale verdeling getabellleerd (zie appendix). Het is de gewoonte om deze

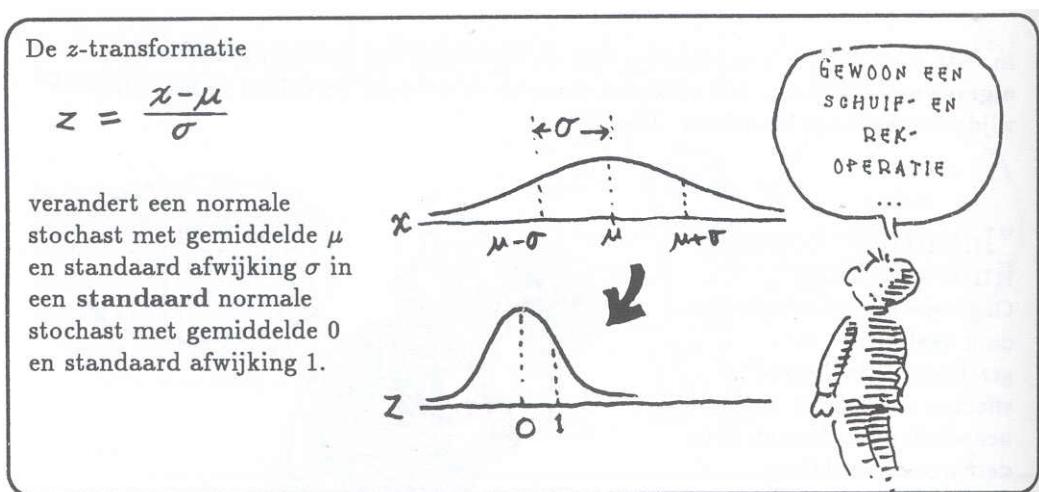
verdelingsfunctie voor te stellen door de hoofdletter Φ en een standaard normale stochastische veranderlijke door Z . Er geldt dan: $\Phi(x) = P(Z \leq x)$ met $Z \sim \mathcal{N}(0;1)$



Figuur 5.18 : $\Phi(x)$ is de aangeduide oppervlakte

De z-transformatie

De volgende figuur illustreert hoe we van de dichtheidsfunctie van een algemene normale verdeling naar de dichtheidsfunctie van een standaard normale verdeling overgaan. Eerst verschuiven we de functie over μ , zodat het maximum van de functie in 0 ligt. We krijgen dan een verdeling met gemiddelde 0 en variantie σ^2 . Daarna herschalen we met σ , zodat we de standaard normale verdeling krijgen. We noemen dit ook wel eens de z-transformatie.



Figuur 5.19 : De z-transformatie

Hieruit volgt dat kansen die verband houden met een algemene normale verdeling kunnen berekend worden door eerst over te stappen op de standaard normale verdeling en dan de tabel te gebruiken. Immers als $X \sim \mathcal{N}(\mu, \sigma^2)$ dan geldt $Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0;1)$. Dus trek van een normaal verdeelde stochastische veranderlijke zijn gemiddelde af en deel het resultaat door zijn standaardafwijking en je bekomt een nieuwe stochastische veranderlijke die standaard normaal verdeeld is. Voor elk reëel getal z geldt immers:

$$\begin{aligned}
 P(Z \leq z) &= P\left(\frac{X - \mu}{\sigma} \leq z\right) = P(X \leq \mu + \sigma z) \\
 &= \int_{-\infty}^{\mu + \sigma z} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
 &= \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad \text{na substitutie } t = (x - \mu) / \sigma \\
 &= \Phi(z) \quad \text{want dit is de verdelingsfunctie van een standaard normale verdeling}
 \end{aligned}$$

Het berekenen van kansen

Hieruit volgt nu voor $X \sim \mathcal{N}(\mu; \sigma^2)$ dat:

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Dit laatste is afleesbaar in de tabel van de standaard normale verdeling.

Bemerk dat in de tabel de verdelingsfunctie $\Phi(z)$ enkel gegeven is voor positieve waarden van z . Vermits de standaard normale verdeling symmetrisch is rond 0 geldt: $\Phi(-z) = 1 - \Phi(z)$. Dus:

$$P(Z \leq -x) = 1 - P(Z \leq x) \quad (\text{N.B.: } P(Z \leq x) = P(Z < x) \text{ voor continue verdelingen})$$

De volgende voorbeelden met bijbehorende figuren geven een goed overzicht van hoe we de kansen met behulp van de tabel moeten berekenen.

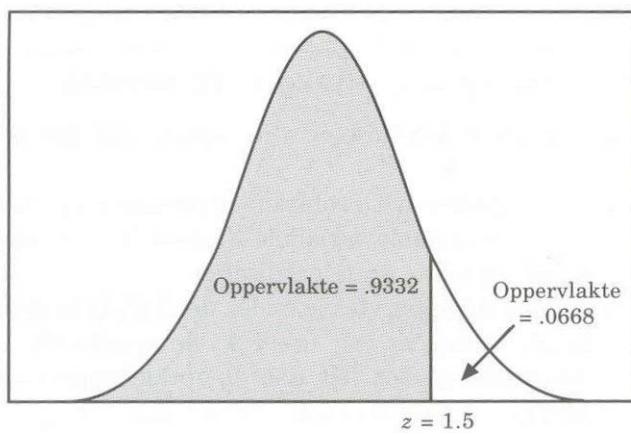
Voorbeeld 1

De lengte van vrouwen X is bij benadering normaal verdeeld met gemiddelde 166.4cm en standaardafwijking 6.4cm. Welk fractie van de vrouwen heeft een lengte van hoogstens 176cm.

$$X \sim \mathcal{N}(166.4; 6.4^2) \rightarrow Z = \frac{X - 166.4}{6.4} \sim \mathcal{N}(0, 1)$$

$$P(X \leq 176) = P\left(Z \leq \frac{176 - 166.4}{6.4}\right) = P(Z \leq 1.5) = \Phi(1.5) = 0.9332$$

93% van de vrouwen zijn niet groter dan 176cm



Figuur 5.20

Voorbeeld 2

Een monitor ten behoeve van grafische computerbeelden heeft een dunmazig scherm achter het beeldoppervlak. Tijdens de assemblage wordt het scherm uitgerekt en op een metalen frame gelast. Bij een te lage spanning zullen in dit stadium plooien ontstaan, bij een te hoge spanning zal het scherm scheuren. De spanning wordt gemeten met een elektrisch instrument waarop men de resultaten in mV afleest. Op dit moment volgen de afgelezen waarden X voor opeenvolgende monitoren de $\mathcal{N}(275; 1849)$ verdeling.

- De minimale acceptabele spanning komt overeen met een aflezing van 200mV. Welk percentage van de monitoren haalt minstens deze limiet?
- Een spanning boven de 375mV doet het scherm scheuren. Welk percentage van de monitoren heeft acceptabele spanningswaarden?

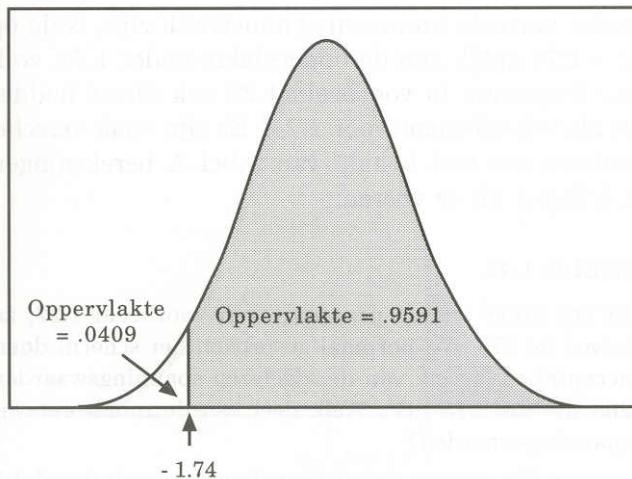
Antwoord:

i) Wij willen de relatieve frequentie van $X \geq 200$ uitrekenen, waarbij $X \sim \mathcal{N}(275, 1849)$ is. $X \geq 200$ betekent voor de standaard normaal verdeelde grootheid $Z = (X - 275)/43$:

$$Z \geq (200 - 275)/43 = -1.74$$

Lees nu in de tabel $\Phi(1.74) = 0.9591$

Dit wil zeggen dat $P(Z \geq -1.74) = 1 - P(Z < -1.74) = P(Z \leq 1.74) = 0.9591$

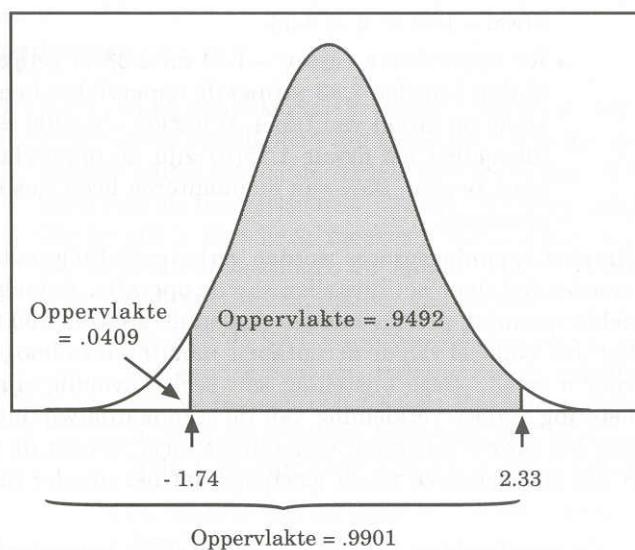


Figuur 5.21

ii) $P(X \leq 375) = P(Z \leq 100/43) = P(Z \leq 2.33) = \Phi(2.33) = 0.9901$

Voor het percentage van acceptabele monitoren moeten we $P(200 \leq X \leq 375)$ kennen en dus de oppervlakte tussen -1.74 en 2.33 (zie figuur). Deze oppervlakte is gelijk aan de oppervlakte beneden 2.33 minus de oppervlakte beneden -1.74 ($= 1 - \text{oppervlakte boven } -1.74$) en dus gelijk aan: $0.9901 - (1 - 0.9591) = 0.9492$.

95% van de monitoren heeft dus een acceptabele spanning.

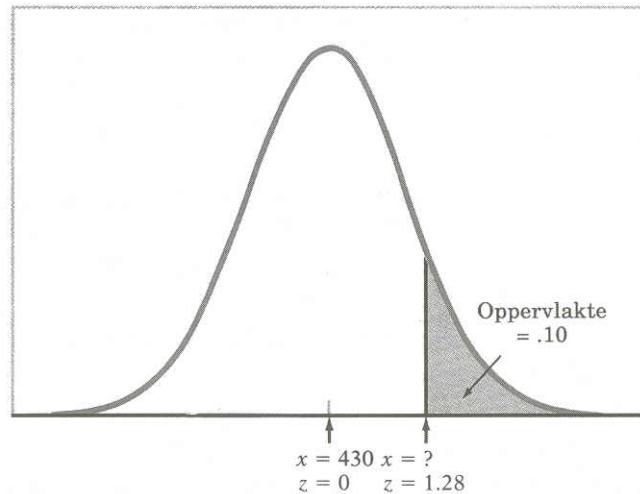


Figuur 5.22

Voorbeeld 3

Zij X de scores in de "American Aptitude Test" (SAT-test) voor verbale vaardigheid van leerlingen uit de hoogste klas van de middelbare school en stel dat $X \sim \mathcal{N}(430; 10000)$. Hoe hoog moet een leerling scoren om bij de top-10% van de leerlingen te horen die aan de test meedoen?

We zoeken eerst z met $P(Z>z)=0.1$ ofwel $P(Z\leq z)=0.9$; in de tabel vind je dat $\Phi(1.28)=0.9$ en dus is z gelijk aan 1.28; nu moeten we terug overgaan naar de stochastische variabele X en dus naar een waarde x i.p.v. z : $(x-430)/100 = 1.28$ ofwel $x = 558$
Een student moet dus ten minste 558 scoren.



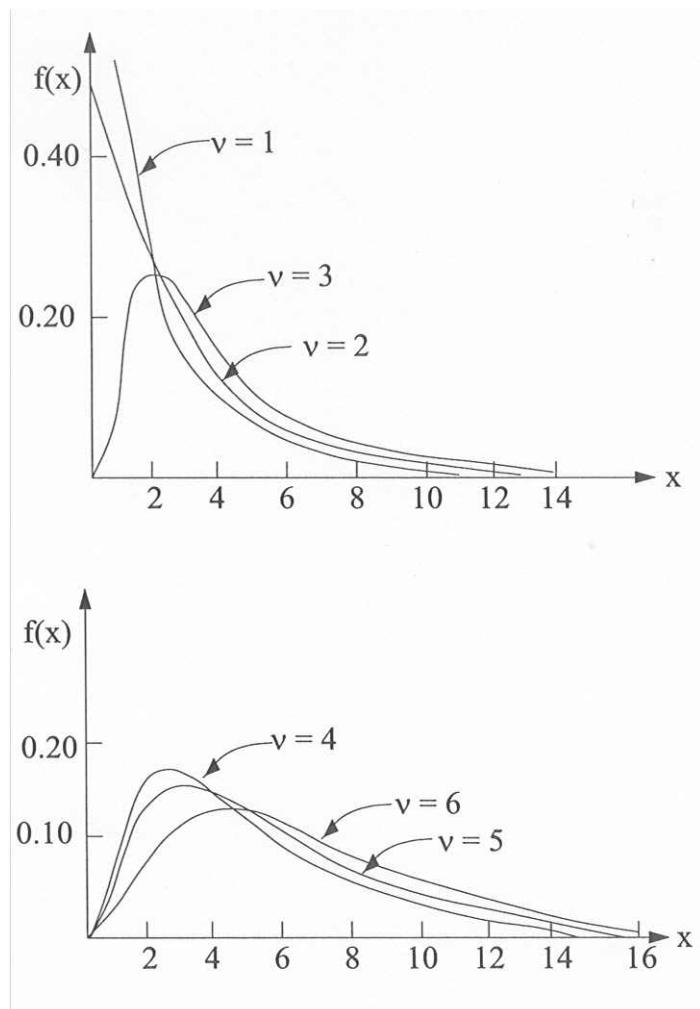
Figuur 5.23

5.2.4 De Chi-kwadraat verdeling

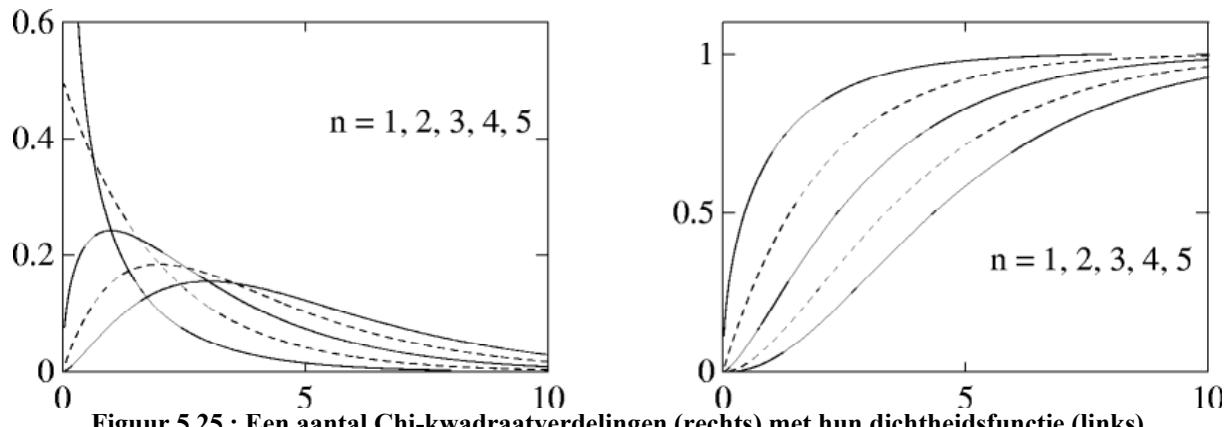
De Chi-kwadraat verdeling (χ^2) komt vooral voor als model voor grootheden bij bepaalde statistische procedures. Als Z_1, Z_2, \dots, Z_v onafhankelijke stochastische veranderlijken zijn die allemaal standaard normaal verdeeld zijn, dan is $X = \sum_{i=1}^v Z_i^2$ verdeeld als een χ^2 verdeling met v vrijheidsgraden (de parameter v in de dichtheidsfunctie wordt het aantal vrijheidsgraden genoemd). Als $X \sim \chi^2(v)$ dan is de dichtheidsfunctie gegeven door:

$$f(x) = \frac{1}{\Gamma(v/2)} \cdot \frac{1}{2^{v/2}} \cdot \left(\frac{x}{2}\right)^{(v/2)-1} \cdot e^{-x/2} \quad \text{voor } x > 0 \text{ en } v = 1, 2, \dots$$

Tevens is: $E[X] = v$ en $Var[X] = 2v$



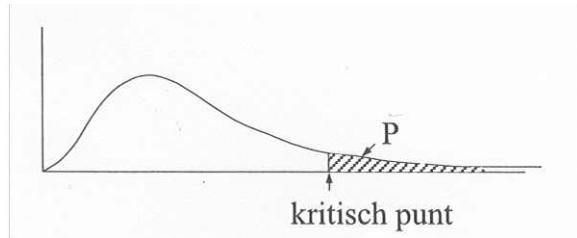
Figuur 5.24 : De dichtheid van de Chi-kwadraatverdeling voor verschillende vrijheidsgraden



Figuur 5.25 : Een aantal Chi-kwadraatverdelingen (rechts) met hun dichtheidsfunctie (links)

In de appendix vind je tabellen met de kwantilen van een Chi-kwadraat verdeling met verschillende vrijheidsgraden.

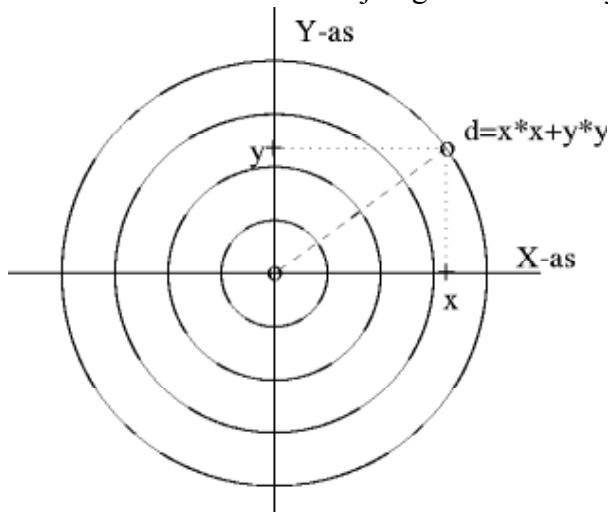
In sommige teksten ga je tabellen vinden met de waarde van het kritisch punt x waarvoor $P(X>x)=p$ en dit voor verschillende waarden van p . (Wat is het verband met onze tabel ?)



Figuur 5.26

Voorbeeld

Een schutter mikte op een roos en treft het punt (x, y) . Als de afwijkingen tot het midden $(0,0)$ zowel in de X- als in de Y-richting normaal verdeeld en onafhankelijk zijn, dan heeft het kwadraat van de afstand tot de roos $d^2 = x^2 + y^2$ een $\chi^2(2)$ verdeling. Indien 2 schutters een pijl afschieten zal de som van de kwadratische afwijkingen verdeeld zijn als $\chi^2(4)$.



Figuur 5.27

5.2.5 De t-verdeling

De t-verdeling (ook Student t-verdeling genoemd) werd in 1908 opgesteld door W. Gosset. Deze verdeling komt voor als model voor bepaalde grootheden, die in de statistische procedures worden gebruikt.

Als Z standaard normaal verdeeld is en $U \sim \chi^2(v)$ waarbij Z en U onafhankelijk zijn, dan heeft

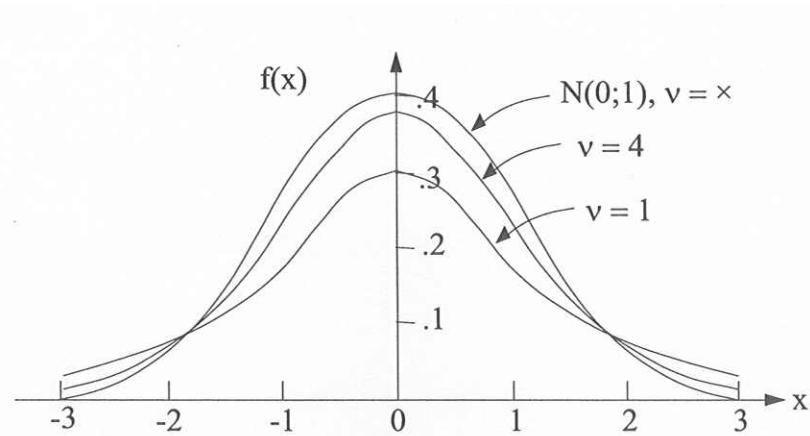
$$X = \frac{Z}{\sqrt{U/v}} \quad \text{een t verdeling met } v \text{ vrijheidsgraden.}$$

Als $X \sim t(v)$ dan geldt:

$$f(x) = \frac{\Gamma((v+1)/2)}{\Gamma(v/2)} \cdot \frac{1}{\sqrt{v\pi}} \cdot \left(1 + \frac{x^2}{v}\right)^{-(v+1)/2} \quad \text{voor } -\infty < x < \infty \text{ en } v = 1, 2, \dots$$

Tevens is: $E[X] = 0 \quad (v > 1) \quad \text{en} \quad Var[X] = \frac{v}{v-2} \quad (v > 2)$

Hoe groter het aantal vrijheidsgraden, hoe beter de verdeling de standaard normale benadert. In de limiet $v \rightarrow \infty$ valt ze er mee samen.

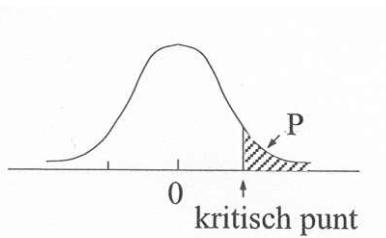


Figuur 5.28 : De dichtheid van de standaard normale en de t-verdeling voor verschillende waarden van v

Dit is nogmaals duidelijk geïllustreerd in onderstaande grafiek. De centrale limietstelling (zie verder) bevestigt deze observatie.

Figuur 5.29 : De dichtheid en de verdelingsfunctie van de standaard normale en de t-verdeling voor verschillende waarden van v

De kritische punten of percentielen q van deze verdeling, waarvoor geldt dat $P(X \leq q) = p$, kun je ook in de tabellen vinden voor verschillende waarden n van het aantal vrijheidsgraden. (Deze worden ook wel eens genoteerd als $q = t_{n,p}$)



Figuur 5.30 $P = 1-p$ is de gearceerde oppervlakte

5.2.6 De $F_{m,n}$ -verdeling (Fisher en Snedecor)

Als X en Y onafhankelijke chi-kwadraat verdeelde stochastische veranderlijken zijn met m respectievelijk n vrijheidsgraden ($X \sim \chi_m^2$ en $Y \sim \chi_n^2$), dan heeft het quotiënt

$$F = \frac{\frac{1}{m}X}{\frac{1}{n}Y}$$

een $F_{m,n}$ -verdeling met m vrijheidsgraden in de teller en n in de noemer en we noteren:

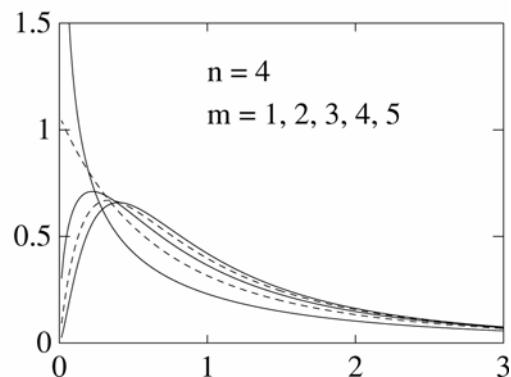
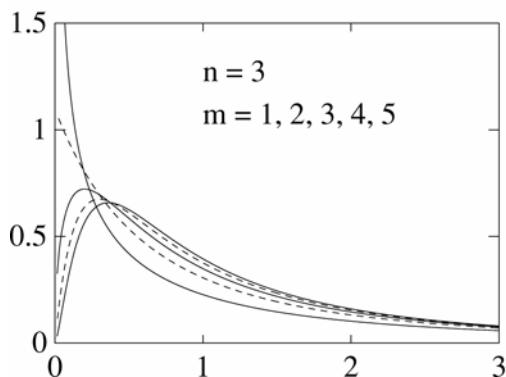
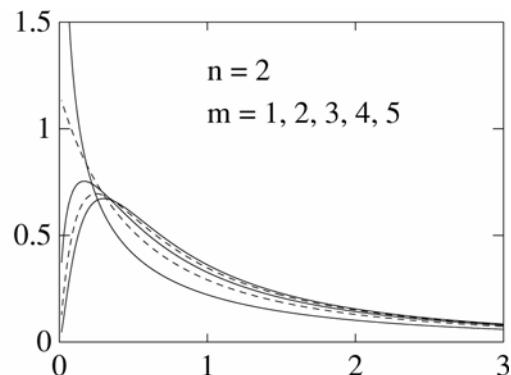
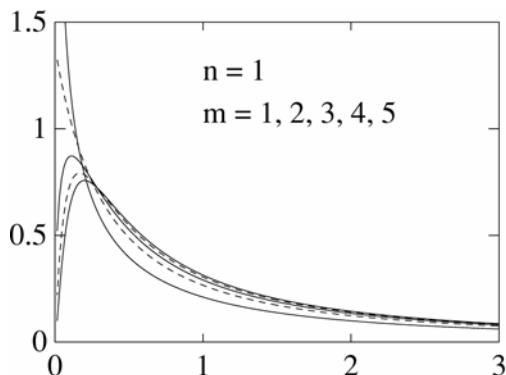
$$F \sim F_{m,n}$$

Er geldt:

$$E[F] = \frac{n}{n-2} \quad (n > 2) \quad \text{en} \quad \text{Var}(F) = \frac{2n^2(m+n-2)}{m(n-4)(n-2)^2} \quad (n > 4)$$

Eigenschappen

- i) De verdelingsfunctie $F_{m,n}$ voldoet aan de symmetrie $F_{m,n}(x) = 1 - F_{n,m}\left(\frac{1}{x}\right)$
- ii) Voor $m=1$ is er een relatie met de t-verdeling: $T \sim t_n \Leftrightarrow T^2 \sim F_{1,n}$



Kansdichthesen van $F_{m,n}$ -verdelingen voor een aantal waarden van n en m .

In de tabellen van de F-verdeling vinden we meestal de kwantieelen van deze verdeling:

$$F_{m,n,\alpha} = F_{m,n}^{-1}(\alpha) \quad \text{zodat} \quad P(F \leq F_{m,n,\alpha}) = \alpha$$

De tabel bevat 3 variabelen: n, m en α . Om de omvang enigszins beperkt te houden volstaat het wegens de bovenstaande symmetrie-eigenschap enkel het bereik $\alpha \geq \frac{1}{2}$ te tabelleren. Er

geldt immers dat: $F_{m,n,\alpha} = \frac{1}{F_{n,m,1-\alpha}}$.

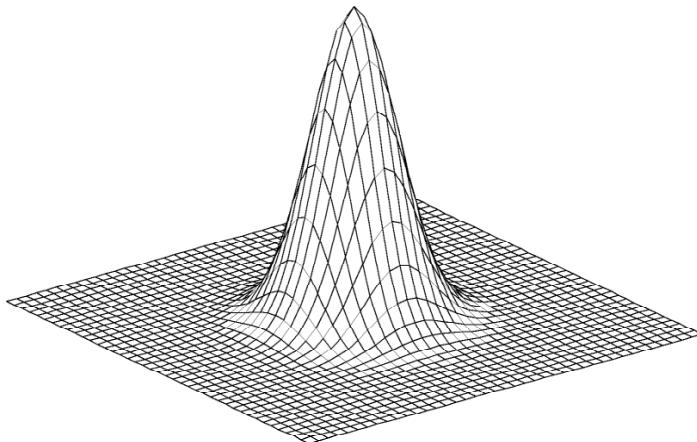
5.3 De meerdimensionale normale verdeling

5.3.1 De meerdimensionale standaard normale verdeling

We noemen een kansvector $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$ standaard normaal verdeeld als de componenten Z_1, Z_2, \dots, Z_n onafhankelijk en standaard normaal verdeeld zijn. We zien onmiddellijk dat de dichtheidsfunctie van de n-dimensionale standaardverdeling gegeven wordt door de formule:

$$f_Z(z_1, z_2, \dots, z_n) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2} \sum_{i=1}^n z_i^2\right) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2} \langle z^T, z \rangle\right)$$

waarbij $\langle z^T, z \rangle$ het scalair product van de vector \mathbf{z} met zichzelf is.



Figuur 5.31 : De dichtheidsfunctie van de tweedimensionale standaard normale verdeling

5.3.2 De meerdimensionale normale verdeling

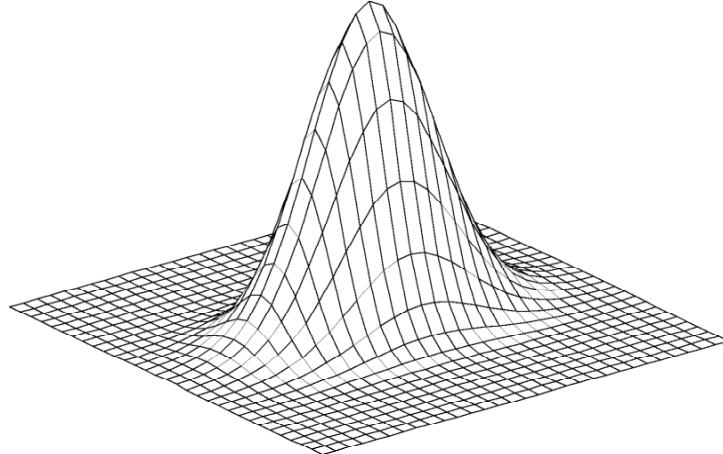
We zeggen dat de kansvector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ normaal verdeeld is als er een n-dimensionale vector \mathbf{m} bestaat en een reguliere nxn-matrix \mathbf{A} zodat $\mathbf{Z} = \mathbf{A}^{-1}(\mathbf{X}-\mathbf{m})$ standaard normaal verdeeld is.

De dichtheidsfunctie ziet er dan als volgt uit :

$$\begin{aligned} f_X(x_1, x_2, \dots, x_n) &= \frac{1}{|\det(A)|\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}(x-m)^T A^{-1}(x-m)\right) \\ &= \sqrt{\frac{\det(B)}{(2\pi)^n}} \exp\left(-\frac{1}{2}(x-m)^T B(x-m)\right) \end{aligned}$$

met $B = (A^{-1})^T A^{-1}$

Er geldt dat $B^{-1} = AA^T$ de covariantiematrix van de kansvector \mathbf{X} is.



Figuur 5.32 : De dichtheidsfunctie van de tweedimensionale normale verdeling met $\sigma_1 = 1$ en $\sigma_2 = \sqrt{3}$

5.4 De centrale limietstelling

5.4.1 Inleiding

In vele gevallen, maar vooral bij het nemen van steekproeven (zie later), zijn we geïnteresseerd in de verdeling Y_n van een som van n onafhankelijke stochastische veranderlijken X_i . De centrale limietstelling geeft hierop een antwoord. Laat ons om de gedachten te vestigen de som van Bernoulli experimenten behandelen. We zagen reeds dat als X_1, X_2, \dots, X_n onafhankelijke Bernoulli verdeelde experimenten zijn met parameter p , dat hun som $Y_n = X_1 + X_2 + \dots + X_n$ binomiaal verdeeld is ($Y_n \sim B(n; p)$) met gemiddelde np en spreiding $\sqrt{np(1-p)}$. De genormaliseerde verdeling $Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}}$ heeft dus gemiddelde 0 en

Figuur 5.33 : Staafdiagrammen van Z_n voor $n=10, 20, 40$ en 80

spreiding 1. Als we het staafdiagram van Z_n voor verschillende waarden van n tekenen (zie Figuur 5.33) dan zien we convergentie naar een mooie symmetrische klokvormige kromme voor $n \rightarrow \infty$ en dit ondanks de asymmetrie van X_i .

Dit laatste was al opgemerkt door de Moivre in 1718, die bewees dat de limietverdeling $n \rightarrow \infty$ de vorm heeft van de functie $\exp\left(-\frac{1}{2}x^2\right)$, de dichtheidsfunctie van de standaard normale verdeling.

Laplace (1812) liet zien dat deze limieteigenschap voor een veel grotere klasse verdelingen geldt. Y_n neigt naar een standaard normale verdeling bij de optelling van een brede waaier van verdelingen. Deze eigenschap heeft de naam centrale limietstelling gekregen. We formuleren ze nu.

5.4.2 De stelling en haar gevolgen

Stelling: De centrale limietstelling

Als X_1, X_2, \dots, X_n onafhankelijke stochastische variabelen zijn die willekeurige kansverdelingen bezitten waarvoor verwachtingswaarden μ_j en varianties σ_j^2 bestaan en uniform begrensd zijn: $|\mu_j| \leq M$ en $\sigma_j \leq V \quad \forall j \in N$

Als Y_n hun som is met verwachtingswaarde $\bar{\mu}_n$ en variantie $\bar{\sigma}_n^2$:

$$Y_n = \sum_{j=1}^n X_j, \quad \bar{\mu}_n = \sum_{j=1}^n \mu_j, \quad \bar{\sigma}_n^2 = \sum_{j=1}^n \sigma_j^2; \quad ;$$

dan convergeert $Z_n = \frac{Y_n - \bar{\mu}_n}{\bar{\sigma}_n}$ naar een standaard normaal verdeelde variabele:

$$\lim_{n \rightarrow \infty} Z_n = \lim_{n \rightarrow \infty} \frac{Y_n - \bar{\mu}_n}{\bar{\sigma}_n} = W \quad \text{met } W \sim \mathcal{N}(0;1)$$

Meestal zullen we in de praktijk gebruik maken van de hieronder vermelde verzwakte vorm van deze stelling.

Gevolg

Als X_1, X_2, \dots, X_n n onafhankelijke, identiek verdeelde stochastische variabelen zijn, met verwachtingswaarde μ en variantie σ^2 , dan zal naarmate n groter wordt de verdeling van de stochastische variabele $Y_n = \sum_{j=1}^n X_j$ steeds beter benaderd worden door de normale verdeling met gemiddelde $n\mu$ en variantie $n\sigma^2$.

Intuïtief

Gegevens die beïnvloed worden door vele kleine, niet gerelateerde, willekeurige effecten, worden bij benadering beschreven door de normale verdeling.

Toepassing: Benadering van de binomiale verdeling

a. De normale benadering ($n \geq 30$, $np \geq 5$ en $nq \geq 5$)

We kunnen deze stelling gebruiken om kansen van de binomiale verdeling voor grote n te benaderen met de normale verdeling. In de praktijk blijkt dit meestal reeds voor $n \geq 30$ een

goede benadering te geven, mits p en $q=1-p$ niet te klein zijn ($np \geq 5$ en $nq \geq 5$). Er geldt dat $B(n,p) \sim \mathcal{N}(np; np(1-p))$

Omdat $B(n,p)$ een discrete verdeling is en $\mathcal{N}(np; np(1-p))$ een continue, is het niet onmiddellijk duidelijk hoe we in een concreet geval de benadering zullen moeten uitrekenen. We kunnen bijvoorbeeld voor $X \sim B(36,0.2)$ de benadering $Y \sim \mathcal{N}(7.2; (2.4)^2)$ gebruiken. Als we echter de complementaire kansen

$$P(X \leq 6) = 0.4007 \text{ en } P(X \geq 7) = P(X > 6) = 0.5993$$

benaderen met de kansen

$$P(Y \leq 6) = \Phi\left(\frac{6 - 7.2}{2.4}\right) = \Phi(-0.5) = 0.3085$$

$$P(Y \geq 7) = \Phi\left(-\frac{7 - 7.2}{2.4}\right) = \Phi(0.0833) = 0.5332$$

hebben we een grote fout gemaakt. De som van beide benaderingen is niet gelijk aan één!

Voor de discrete binomiale verdeling is de kans $P(6 < X < 7)$ gelijk aan nul, maar voor de continue benadering is de kans $P(6 < Y < 7) = 0.1582$ en dus niet nul. We kunnen dit probleem oplossen door het bewuste interval $[6,7]$ eerlijk te verdelen tussen beide zijden en dus door de volgende benaderingen te gebruiken:

$$P(X \leq 6) \approx P(Y \leq 6.5) = \Phi\left(\frac{6.5 - 7.2}{2.4}\right) = 0.3853$$

$$P(X \geq 7) \approx P(Y \geq 6.5) = \Phi\left(-\frac{6.5 - 7.2}{2.4}\right) = 0.6147$$

We noemen dit de **continuïteitscorrectie**.

Analoog heeft het geen zin om de kans $P(X=6)=0.1543$ te benaderen met de kans $P(Y=6)$, omdat de kans op een gegeven uitkomst bij een continue verdeling altijd nul is. Voor een correcte benadering zullen we de discrete kans op $X=6$ moeten benaderen met een continue kans voor Y op een interval rond de waarde 6. Omdat X discreet is en alleen de waarden ..., 5, 6, 7, ... kan aannemen, ligt het opnieuw voor de hand om de intervallen $[5,6]$ en $[6,7]$ eerlijk te verdelen en $P(X=6)$ te benaderen met

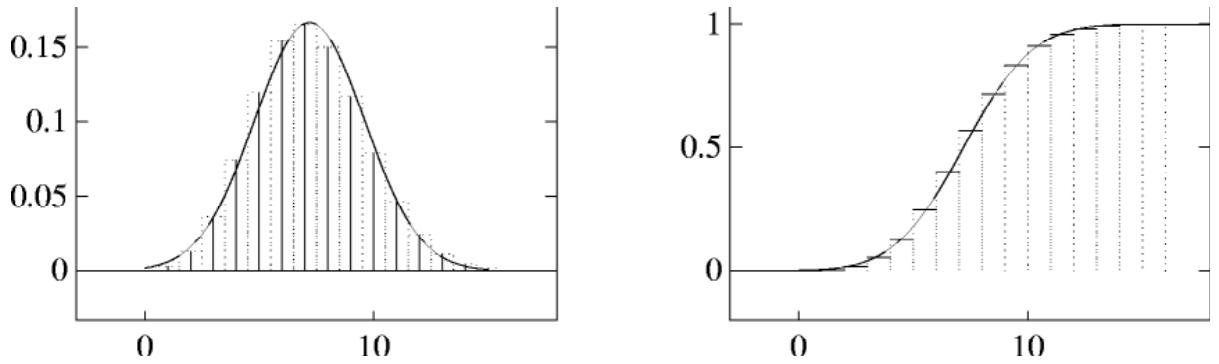
$$P(5.5 < Y < 6.5) = \Phi\left(\frac{6.5 - 7.2}{2.4}\right) - \Phi\left(\frac{5.5 - 7.2}{2.4}\right) = 0.3853 - 0.2394 = 0.1459$$

In het algemeen moeten we dus voor $X \sim B(n,p)$ en $Y \sim \mathcal{N}(np; np(1-p))$ de volgende benadering gebruiken:

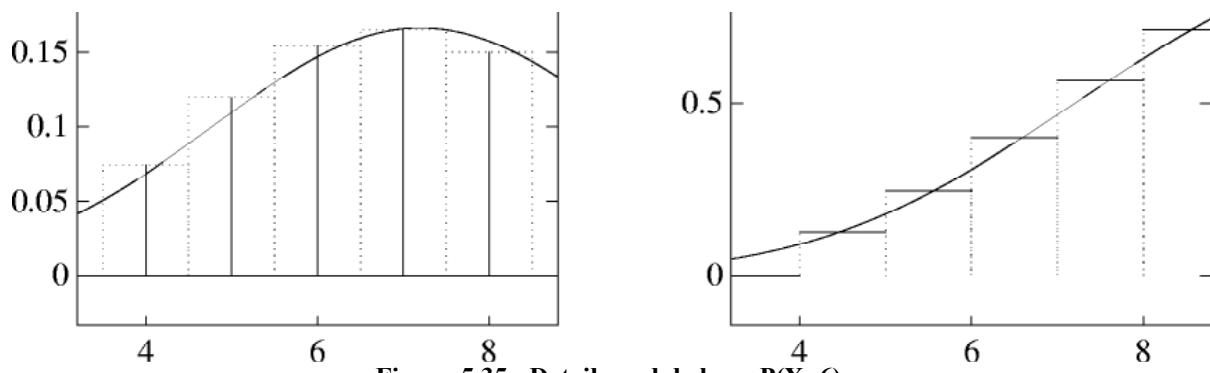
$$P(j \leq X \leq k) \approx P(j-0.5 < Y < k+0.5)$$

$$P(X \leq 0) \approx P(Y \leq 0.5) \text{ en } P(X \geq n) \approx P(Y \geq n-0.5)$$

Dit wordt nog eens duidelijk geïllustreerd door de onderstaande figuren.



Figuur 5.34 : Staafdiagram van $B(36; 0.2)$ en kansdichtheid van de benaderende $N(7.2; (2.4)^2)$ en hun kansverdelingen.



Figuur 5.35 : Detail rond de kans $P(X=6)$.

We zien op de grafieken van $X \sim B(36, 0.2)$ en $Y \sim \mathcal{N}(7.2; (2.4)^2)$ dat $P(X=6)$, de oppervlakte van de gestippelde rechthoek rond de staaf op $x=6$, goed benaderd wordt door de oppervlakte onder de continue kromme tussen 5.5 en 6.5.

(Uit de rechterfiguur zien we, dat we de continue verdeling 0.5 naar links moeten schuiven om in de gehele punten 0, 1, 2, ... een goede benadering te krijgen van $P(X \leq k)$.)

De benadering is natuurlijk niet perfect, maar zij wordt beter naarmate n groter wordt en zij heeft de eigenschap dat de som van de benaderingen van complementaire kansen steeds gelijk is aan één.

b. De Poisson benadering van de binomiale verdeling ($n \geq 30$, $np < 5$ of $nq < 5$)

Bij de benadering van $B(n, p)$ moeten we er wel op letten dat $B(n, \lambda/n)$ naar de Poissonverdeling $\mathcal{P}(\lambda)$ convergeert voor $n \rightarrow \infty$ en dus dat de benadering slechter wordt naarmate het produkt np (of het complement $n(1-p)$) kleiner wordt. Als np te klein is, is een benadering van $B(n, p)$ door $\mathcal{P}(np)$ beter.

Als vuistregel nemen we, dat we $B(n, p)$ voor $n \geq 30$ en $np < 5$ voldoende goed kunnen benaderen door $\mathcal{P}(np)$:

als $n \geq 30$, $X \sim B(n, p)$, $Y \sim \mathcal{P}(np)$ en $Z \sim \mathcal{P}(n(1-p))$ dan

$$np < 5 \rightarrow P(X=k) \approx P(Y=k)$$

$$nq < 5 \rightarrow P(X=n-k) \approx P(Z=k)$$

Omdat de Poissonverdeling eveneens discreet is, hebben we hierbij natuurlijk geen problemen met de bovenvermelde continuïteitscorrectie.

Toepassing: Benadering van de Poisson verdeling

Op grond van de centrale limietstelling en de eigenschap $\mathcal{P}(\lambda+\mu) = \mathcal{P}(\lambda) + \mathcal{P}(\mu)$ voor onafhankelijke Poisson verdelingen, weten we dat de Poisson verdeling zelf voor grote waarden van λ naar de normale verdeling convergeert. Als $X_\lambda \sim \mathcal{P}(\lambda)$, dan geldt $E[X_\lambda] = \lambda$ en $\text{Var}[X_\lambda] = \lambda$ zodat:

$$\lim_{\lambda \rightarrow \infty} \frac{X_\lambda - \lambda}{\sqrt{\lambda}} \sim \mathcal{N}(0;1) \text{ ofwel } \mathcal{P}(\lambda) \approx \mathcal{N}(\lambda; \lambda) \text{ als } \lambda \text{ voldoende groot.}$$

Als vuistregel nemen we opnieuw dat $\mathcal{P}(\lambda)$ op te zoeken is in een tabel voor $\lambda \leq 30$; anders is ze te benaderen met bovenstaande formule.

Aangezien de Poissonverdeling discreet is en de normale continu, moeten we ook hier aan de continuïteitscorrectie denken, dus:

$$P(X_\lambda \leq k) \approx \Phi\left(\frac{k + 0.5 - \lambda}{\sqrt{\lambda}}\right)$$

$$P(X_\lambda = k) \approx \Phi\left(\frac{k + 0.5 - \lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{k - 0.5 - \lambda}{\sqrt{\lambda}}\right)$$

Voorbeeld

Een maandblad schreef dat 6% van de Amerikaanse chauffeurs lezen terwijl ze autorijden. Indien je willekeurig 300 chauffeurs kiest, hoe groot is dan de kans dat er exact 25 hiervan lezen terwijl ze rijden ?

Zij X het aantal lezende chauffeurs $\rightarrow X \sim B(300; 0.06)$

$p = 0.06$ en dus $q = 0.94$ en $n = 300$

- Kan de normale benadering hier gebruikt worden ? Ja want:

$$\begin{aligned} np &= 18 \text{ en } nq = 282 \text{ zijn beiden groter dan 5} \\ n &\text{ is groter dan 30} \end{aligned}$$

- $\mu = np = 18$ en $\sigma = \sqrt{npq} = \sqrt{16.92} = 4.11$

- We kunnen X benaderen door $Y \sim \mathcal{N}(18; 16.92)$

- We zijn op zoek naar $P(X=25)$ of na continuïteitscorrectie $P(24.5 < Y < 25.5)$ en na z transformatie

$$P\left(\frac{24.5 - 18}{4.11} < Z < \frac{25.5 - 18}{4.11}\right) = P(1.58 < Z < 1.82) = \Phi(1.82) - \Phi(1.58) = 0.9656 - 0.9429 = 0.0227$$

NB: Bij de benaderingen hierboven hebben we als grens voor toepassing steeds $n \geq 30$ genomen. In vele gevallen zal de benadering bij $n \geq 20$ reeds zeer goede resultaten geven. Je zult merken dat in verschillende tekstboeken verschillende grenzen worden gehanteerd.

Hoofstuk 6: Verklarende statistiek

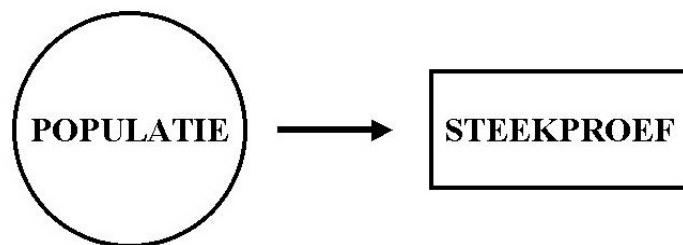
"In solving a problem of this sort, the grand thing is to be able to reason backwards. That is a very useful accomplishment, and a very easy one, but people do not practice it much... Most people, if you describe a train of events to them, will tell you what the result would be. They can put those events together in their minds, and argue from them that something will come to pass. There are few people, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were, which led to that result. This power is what I mean when I talk of reasoning backward." Sherlock Holmes

6.1 Inleiding

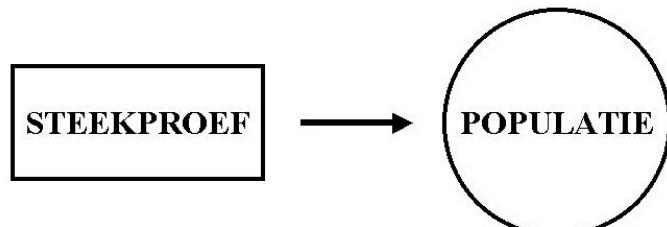
Vaak wil men van een karakteristiek weten welke waarde men kan verwachten, met andere woorden wat de gemiddelde uitkomst is in de populatie. Bijvoorbeeld, wat is de gemiddelde neerslag van calcium in België? Wat is het percentage aan zware metalen in de Belgische bodem? Wat is het gemiddeld gewicht van volwassen huismussen? Wat is de proportie van succesvolle proefboringen? Heeft een geneesmiddel een positief effect als er 12 van degenen, die het geneesmiddel gebruikten, verbetering tonen en maar 8 van degenen die het placebo gebruikten?

In de verklarende statistiek worden op basis van een beperkt aantal observaties, die afkomstig zijn uit een groter geheel (resultaten van een steekproef uit een populatie), uitspraken gedaan over (bepaalde karakteristieken van) het grotere geheel. Bovendien moet worden aangegeven "hoe betrouwbaar" die uitspraken zijn.

Als algemeen schema vertrekken we van een populatie waarvan we bepaalde eigenschappen willen te weten komen. Over deze populatie maken we modelveronderstellingen en dit levert de basis voor een model dat het gedrag van een steekproef beschrijft bv $X \sim B(1,p)$ of $X \sim \mathcal{N}(\mu, \sigma^2)$ of $X \sim \text{Uniform}(a,b)$.



Met deze steekproef gaan we nu grootheden construeren in functie van de vraag die we over de populatie hebben gesteld. Uit het gedrag van deze grootheden gaan we uiteindelijk een bewering over de populatie afleiden, samen met een uitspraak over de nauwkeurigheid van deze bewering.



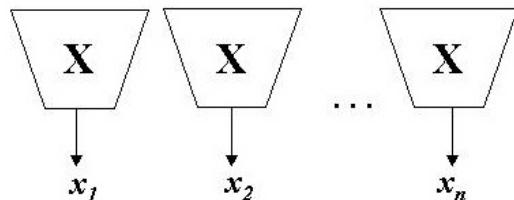
Meer concreet: een veel voorkomend probleem in de experimentele wetenschappen is het volgende. We willen de waarde van een grootheid X , b.v. de pH van een oplossing, door metingen bepalen. Hiertoe doen we een aantal onafhankelijke metingen x_1, x_2, \dots, x_n en we bepalen het steekproefgemiddelde \bar{x}_n en de standaardafwijking s_n . We vragen ons nu af: "Wat betekenen deze getallen?" en "Hoe betrouwbaar is het gemiddelde \bar{x}_n als benadering van de echte waarde?"

5.68	5.32	5.16	5.30	5.36
5.59	5.07	5.03	5.15	5.46
5.23	5.23	5.52	5.17	5.51
5.52	5.39	5.20	5.34	5.84

**Voorbeeld van 20 onafhankelijke metingen van de pH van een oplossing
met steekproefgemiddelde $\bar{x}_n = 5.35$ en standaardafwijking $s_n = 0.21$**

Hiertoe moeten we een aanname maken over de meetgegevens x_i . Zij moeten onafhankelijke trekkingen zijn uit een stochastische variabele X met verwachtingswaarde μ en spreiding σ . μ is dus de exacte waarde van de te schatten pH. Een meting is de uitkomst van een "kansspel", waarvan de uitkomst een zekere kansverdeling heeft.

Onze n experimenten vormen een steekproef $\{X_1, X_2, \dots, X_n\}$ van n onafhankelijke stochastieken, allen met dezelfde verdeling. De metingen $\{x_1, x_2, \dots, x_n\}$ vormen een trekking hieruit (of realisatie; X_i is het meetproces en x_i het toevallige resultaat, de meting).



In het vervolg zullen we steeds de stochastiek aanduiden met een hoofdletter en de verkregen getalwaarde of realisatie met een kleine letter. Als x_i een trekking is uit X_i , dan is het gemiddelde \bar{x}_n kennelijk een trekking uit $\bar{X}_n := (X_1 + \dots + X_n)/n$ en zal een uitspraak over de betrouwbaarheid van \bar{x}_n afhangen van de kansverdeling van \bar{X}_n , die een functie is van de steekproef. We zullen dit in één van de volgende paragrafen verder bespreken. Eerst formuleren we nog de definities van statistiek, schatter en zuivere schatter.

6.2 Statistieken en schatters

Een statistiek is een stochastische variabele die alleen een functie is van de steekproef en niet van onbekende parameters (zoals μ en σ).

Voorbeeld

De grootheden $\sum_{i=1}^n X_i$ en $\sum_{i=1}^n (X_i - \bar{X}_n)^2$ zijn statistieken, maar $\sum_{i=1}^n (X_i - \mu)^2$ niet.

Definitie

Een schatter is een statistiek die gebruikt wordt om een onbekende parameter te benaderen. Een schatting is de getalwaarde van de schatter in een concreet experiment.

Voorbeeld

Het steekproefgemiddelde $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$ is een **schatter** voor μ , immers $E[\bar{X}_n] = \mu$. De

getalwaarde $\bar{x}_n = 5.35$ is een **schatting** voor de exacte waarde μ van de pH in bovengenoemd experiment. Omdat \bar{x}_n een getal is, spreekt men ook wel over een **puntschatting**.

Voorbeeld

Ook $W := \frac{2}{n(n+1)} \sum_{k=1}^n kX_k$ is een schatter voor μ (maar minder goed). De bijbehorende schatting van de pH levert de getalwaarde 5.37 op.

Definitie

Een schatter T van een parameter λ heet **zuiver** (Eng. unbiased), als $E[T] = \lambda$, dus als de verwachtingswaarde van de schatter gelijk is aan de (gezochte) parameter.

Gevolg

Het steekproefgemiddelde \bar{X}_n en het gewogen gemiddelde $W = \frac{2}{n(n+1)} \sum_{k=1}^n kX_k$, zijn beide zuivere schatters van μ (bewijs zelf). De eerste is beter omdat de tweede een grotere variantie heeft:

$$Var[\bar{X}_n] = \sum_{i=1}^n Var\left[\frac{X_i}{n}\right] = \frac{\sigma^2}{n} \quad \text{en} \quad Var[W] > \frac{\sigma^2}{n}$$

Voorbeeld

Als $X \sim B(1,p)$ Bernoulli verdeeld is (met uitkomsten 0 en 1), ligt het voor de hand om de onbekende fractie p te schatten met de relatieve frequentie van het aantal waarnemingen $x_i = 1$, ($i=1 \dots n$). We gebruiken dus als schatter:

$$F = \#\{X_i = 1 | i=1 \dots n\} / n = \frac{1}{n} \sum_{i=1}^n X_i.$$

We weten dat $Y := X_1 + \dots + X_n \sim B(n,p)$ en $E[Y] = np$. De schatter F heeft dus verwachtingswaarde $E[F] = p$ en is dus zuiver en we vinden zo de schatting $\hat{p} = (x_1 + \dots + x_n) / n$

Op grond van onze kennis van de binomiale verdeling kunnen we dan betrouwbaarheids-grenzen berekenen voor de afwijking tussen de schatting \hat{p} en de (onbekende) parameter p (zie verder).

Analoog kunnen we zo voor een discrete verdeling steeds de relatieve frequentie van een waarde gebruiken als schatting voor de kans op die waarde.

Voorbeeld. Let op !

Als X uniform verdeeld is op $[0, b]$, dan kunnen we de verwachtingswaarde \hat{b} van b schatten m.b.v. het eerste empirische moment $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ en dus zou $\hat{b} := 2\bar{x}$ een goede schatting zijn van b .

Het is gemakkelijk in te zien dat de bijbehorende schatter zuiver is. Deze schatter geeft echter het volgende probleem. Als $\{1, 2, 9\}$ drie waarnemingen zijn van $X \sim \text{Uniform}(0, b)$, dan verwachten we $b > 9$ op grond van de derde waarneming; de schatting \hat{b} heeft echter de te kleine waarde $\hat{b} = 2/3(1+2+9)=8$.

Een schatter die wel rekening houdt met de grootste waarneming, vinden we via de “ordestatistiek” van X . De verdelingsfunctie van X is $P(X \leq x) = F_X(x) = x/b$ als $0 \leq x \leq b$. Voor de grootste waarneming in $\{X_1, \dots, X_n\}$ geldt dus wegens de onderlinge onafhankelijkheid:

$$P\left(\max_i X_i \leq x\right) = P(X_1 \leq x \& X_2 \leq x \& \dots \& X_n \leq x) = \left(\frac{x}{b}\right)^n \quad \text{als } 0 \leq x \leq b$$

zodat

$$E[\max\{X_1, \dots, X_n\}] = \int_0^b xn\left(\frac{x}{b}\right)^{n-1} \frac{1}{b} dx = \frac{nb}{n+1} \quad \text{en} \quad \hat{b} := \frac{n+1}{n} \max\{x_1, \dots, x_n\}$$

Uit het eerste moment van $\max\{X_1, \dots, X_n\}$ vinden we dus wel een bevredigende schatter voor b .

Bepaling van schatters

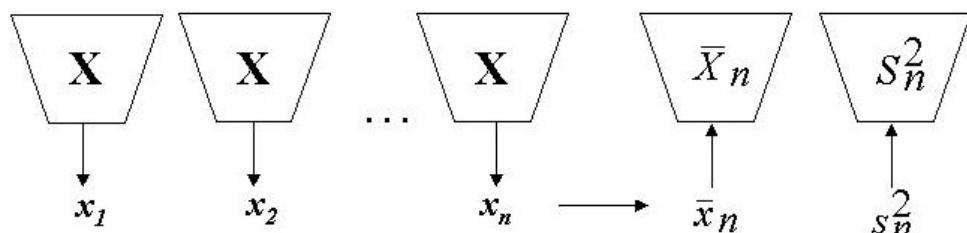
Er bestaan meerdere methoden voor het bepalen van schatters.

- momenten methode: het schatten van momenten van een verdeling om daaruit de onbekende parameters te schatten (denk maar aan de schatting van \hat{p} en \hat{b} in de bovenstaande voorbeelden.)
- methode van de maximum kans (“maximum likelihood method”)

We gaan er hier niet dieper op in.

6.3 Het schatten van de verwachtingswaarde en de variantie

In wat volgt zullen we trachten schatters te vinden voor de verwachtingswaarde en de variantie van een verdeling. We vermeldden reeds dat als x_i een trekking is uit X_i , het gemiddelde \bar{x}_n een trekking is uit $\bar{X}_n := (X_1 + \dots + X_n)/n$ en een uitspraak over de betrouwbaarheid van \bar{x}_n zal dus afhangen van de kansverdeling van \bar{X}_n , die een functie is van de steekproef. Een analoge redenering zal gevuld worden voor de steekproefvariantie.



Alvorens over te gaan tot uitspraken over betrouwbaarheid behandelen we enkele eigenschappen van het steekproefgemiddelde \bar{X}_n , van de steekproefvariantie en van het verschil van twee steekproefgemiddelden.

6.3.1 Het steekproefgemiddelde

a) Eigenschappen.

- De verwachtingswaarde van het steekproefgemiddelde is het populatiegemiddelde (We bewezen immers dat het steekproefgemiddelde een zuivere schatter voor het populatiegemiddelde is.)

$$E[\bar{X}_n] = \mu$$

- De variantie van het steekproefgemiddelde is gelijk aan de populatievariantie gedeeld door de steekproefgrootte.

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

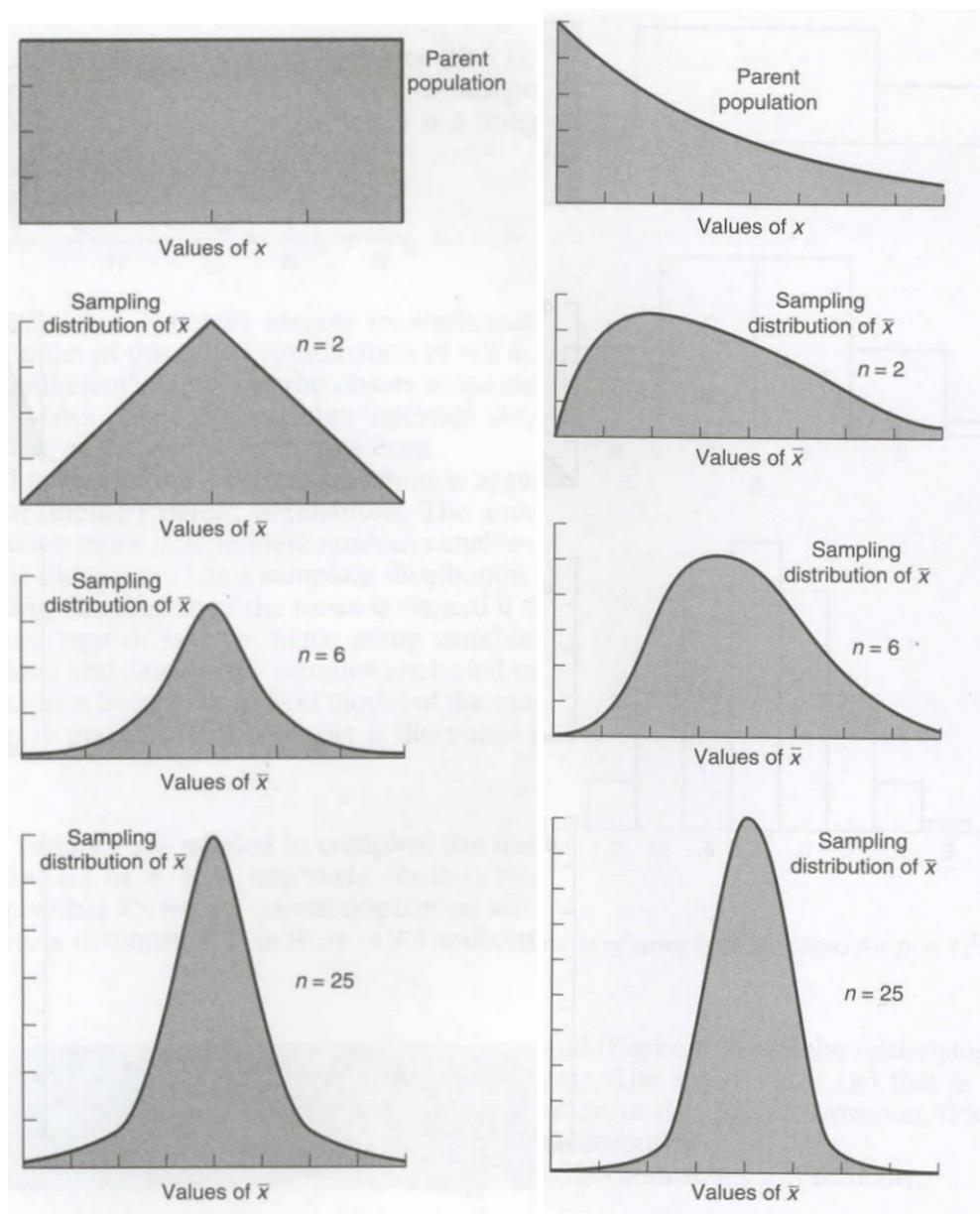
- De **standaardfout** van het steekproefgemiddelde is dus $\frac{\sigma}{\sqrt{n}}$. Als σ niet gekend is, wordt de standaardfout geschat door $\frac{s_n}{\sqrt{n}}$.

N.B.: Bemerk dat we hier de term standaardfout gebruikten; we hadden het hier immers over de standaardafwijking van het steekproefgemiddelde en niet de standaardafwijking van de populatie. De standaardfout geeft bij herhaalde metingen een goed idee over de experimentele fout (cfr voorbeeld met pH metingen). Men zegt dikwijls dat de fout bij het n maal herhalen van een meting daalt met een factor \sqrt{n} .

b) De verdeling

De centrale limietstelling vertelt ons dat voor grote n ($n > 30$) $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ en dit ongeacht de verdeling van X . Deze verdeling concentreert zich met groeiende n steeds meer rond μ .

Op de volgende figuren wordt dit duidelijk geïllustreerd indien X verdeeld is volgens de uniforme en de Poisson verdeling. Merk dat voor groter wordende n de verdeling van \bar{X}_n inderdaad de klokvorm benadert en de spreiding kleiner wordt.



6.3.2 De steekproefvariantie

a) Eigenschap

De steekproefvariantie $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is een zuivere schatter van de populatievariantie σ^2 .

$$\begin{aligned}
 \text{Bewijs: } S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n ((X_i - \mu) - (\bar{X}_n - \mu))^2 && \text{(dubbel produkt valt NIET weg)} \\
 &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 - \sum_{i=1}^n (\bar{X}_n - \mu)^2 \right] \\
 &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - E[X_i])^2 - n(\bar{X}_n - E[\bar{X}_n])^2 \right]
 \end{aligned}$$

$$\begin{aligned}
 \text{dus: } E[S_n^2] &= \frac{1}{n-1} \left[\sum_{i=1}^n E[(X_i - E[X_i])^2] - nE[(\bar{X}_n - E[\bar{X}_n])^2] \right] \\
 &= \frac{1}{n-1} \left[\sum_{i=1}^n Var(X_i) - nVar(\bar{X}_n) \right] \\
 &= \frac{1}{n-1} \left[n\sigma^2 - n\frac{\sigma^2}{n} \right] = \sigma^2
 \end{aligned}$$

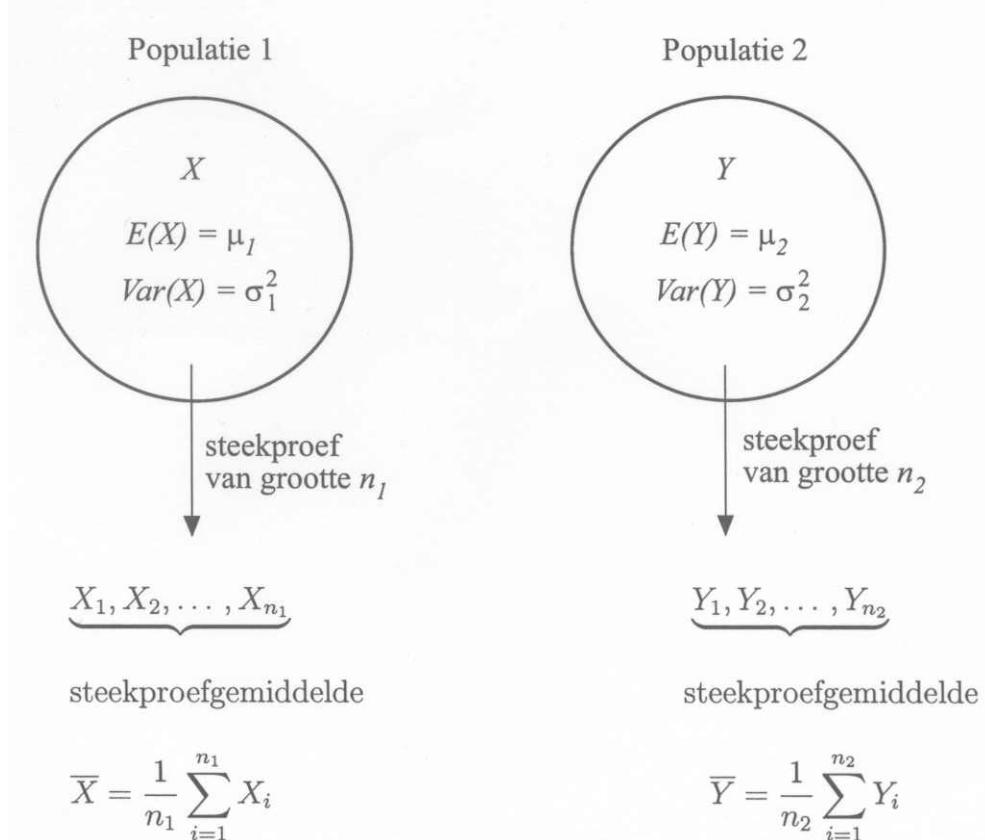
b) De verdeling

Indien de populatie normaal verdeeld is ($X \sim \mathcal{N}(\mu, \sigma^2)$), dan is de stochastiek $\frac{(n-1)S_n^2}{\sigma^2}$ chikwadraat verdeeld met $n-1$ vrijheidsgraden: $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$.

Bovendien zijn het steekproefgemiddelde en de steekproefvariantie onafhankelijke statistieken.

6.3.3 Het verschil van twee steekproefgemiddelden.

Om het gemiddelde van twee populaties met elkaar te vergelijken, zullen we ons baseren op het verschil tussen de gemiddelden van steekproeven uit deze populaties.



N.B.: Voor het steekproefgemiddelde wordt in sommige figuren de notatie zonder index n gebruikt.

Enkele eigenschappen van het verschil van steekproefgemiddelden ($\bar{Y}_{n_2} - \bar{X}_{n_1}$) worden in deze paragraaf bestudeerd. Het is duidelijk dat dit verschil ook een stochastische veranderlike is.

a) De verwachtingswaarde en de variantie

Voor de verwachtingswaarde en de variantie van $\bar{Y}_{n_2} - \bar{X}_{n_1}$ geldt:

$$E[\bar{Y}_{n_2} - \bar{X}_{n_1}] = \mu_2 - \mu_1$$

$$Var(\bar{Y}_{n_2} - \bar{X}_{n_1}) = Var(\bar{Y}_{n_2}) + Var(\bar{X}_{n_1}) = \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}$$

b) De verdeling

Wanneer de onderliggende populaties normaal verdeeld zijn ;
 $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ en $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$

dan geldt: $\bar{Y}_{n_2} - \bar{X}_{n_1} \sim \mathcal{N}\left(\mu_2 - \mu_1; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

ofwel: $\frac{(\bar{Y}_{n_2} - \bar{X}_{n_1}) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$

Wanneer de steekproeven groot zijn (>30) dan mag je bovenstaande eigenschap ook gebruiken, zelfs als de onderliggende populaties niet normaal verdeeld zijn.

c) De standaardfout

De standaardfout van $\bar{Y}_{n_2} - \bar{X}_{n_1}$ is dus gelijk aan $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Als de populatievarianties niet gekend zijn, dan schat men de standaardfout door $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ waarbij s_1^2 en s_2^2 de waarden van de steekproefvarianties zijn.

6.4 Het schatten van parameters en betrouwbaarheidsintervallen

Nu beschikken we over alle informatie om de parameters μ en σ te schatten en een betrouwbaarheidsmarge te geven van de schatting. We zullen aannemen, dat X normaal verdeeld is: $X \sim \mathcal{N}(\mu, \sigma^2)$; zoniet zal dit expliciet vermeld worden. Bedenk dat deze veronderstelling niet zo cruciaal is indien onze steekproefgrootte groter dan 30 is.

We gaan de volgende gevallen bestuderen:

- i) Het schatten van μ indien σ bekend is
- ii) Het schatten van μ indien σ niet bekend is
- iii) Het schatten van σ

Nadien (in punt iv) zullen we het schatten van een populatieproportie behandelen, een geval waarbij X niet normaal verdeeld is.

Gaandeweg zullen we de begrippen zoals onzekerheidsniveau en betrouwbaarheidsinterval definiëren.

6.4.1 Het schatten van μ indien σ bekend is

Laat ons de werkwijze en de begrippen die aan bod komen illustreren aan de hand van een voorbeeld.

Voorbeeld

500 willekeurig gekozen Belgische eindejaarsstudenten leggen een examen af. Zij behalen gemiddeld $\bar{x}_n = 461$. Wat kun je zeggen over μ , de gemiddelde score van alle Belgische eindejaarsstudenten? Je weet dat $\sigma = 100$.

We weten dat $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Dit betekent voor de score dat $\bar{X}_n \sim \mathcal{N}(\mu, 20)$. Dus:

$$P\left(\frac{|\bar{X}_n - \mu|}{\sqrt{20}} \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha \text{ voor elke } \alpha.$$

Kies nu α gelijk aan 0.05. Dan krijg je:

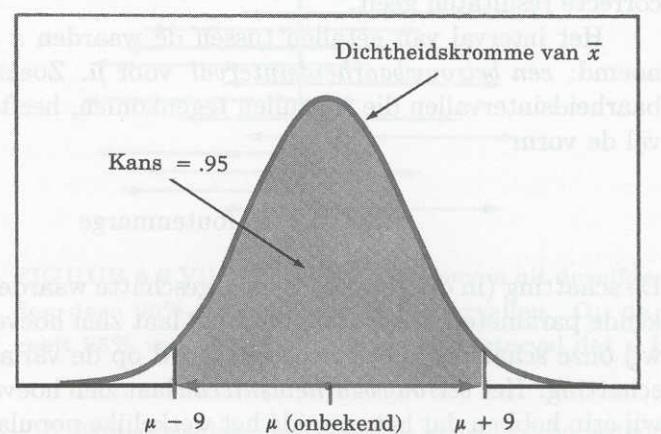
$$P\left(\frac{|\bar{X}_n - \mu|}{4.5} \leq 1.96\right) = 0.95$$

of $P(\bar{X}_n - 9 \leq \mu \leq \bar{X}_n + 9) = 0.95$

Het steekproefgemiddelde 461 gaat dus met een kans 95% binnen een afstand 9 van μ liggen: $P(\mu - 9 \leq 461 \leq \mu + 9) = 0.95$. Zeggen dat \bar{x}_n binnen een afstand van 9 punten van μ ligt, is hetzelfde als zeggen dat μ binnen een afstand van 9 punten van \bar{x}_n ligt. Dit betekent dat in 95% van alle steekproeven het interval $\bar{x}_n - 9$ tot $\bar{x}_n + 9$ de te schatten μ zal bevatten of dat μ met een kans van 95% in het interval [452,470] ligt. Dit interval wordt het betrouwbaarheidsinterval (BI) voor μ genoemd met een betrouwbaarheid van 95%.

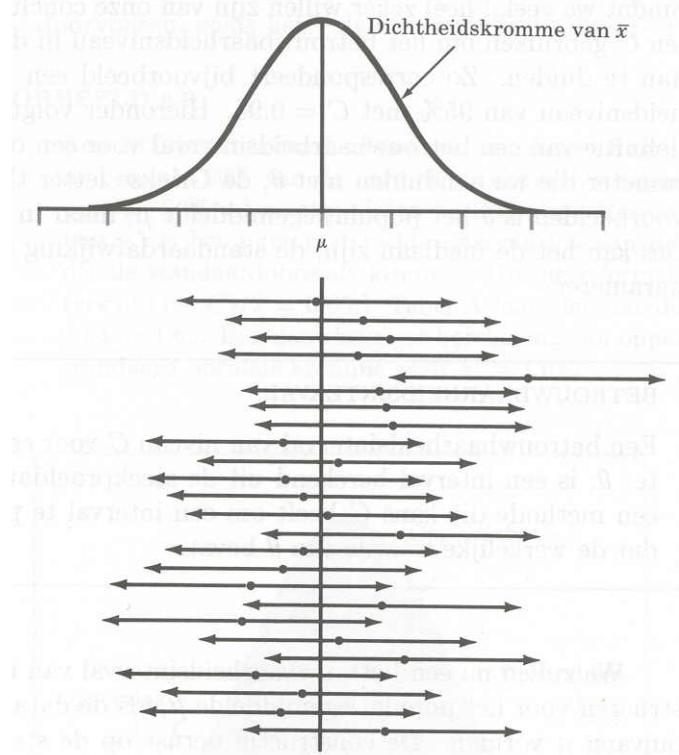
Het bestaat dus uit: [schatting – foutmarge, schatting + foutmarge]

Bovenstaande redenering wordt in onderstaande grafiek weergegeven.



De foutmarge vertelt ons iets over de accuraatheid van onze schatting. Het betrouwbaarheidsniveau 95% laat zien hoeveel vertrouwen wij erin hebben dat het interval μ bevat.

Bij een nieuwe steekproef zal \bar{x}_n een ander getal zijn en zal het interval wat verschoven worden. Bemerk dat de lengte van het interval dezelfde blijft. In onderstaande grafiek werd het resultaat van 25 steekproeven afgebeeld. Bemerk dat er 1 interval μ niet bevatt.



Definitie

Het 2-zijdig betrouwbaarheidsinterval (BI) voor μ op het onzekerheidsniveau α of het betrouwbaarheidsinterval met betrouwbaarheid $1-\alpha$ (confidence interval) is het interval:

$$\left[\bar{x}_n - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}} \right]$$

De meest gebruikte betrouwbaarheidsintervallen zijn:

- a) Het 99%-betrouwbaarheidsinterval : $\left[\bar{x}_n - 2.576 \frac{\sigma}{\sqrt{n}}, \bar{x}_n + 2.576 \frac{\sigma}{\sqrt{n}} \right]$
- b) Het 95%-betrouwbaarheidsinterval : $\left[\bar{x}_n - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x}_n + 1.96 \frac{\sigma}{\sqrt{n}} \right]$
- c) Het 90%-betrouwbaarheidsinterval : $\left[\bar{x}_n - 1.645 \frac{\sigma}{\sqrt{n}}, \bar{x}_n + 1.645 \frac{\sigma}{\sqrt{n}} \right]$

want $\Phi^{-1}(0.995) = 2.576$, $\Phi^{-1}(0.975) = 1.96$ en $\Phi^{-1}(0.95) = 1.645$

Gevolg

- 1) Indien je de betrouwbaarheid vergroot, d.w.z. α verkleint, dan zal het BI vergroten. Je wil immers met een grotere betrouwbaarheid zeggen dat μ in het BI ligt.
- 2) Indien je de steekproefgrootte vergroot zal het BI voor eenzelfde niveau α een kleinere lengte hebben.

We berekenen deze situaties in het volgende voorbeeld.

Voorbeeld

In een laboratorium worden monsters van een farmaceutisch produkt geanalyseerd om de concentratie van het actieve bestanddeel te bepalen. Vermits de chemische analyses niet heel nauwkeurig zijn worden meerdere metingen op hetzelfde monster herhaald. De resultaten van herhaald meten volgen vrij nauwkeurig een normale verdeling. De standaardafwijking van deze verdeling is een inherente eigenschap van het analytisch procédé en heeft de bekende waarde 0.0068 gram per liter. Drie analyses van hetzelfde monster leveren de concentraties:

$$0.8403 \quad 0.8363 \quad 0.8447$$

- i) Geef een 99% betrouwbaarheidsinterval voor de werkelijke concentratie.
- ii) Veronderstel dat één enkele meting 0.8404 geeft en je enkel over deze meting beschikt, geef dan een 99%-BI voor de werkelijke concentratie
- iii) Geef een 90%-BI voor de werkelijke concentratie gebaseerd op je eerste 3 analyses.

Oplossing

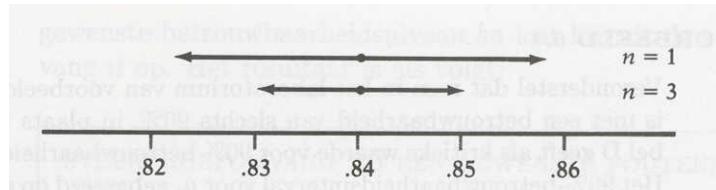
- i) De concentratie (X) is normaal verdeeld. $\sigma = 0.0068$ is gekend. De werkelijke concentratie μ is de te schatten parameter. Het steekproefgemiddelde $\bar{x}_3 = 0.8404$

Het 99%-BI is dus:

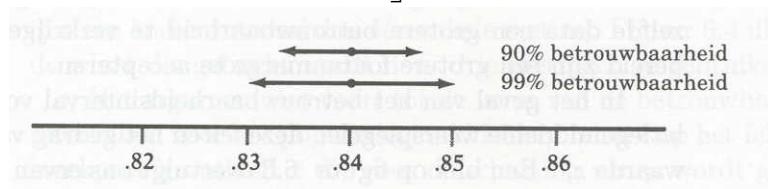
$$\left[\bar{x}_n - 2.576 \frac{\sigma}{\sqrt{n}}, \bar{x}_n + 2.576 \frac{\sigma}{\sqrt{n}} \right] = \left[0.8404 - 2.576 \frac{0.0068}{\sqrt{3}}, 0.8404 + 2.576 \frac{0.0068}{\sqrt{3}} \right] = [0.8303, 0.8505]$$

- ii) analoog: $n=1$ en $x_1 = \bar{x}_1 = 0.8404$

$$\left[0.8404 - 2.576 \frac{0.0068}{\sqrt{1}}, 0.8404 + 2.576 \frac{0.0068}{\sqrt{1}} \right] = [0.8229, 0.8579]$$



$$\text{iii) } \left[0.8404 - 1.645 \frac{0.0068}{\sqrt{3}}, 0.8404 + 1.645 \frac{0.0068}{\sqrt{3}} \right] = [0.8339, 0.8469]$$



Gevolg

Je kunt nu ook de steekproefgrootte berekenen die voor een bepaald onzekerheidsniveau een bepaalde foutenmarge geeft. De lengte van het BI bij een gegeven niveau α is:

$$2 \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \frac{\sigma}{\sqrt{n}} \text{ en de foutmarge } \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \frac{\sigma}{\sqrt{n}}$$

Om een foutmarge m te hebben moeten we dus een steekproefgrootte n hebben met:

$$n \geq \left(\Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \frac{\sigma}{m} \right)^2$$

Voorbeeld

Indien een nieuwe klant van het laboratorium (van het vorige voorbeeld) met een betrouwbaarheid van 95% resultaten wil die tot op 0.005 nauwkeurig zijn, hoeveel metingen zullen er dan moeten uitgevoerd worden op zijn monster?

Antwoord: $m=0.005$; $\alpha=0.05$; $\Phi^{-1}(0.975)=1.96$

$$\text{Het aantal uit te voeren metingen is dus: } n \geq \left(\frac{1.96 \cdot 0.0068}{0.005} \right)^2 = 7.1$$

7 metingen zijn dus te weinig, maar 8 metingen zullen zeker de gevraagde accuraatheid geven met een betrouwbaarheid van 95%.

6.4.2 Het schatten van μ indien σ niet bekend is

Het kader waarin we tot nu toe betrouwbaarheidsintervallen hebben opgesteld, steunt op de onderstelling dat je veel weet over de populatie; je moet namelijk niet enkel weten dat ze normaal verdeeld is, maar ook dient haar variantie σ^2 gekend te zijn. In de meeste toepassingen is dit behoorlijk onrealistisch. Zoals reeds gemeld is de onderstelling over het normaal verdeeld zijn van de populatie niet zo cruciaal van zodra we een redelijk grote steekproef hebben ($n>30$). σ^2 daarentegen is een karakteristiek van de populatie die we meestal niet kennen. Een oplossing hiervoor bestaat in het vervangen van deze populatieparameter door een “goede schatter”. Het invullen van de steekproefresultaten zal dan leiden tot een schatting waarvan we hopen dat het een goede benadering voor σ^2 is. We hebben reeds gezien dat de steekproefvariantie s_n^2 een onvertekende of zuivere schatter is van σ^2 . Het vervangen van σ^2 door s_n^2 leidt echter tot een nieuw model met een eigen verdeling.

$$\text{We weten dat: } Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \quad \text{en} \quad Y = \frac{(n-1)s_n^2}{\sigma^2} \sim \chi_{n-1}^2$$

Tevens zijn Y en Z onafhankelijk vanwege de onafhankelijkheid van het steekproefgemiddelde en de steekproefvariantie.

$$\text{Er geldt dus dat: } \frac{Z}{\sqrt{\frac{Y}{n-1}}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \frac{\sigma}{s_n} = \frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} = T \sim t_{n-1}$$

zodat we bij het opstellen van het BI in dit geval een Student t-verdeling met ($n-1$) vrijheidsgraden moeten gebruiken.

Analoog als hiervoor volgt nu:

$$P\left(\left|\frac{\bar{X}_n - \mu}{\sqrt{s_n^2/n}}\right| \leq t_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

$$\text{zodat: } P\left(\bar{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,1-\alpha/2} \leq \mu \leq \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

Het betrouwbaarheidsinterval voor μ met betrouwbaarheid $1-\alpha$ is het interval:

$$\left[\bar{x}_n - \frac{s_n}{\sqrt{n}} t_{n-1,1-\alpha/2}, \bar{x}_n + \frac{s_n}{\sqrt{n}} t_{n-1,1-\alpha/2} \right]$$

Het is duidelijk dat dit interval breder is dan het overeenkomstige interval voor μ in het geval dat σ bekend is. Het verschil neemt echter af met toenemende n en het verdwijnt geheel in de limiet $n \rightarrow \infty$.

In gevallen waarbij de steekproef voldoende groot is, kunnen we de t-verdeling benaderen door de standaard normale verdeling. Er geldt dat $\frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} \sim \mathcal{N}(0,1)$ voor grote n .

Een $(1-\alpha)$ -BI voor μ is dan:

$$\left[\bar{x}_n - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{s_n}{\sqrt{n}}, \bar{x}_n + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{s_n}{\sqrt{n}} \right]$$

Dit is hetzelfde resultaat als voor het geval waarin σ gekend is. In de formule is σ enkel vervangen door s_n .

Besluit

Bemerkt dat we in het bovenstaande telkens een betrouwbaarheidsinterval bekwamen gaande van het steekproefgemiddelde min een “aantal keer” de standaardfout van het steekproefgemiddelde ($\frac{\sigma}{\sqrt{n}}$) tot het steekproefgemiddelde plus een “aantal keer” de standaardfout van het steekproefgemiddelde

$$\Rightarrow \text{steekproefgemiddelde} \pm \text{“aantal keer” standaardfout}$$

Het “aantal keer” wordt berekend als het $1-\alpha/2$ kwantiel van de standaard normale verdeling indien de standaardfout van het steekproefgemiddelde ($\frac{\sigma}{\sqrt{n}}$) of dus de populatie variantie

$$(\sigma^2) \text{ gekend is} \quad \Rightarrow \bar{x} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}}$$

Zoniet wordt deze standaardafwijking geschat door $\frac{s_n}{\sqrt{n}}$ en wordt het “aantal keer” bepaald

$$\text{als het } 1-\alpha/2 \text{ kwantiel van de t verdeling met } n \text{ vrijheidsgraden.} \Rightarrow \bar{x} \pm t_{n-1,1-\alpha/2} \frac{s_n}{\sqrt{n}}$$

Voor grote steekproeven (grote n) convergeert de t-verdeling naar de standaard normale verdeling, zodat het gebruik van de kwantielen van de t-verdeling en van de standaard normale verdeling dezelfde resultaten geven.

6.4.3 Het schatten van σ

We vermeldden reeds dat als een populatie normaal verdeeld is ($X \sim \mathcal{N}(\mu, \sigma^2)$), de stochastiek $\frac{(n-1)s_n^2}{\sigma^2}$ chi-kwadraat verdeeld is met $n-1$ vrijheidsgraden: $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi_{n-1}^2$.

We zullen dit nu gebruiken om een $(1-\alpha)\%$ -BI voor σ te construeren.

Er geldt immers:

$$P\left(\chi_{n-1, \frac{\alpha}{2}}^2 \leq \frac{(n-1)s_n^2}{\sigma^2} \leq \chi_{n-1, 1-\frac{\alpha}{2}}^2\right) = 1 - \alpha$$

of: $P\left(\frac{(n-1)s_n^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \leq \sigma^2 \leq \frac{(n-1)s_n^2}{\chi_{n-1, \frac{\alpha}{2}}^2}\right) = 1 - \alpha$

Definitie

Het betrouwbaarheidsinterval voor σ^2 met betrouwbaarheid $1-\alpha$ is het interval:

$$\left[\frac{(n-1)s_n^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)s_n^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right]$$

Merk op dat dit BI niet symmetrisch is rond s_n^2 . Dit komt omdat de verdeling niet symmetrisch is.

Een $(1-\alpha)\%$ -BI voor σ zelf vinden we door de wortel te nemen:

$$\left[\sqrt{\frac{(n-1)s_n^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}}, \sqrt{\frac{(n-1)s_n^2}{\chi_{n-1, \frac{\alpha}{2}}^2}} \right]$$

Voorbeeld

De steekproefvariantie van de lengte van 16 willekeurig gekozen soldaten van de lichting 1992 is 51.8 cm^2 . We construeren een 90% -BI voor σ .

We weten: $n-1 = 15$; $\alpha = 0.1$; $\chi_{15, 0.05}^2 = 7.261$; $\chi_{15, 0.95}^2 = 24.996$

$$\text{BI} = \left[\sqrt{\frac{15 \cdot 51.8}{24.996}}, \sqrt{\frac{15 \cdot 51.8}{7.261}} \right] = [5.58, 10.34]$$

Voorbeeld

We voeren 12 onafhankelijke metingen uit van het gewicht van een bol en vinden de volgende waarden (in gram):

170 ; 183 ; 185 ; 175 ; 177 ; 173 ; 172 ; 181 ; 183 ; 177 ; 176 ; 180

Construeer een 90%-BI voor μ en σ .

We veronderstellen dat de metingen normaal verdeeld zijn.

$$\text{We hebben: } \bar{x}_{12} = \frac{2132}{12} = 177.67 \text{ g} \quad \text{en} \quad s_{12}^2 = \frac{1}{11} \sum_{i=1}^{12} (x_i - \bar{x}_{12})^2 = 22.788 \text{ g}^2$$

$$\text{Zodat: } s_{12} = 4.774 \text{ g} \quad \text{en} \quad \frac{S_{12}}{\sqrt{12}} t_{11,0.95} = \frac{4.774}{\sqrt{12}} 1.796 = 2.4751$$

Een 90%-BI voor μ is dus [175,180]

$$\text{Tevens: } \sqrt{\frac{11s_{12}^2}{\chi^2_{11,0.95}}} = \sqrt{\frac{11 \cdot 22.788}{19.675}} = 3.5694 \quad \text{en} \quad \sqrt{\frac{11s_{12}^2}{\chi^2_{11,0.05}}} = \sqrt{\frac{11 \cdot 22.788}{4.575}} = 7.4021$$

Een 90%-BI voor σ is dus [3.6,7.4]

6.4.4 Het schatten van een proportie van een populatie

Wie krijgt kanker ?

Veertig procent van alle rokers, twee procent van de arbeiders in chemische bedrijven, één procent van de mensen die bij een roker wonen, één per miljoen omwonenden van een kerncentrale,

Een grote risicofactor voor de algehele gezondheid is ook de geboorte: honderd procent van allen die geboren worden gaat dood.

Allan Smith

In deze paragraaf spitsen we ons toe op de situatie waarbij onze interesse in een populatie uitgaat naar het al dan niet bezitten van een specifieke eigenschap. Bij een eindige populatie zou men kunnen gaan tellen hoeveel individuen de eigenschap bezitten. Als men dit uitdrukt ten opzichte van de totale populatie bekomt men een getal (tussen 0 en 1) dat een karakteristiek van de populatie is (populatieparameter). Dat getal noemen we de **populatieproportie** en we stellen het voor door p . (Bedenk dat je, om consistent te zijn, dit getal door π zou moeten voorstellen. In sommige teksten zal dit ook het geval zijn.)

Wanneer we nu omgekeerd redeneren, dan zien we dat dit getal p ook de kans aangeeft dat een willekeurig getrokken individu uit deze populatie de eigenschap heeft. We hebben dus een binair kenmerk (is ziek / is niet ziek ; roker / niet roker ; man / vrouw ; ...). Als we afspreken dat het getal 0 overeenstemt met één van beide categorieën en het getal 1 met de andere, kunnen we onze populatie beschrijven door een stochastische variabele X die Bernoulli verdeeld is met parameter p ; $X \sim B(1,p)$ (vandaar de keuze voor p i.p.v. π).

Aan de hand van de steekproef $\{X_1, \dots, X_n\}$ wensen we p te schatten. Het is duidelijk dat de steekproefproportie \bar{P} een zuivere schatter voor p is. In feite is de stochastiek \bar{P} gelijk aan $\bar{X}_n := (X_1 + \dots + X_n)/n$. Tevens geldt dat $n\bar{X}_n = n\bar{P} \sim B(n,p)$.

Vermits $n > 30$, $np \geq 5$ en $nq \geq 5$ mogen we $B(n,p)$ benaderen door een normale verdeling:

$$n\bar{P} \sim \mathcal{N}(np, npq) \quad \text{en dus} \quad \bar{P} \sim \mathcal{N}\left(p, \frac{pq}{n}\right)$$

Dit betekent dat de steekproefproportie voor grote n normaal verdeeld is met een gemiddelde gelijk aan de populatieproportie p (\bar{P} is immers een zuivere schatter van p) en een variantie gelijk aan $\frac{pq}{n}$. Analoog als bij het steekproefgemiddelde hebben we dus een standaardfout op

de proportie $\sqrt{\frac{pq}{n}}$, die omgekeerd evenredig is met de wortel uit de steekproefgrootte.

Vermits we de populatieproportie niet kennen, kennen we deze standaardfout niet. Voor grote n maken we geen grote fout indien we de steekproefwaarden \bar{p} en $\bar{q} = 1 - \bar{p}$ invullen.

We hebben dan bij benadering dat : $\sqrt{\frac{pq}{n}} \approx \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

We kunnen dus stellen dat de verdeling $Z = \frac{\bar{P} - p}{\sqrt{\frac{pq}{n}}} \approx \mathcal{N}(0,1)$

en dus ook: $P\left(|Z| \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \approx 1 - \alpha$

We leiden hieruit het betrouwbaarheidsinterval voor p af:

$$P\left(\left|\bar{P} - p\right| \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right) = 1 - \alpha$$

of $P\left(\bar{P} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \leq p \leq \bar{P} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right) = 1 - \alpha$

Het volgende interval is dus een benaderd BI voor p met betrouwbaarheid $1 - \alpha$:

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

We hebben dus weer een betrouwbaarheidsinterval van de vorm:

Steekproefproportie $\pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ keer de standaardfout op de steekproefproportie

Voorbeeld

De paardehorzel legt eitjes in wonderen van warmbloedige dieren, wat ernstige infecties tot gevolg kan hebben. In een experiment werden larven van de paardehorzel blootgesteld aan een stralingsdosis van 2500 rad, met de bedoeling de meeste mannetjes steriel te maken. Aangezien de vrouwtjes slecht éénmaal paren, resulteert het paren met een steriel mannetje in steriele eitjes. Na de bestraling observeerde men dat bij 415 van de 500 paringen de eitjes steriel waren. Stel het 95% betrouwbaarheidsinterval op voor de proportie steriele paringen, die teweeggebracht wordt door een stralingsdosis van 2500 rad.

Oplossing:

Beschouw de stochastische variabele X met $X=1$ als de paring steriel is na bestraling
 $X=0$ als de paring niet steriel is na bestraling

$\rightarrow X \sim B(1,p)$ met p de populatieproportie van steriele paringen na bestraling.

Puntschatting voor p : $\bar{p} = \frac{415}{500} = 0.83$ en $n > 30$, $np \geq 5$ en $nq \geq 5 \rightarrow$ normale

benadering O.K.

Het 95% BI voor p wordt dan:

$$\left[\bar{p} - 1.96 \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + 1.96 \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right] = [0.7971, 0.8629] \approx [80\%, 86\%]$$

N.B.: Men kan bewijzen dat $\frac{1}{4}$ een bovengrens is voor $p(1-p)$ (Doe dit). Zonder de benadering \bar{p} voor p te gebruiken kun je dan toch een BI opstellen, dat vanwege het gebruik van deze bovengrens groter zal zijn en dus zeker 0.95 betrouwbaar zal zijn:

$$\left[\bar{p} - 1.96\sqrt{\frac{1}{4n}}, \bar{p} + 1.96\sqrt{\frac{1}{4n}} \right] = [0.7862, 0.8738] \approx [78.6\%, 87.4\%]$$

Je kunt deze bovengrens ook gebruiken om je steekproefgrootte te bepalen; bedenk dat een bovengrens voor de lengte $2d$ van een $(1-\alpha)\%$ BI gegeven wordt door:

$$2d = 2\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{1}{4n}}$$

(d geeft een idee van de precisie; hoe kleiner d, hoe preciezer, dus hoe kleiner het interval)

$$\text{De steekproefgrootte } n \text{ moet dus voldoen aan: } \left(\frac{\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)}{2d} \right)^2 \leq n$$

Dit betekent dat bijvoorbeeld voor een 95% BI voor p met een precisie van 0.04 (=4%) de steekproefgrootte n minstens 601 moet zijn.

6.4.5 Het schatten van het verschil van twee populatiegemiddeldes

We nemen nu een steekproef van grootte n_1 uit een eerste populatie en van grootte n_2 uit een tweede populatie. Wat kunnen we nu zeggen over het verschil van de twee populatiegemiddeldes $\mu_2 - \mu_1$?

In paragraaf 6.3.3 zagen we dat wanneer de onderliggende populaties normaal verdeeld zijn;

$$X \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad \text{en} \quad Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

$$\text{er geldt dat: } \bar{Y}_{n_2} - \bar{X}_{n_1} \sim \mathcal{N}\left(\mu_2 - \mu_1, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

(Voor grote n hoeven de onderliggende populaties niet normaal verdeeld te zijn.)

We kunnen nu een betrouwbaarheidsinterval construeren voor $\mu_2 - \mu_1$ met betrouwbaarheid $1 - \alpha$:

i) Indien σ_1 en σ_2 bekend zijn en dus ook de standaardfout $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ bekomen we:

$$\left[(\bar{x}_{n_2} - \bar{x}_{n_1}) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_{n_2} - \bar{x}_{n_1}) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

waarbij \bar{x}_{n_1} en \bar{x}_{n_2} de steekproefgemiddeldes zijn.

ii) i) Indien σ_1 en σ_2 niet bekend zijn en we de standaardfout benaderen door $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

(met s_1^2 en s_2^2 de steekproefvarianties) bekomen we:

$$\left[(\bar{x}_{n_2} - \bar{x}_{n_1}) - t_{n_1+n_2-2, 1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_{n_2} - \bar{x}_{n_1}) + t_{n_1+n_2-2, 1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

De kwantielen van de normaal verdeling worden vervangen door deze van de t-verdeling met n_1+n_2-2 vrijheidsgraden. Er geldt weer dat indien n_1 en n_2 groot zijn de t verdeling naar de normaal verdeling convergeert en we dus bij benadering de kwantielen van de normaal

verdeling mogen gebruiken ook in het laatste geval waarbij we $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ benaderden door

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

6.5 Toetsen van hypothesen

6.5.1 Inleiding

Binnen de wetenschap moeten vaak binaire beslissingen genomen worden. Enkele voorbeelden hiervan zijn:

- Overschrijdt de neerslag van Ca in de V.S. de vooropgestelde norm ?
- Blijft de CO₂-uitstoot in België binnen de norm van het Kyoto verdrag ?
- Is een antibioticum de aangewezen behandeling voor patiënten met een neusinfectie ?
- Is deze patiënt HIV-positief ?
- Zijn de klanten tevreden over het aangekochte product ?
- Is deze man de echte vader van dit kind ?
- Is er een dosis-respons relatie tussen uren studie en examenresultaat ?

Ook in de industrie komt men dit soort problemen tegen, bijvoorbeeld in kwaliteitscontrole: een firma produceert batterijen voor draagbare radio's en beweert in haar advertentie dat elke batterij goed is voor 30 uur muziek. Een consumentenorganisatie wil deze bewering nagaan en test 40 batterijen, die op diverse plaatsen in het land gekocht werden. De gemiddelde speelduur van deze 40 batterijen is 28 uur met een spreiding van 2 uur. Kunnen we hieruit besluiten dat de reclame van de firma overdreven is ?

6.5.2 Werkwijze

Met behulp van de schattingstheorie uit de voorgaande paragraaf is het mogelijk om vanuit een steekproef een statistische uitspraak te doen over een populatieparameter (gemiddelde, variantie, propotie). Een andere manier om over deze populatieparameter een statistisch gefundeerde bewering te maken, is gebaseerd op de toetsingstheorie. Het geven van een ja/nee antwoord op de gestelde vragen is in zeker opzicht eenvoudiger dan een kwantitatieve schatting maken (BI) van de grootte van de effecten die ter discussie zijn.

6.5.3 De verschillende stappen

Concreet worden bij een statistische toetsingsprocedure de volgende stappen uitgevoerd:

1. Het opstellen van een nulhypothese (H_0) en een alternatieve hypothese (H_A). De nulhypothese zal men trachten te verwerpen ten voordele van de alternatieve hypothese. Voor enkele van de gestelde vragen zijn de hypotheses als volgt:

a) H_0 : De gemiddelde hoeveelheid neerslag van Ca bedraagt 1.5 kg per hectare.

H_A : De gemiddelde hoeveelheid neerslag van Ca bedraagt meer dan 1.5 kg per hectare.

b) H_0 : Antibiotica bevorderen de genezing van de neusinfectie niet.

H_A : Antibiotica bevorderen de genezing van de neusinfectie.

c) H_0 : Deze bodem bevat 5% kalk.

H_A : Deze bodem bevat geen 5% kalk.

2. Het definiëren van een toetsingsgrootheid T , een variabele die men op basis van de steekproef kan berekenen. T moet zodanig zijn dat de verdeling van T (bij benadering) gekend is in de veronderstelling dat de nulhypothese waar is. In de hierboven geformuleerde problemen in de toetsingsgrootheid gebaseerd op:

a) de gemiddelde hoeveelheid gemeten neerslag van Ca.

b) de geobserveerde propotie mensen waarbij antibiotica de genezing bevorderde.

c) de gemiddelde gemeten hoeveelheid kalk in de bodemstalen.

3. Het definiëren van een beslissingscriterium voor de toets. Als de geobserveerde waarde van T "te extreem" ligt ten opzichte van haar verdeling onder H_0 (d.w.z. als H_0 waar is), zal men H_0 verwerpen ten voordele van het alternatief. Men zal hierbij een betrouwbaarheidsdrempel α definiëren: H_0 zal verworpen worden van zodra men onder H_0 hoogstens $\alpha\%$ kans heeft op

een uitkomst die minstens zo extreem is als de geobserveerde T-waarde (de geobserveerde waarde ligt niet in het $(1-\alpha)\%$ -BI). Door de beslissing op deze laatste manier te nemen, aanvaardt men een gecontroleerde kans op een **type I fout**: H_0 verwerpen op basis van de steekproefgegevens terwijl H_0 in de doelpopulatie in werkelijkheid toch waar is. Deze kans is dus kleiner dan α .

4. Een goed opgezette studie moet ook de kans op een **type II fout** beperken : H_0 niet verwerpen op basis van de steekproefgegevens terwijl het alternatief in de populatie waar is en niet H_0 . De kans op een type II fout wordt meestal met β aangeduid en $1-\beta$ wordt het onderscheidingsvermogen (power) van een toets genoemd. Onderstaande tabel vat de beslissingsmogelijkheden nog eens samen:

		Werkelijkheid	
		H_0 is waar	H_A is waar
H_0 wordt verworpen	Type I fout kans α	Correcte uitspraak kans $1-\beta$	
	Correcte uitspraak kans $1-\alpha$	Type II fout kans β	

We zullen op de verschillende begrippen (hypotheses, toetsingsgroothed, betrouwbaarheidsdrempel, type I en II fouten, het onderscheidingsvermogen van een toets) en hun definities dieper ingaan. We zullen dit doen voor de volgende gevallen:

- toetsen van het gemiddelde als de variantie gekend is
- toetsen van het gemiddelde als de variantie onbekend is
- toetsen van de proportie
- toetsen van de variantie
- toetsen voor het vergelijken van de varianties in 2 groepen
- toetsen voor het vergelijken van de gemiddelden in 2 groepen
 - a. ongepaard
 - b. gepaard
- toetsen van frequenties

In de meeste gevallen zal er sprake zijn van een eenzijdige en een tweezijdige toets.

6.5.4 Het toetsen van het gemiddelde als de variantie gekend is

We starten met het formuleren van de hypotheses omtrent het populatiegemiddelde.

Voorbeelden zijn: $\begin{cases} H_0: \mu = 4 \\ H_A: \mu = 5 \end{cases}$ of $\begin{cases} H_0: \mu = 4 \\ H_A: \mu > 4 \end{cases}$ of $\begin{cases} H_0: \mu = 4 \\ H_A: \mu < 4 \end{cases}$ of $\begin{cases} H_0: \mu = 4 \\ H_A: \mu \neq 4 \end{cases}$

De keuze van de hypotheses hangt af van verschillende factoren. De nulhypothese moet zodanig gekozen worden dat het gedrag van statistische grootheden (de toetsingsgroothed T), die gebruikt worden bij het uitvoeren van de toets, te bepalen is in de onderstelling dat H_0 waar is. Bij problemen waar je een uitspraak in tegenstelling tot een andere wil bewijzen moet je de te bewijzen uitspraak als alternatieve hypothese formuleren. Het uitvoeren van een toets bewijst immers nooit de nulhypothese, wel eventueel de alternatieve.

We hebben reeds aangegeven dat we bij het toetsen heel lang aan de nulhypothese blijven vasthouden. We stappen slechts over op de alternatieve hypothese wanneer we in onze steekproef observaties waarnemen die we "helemaal niet" hadden verwacht of, anders

uitgedrukt, die “significant verschillend zijn” van wat we hadden verwacht onder H_0 . Hier moeten we op een éénduidige manier vastleggen wat “helemaal niet verwacht” of “significant verschillend van de verwachting onder H_0 ” betekent. We kiezen een maximaal toelaatbare kans α op een type I fout. Deze heet het significantieniveau. Met observaties die “helemaal niet verwacht” zijn, zullen we nu bedoelen dat de kans om deze waarden te observeren, kleiner dan α is. Het **significantieniveau (α)** wordt dus gedefinieerd als de kans om een type I fout te maken. Bij het uitvoeren van een toets zegt men dat “de nulhypothese (niet) verworpen wordt op significantieniveau α “.

Laat ons de verdere procedure illustreren met het volgende voorbeeld:

Onderstel dat uit onderzoeken van vorige jaren gebleken is dat de IQ-scores van eerstejaarsstudenten normaal verdeeld zijn met gemiddelde 113 en variantie 81. Dit academiejaar echter beweert het onderwijskundig studiebureau dat de nieuwe eerstejaarsstudenten uitzonderlijk goed zijn en een gemiddeld IQ hebben dat zeker groter is dan 113. Een lukraak getrokken steekproef van 9 eerstejaarsstudenten leverde volgende IQ-scores op: 129, 106, 112, 121, 118, 115, 106, 129 en 126.

Levert dit, op het 5% significantieniveau, een statistisch bewijs dat de uitspraak van het onderwijskundig studiebureau juist is ? Onderstel dat ook dit jaar de IQ-scores normaal verdeeld zijn met variantie 81.

Opstellen van het theoretisch model

X = de IQ-score van nieuwe eerstejaarsstudenten

$\mu = E[X]$ = het gemiddelde van deze IQ-scores

De hypotheses

$$H_0 : \mu = 113$$

$$H_A : \mu > 113 \text{ (de bewering van het studiebureau)}$$

Keuze van de toetsingsgrootheid (cf. het opstellen van BI)

Uit de opgave blijkt dat de variantie gekend is en dat de populatie normaal verdeeld is \rightarrow de

$$\text{toetsingsgrootheid } Z = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0,1)$$

Het beslissingcriterium

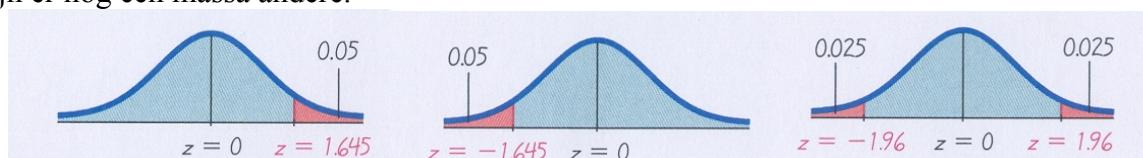
We werken in de veronderstelling dat de nulhypothese waar is, d.w.z. $\mu = 113$; tevens $\sigma^2 = 81$

$$\text{en } n = 9 \rightarrow \bar{X}_n \sim \mathcal{N}(113, 9) \quad \text{of} \quad Z = \frac{\bar{X}_n - 113}{3} \sim \mathcal{N}(0,1) \quad \text{onder } H_0$$

We kiezen 0.05 als significantieniveau.

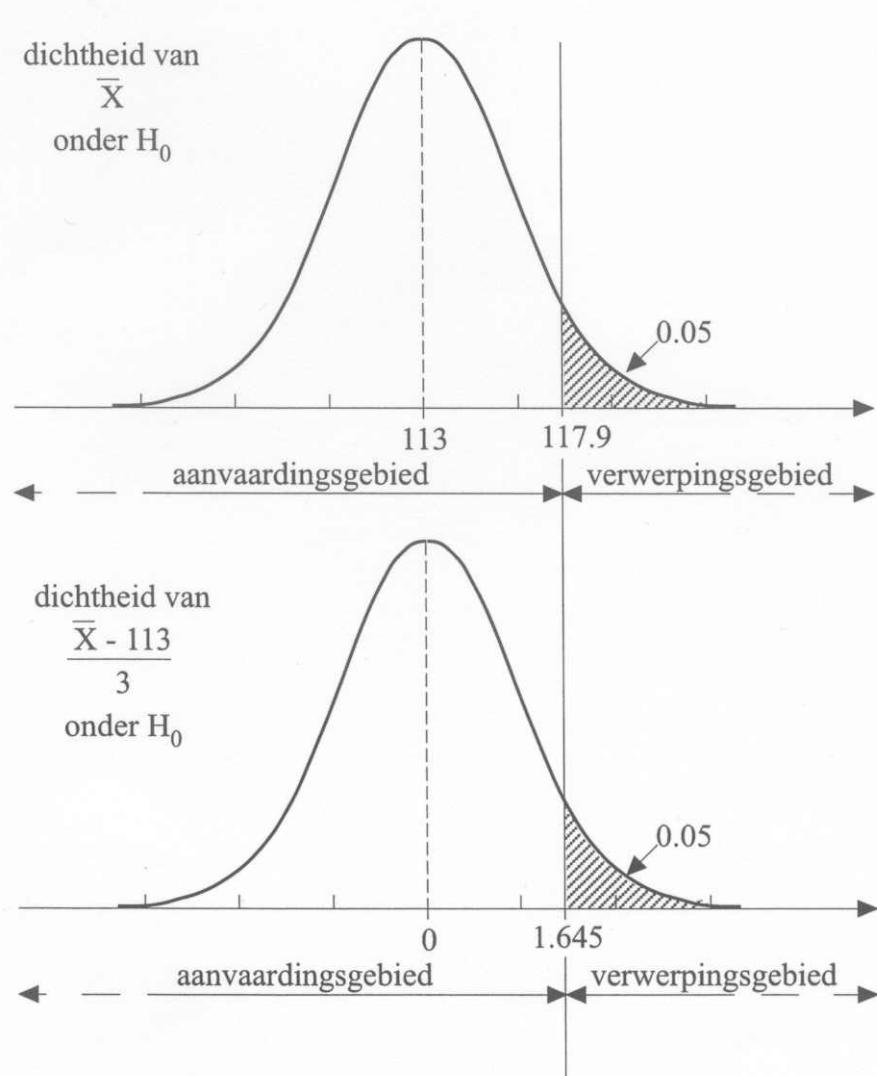
Aangezien $\frac{\bar{X}_n - 113}{3}$ zich onder H_0 gedraagt als een standaard normale verdeling kunnen we

een gebied afbakenen waarin $\frac{\bar{X}_n - 113}{3}$ slechts met kans 0.05 terechtkomt. Voorbeelden hiervan zijn (kijk in de tabel): $[1.645, \infty)$ ofwel $(-\infty, -1.645]$ ofwel $(-\infty, -1.96] \cup [1.96, \infty)$ en er zijn er nog een massa andere.



We moeten nu eerst afspreken welk gebied we nemen. Hiervoor laten we ons leiden door de alternatieve hypothese, er tevens aan denkend dat het verworpen van de nulhypothese

resulteert in het aanvaarden van de alternatieve hypothese. Onder H_0 verwachten we dat realisaties van \bar{X}_n in meerderheid rond 113 vallen en zelden ver verwijderd zijn van 113. Het zijn juist deze "bijna niet verwachte" extreme observaties die aanleiding gaan geven tot het verwerpen van H_0 . Maar onderstel nu eens dat we een extreem kleine observatie hebben, bijvoorbeeld: het steekproefgemiddelde is gelijk aan 105. Zelfs al hadden we dit onder H_0 niet verwacht, toch is het hier absurd om $H_0: \mu=113$ te verwerpen ... en dus te aanvaarden dat $\mu>113$, want dit zegt de alternatieve hypothese. Een zinvolle procedure voor dit voorbeeld zal er dus in bestaan dat we de nulhypothese verwerpen als we een "extreme", "bijna niet te verwachten" observatie van \bar{X}_n waarnemen die bovendien extreem in de richting van de alternatieve hypothese is ! Voor \bar{X}_n betekent dit waarden die veel groter zijn dan 113 en voor $\frac{\bar{X}_n - 113}{3}$ komt dit overeen met realisaties die "veel" groter zijn dan nul. Hoeveel groter halen we uit het significantieniveau. Het **verwerpingsgebied voor Z** wordt dus $[1.645, \infty)$ in dit voorbeeld. Bemerkt dat we het gebied waar we de nulhypothese niet kunnen verwerpen **aanvaardingsgebied** noemen.



NB: Met \bar{X} wordt \bar{X}_n bedoeld in deze figuur

$$P\left(Z = \frac{\bar{X}_n - 113}{3} > 1.645\right) = 1 - \Phi(1.645) = 1 - 0.95 = 0.05$$

of voor \bar{X}_n :

$$P(\bar{X}_n > 3 \cdot 1.645 + 113) = 0.05$$

$$\approx P(\bar{X}_n > 117.9)$$

Het punt dat de scheiding aangeeft tussen het aanvaardingsgebied en het verwerpingsgebied wordt het **kritisch punt** genoemd (hier 117.9 voor \bar{x}_n).

Wanneer het verwerpingsgebied zich slechts naar één kant uitstrekkt (hier alleen naar rechts) dan spreekt men over een éénzijdige toets. Als het verwerpingsgebied de unie van twee staarten is (zowel links als rechts) dan heet de toets tweezijdig.

Het nemen van een steekproef, berekeningen en besluit

Onze steekproef leverde $\bar{X}_9 = 118$. Er geldt dat $z = \frac{\bar{x}_9 - 113}{3} = \frac{118 - 113}{3} = 1.67$. Aangezien

1.67 behoort tot het verwerpingsgebied besluiten we dat de nulhypothese kan verworpen worden op het 5% significantieniveau. Dit betekent tevens dat op het 5% significantieniveau, statistisch bewezen is dat het studiebureau gelijk had.

We hadden ook kunnen zien dat $118 > 117.9$ is.

De tweezijdige toets

Onderstel terug de situatie van het voorbeeld, waarbij een steekproef van de IQ-scores van 9 studenten wordt genomen. Er wordt aangenomen dat de IQ-scores normaal verdeeld zijn met variantie 81. Nu vraagt het studiebureau na te gaan of de nieuwe lichting eerstejaarsstudenten al dan niet overeenstemt met het algemeen aanvaarde gemiddelde IQ van 113. Toets op het 5% significantieniveau.

Opstellen van het theoretisch model

X = de IQ-score van nieuwe eerstejaarsstudenten

$\mu = E[X]$ = het gemiddelde van deze IQ-scores

$\alpha = 0.05$

De hypotheses

$H_0 : \mu = 113$

$H_A : \mu \neq 113$

Keuze van de toetsingsgrootheid

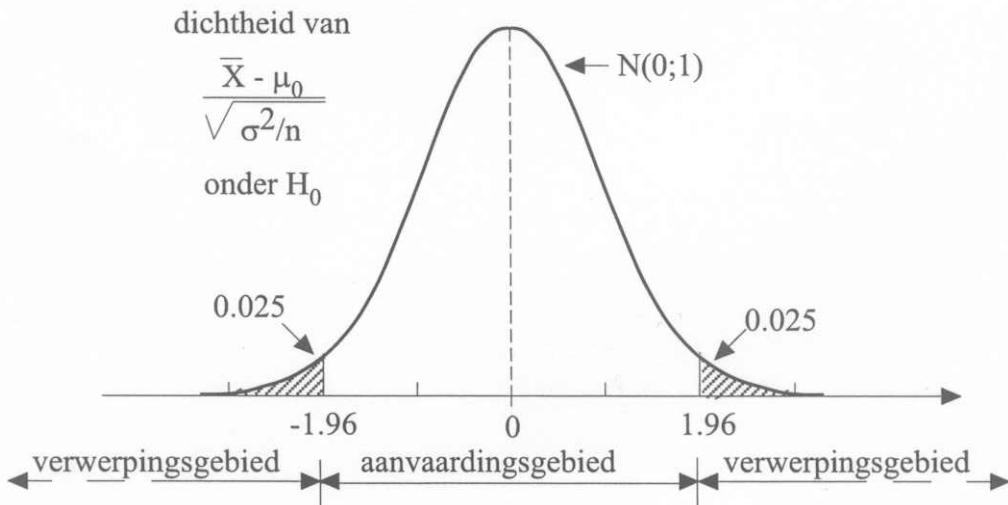
Uit de opgave blijkt dat de variantie gekend is en dat de populatie normaal verdeeld is

$$\rightarrow Z = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0,1) \rightarrow \text{de toetsingsgrootheid } Z = \frac{\bar{X}_9 - 113}{3} \sim \mathcal{N}(0,1) \text{ onder } H_0$$

of $P\left(\Phi^{-1}(\alpha/2) < \frac{\bar{X}_9 - 113}{3} < \Phi^{-1}(1-\alpha/2)\right) = 1 - \alpha = 0.95$

of $P\left(-1.96 < \frac{\bar{X}_9 - 113}{3} < 1.96\right) = 0.95$

of $P(107.1 < \bar{X}_9 < 118.9) = 0.95$



NB: In deze figuur wordt met \bar{X} , \bar{X}_n bedoeld en met μ_0 , μ .

De waarde van de toetsingsgrootte (1.67) ligt in het aanvaardingsgebied of analoog het steekproefgemiddelde (118) ligt in het 95%-BI ([107.1, 118.9]). Je kunt de nulhypothese dus niet verwijzen op het 5% significantieniveau. Er is dus op het 5% significantieniveau geen statistisch bewijs dat de nieuwe generatie studenten een gemiddeld IQ heeft dat verschilt van het standaard IQ van 113.

Opmerking

Het formuleren van een hypothese moet gebeuren voordat je naar de steekproefresultaten kijkt. Onderstel even dat in de vorige situatie het onderwijskundig studiebureau eigenlijk alleen maar zinnens was de vraag te stellen of de nieuwe generatie al dan niet verschillend is, maar dat het eerst de steekproef had getrokken en naar de getallen gekeken. Op basis van deze getallen berekent het een steekproefgemiddelde van 118, dat groter is dan 113. Het verandert dus zijn vraag en toestelt of het gemiddelde IQ groter is dan 113. Dezelfde getallen die geleid hebben tot een verandering in vraagstelling, bevestigen hier ook die nieuwe vraag? Dergelijke statistische procedures zijn **onaanvaardbaar**!

Alvorens over te gaan op de andere toetsen introduceren we nog een belangrijke grootte bij de toetsingsprocedures namelijk de p-waarde.

De p-waarde

Tot nu toe hebben we bij het toetsen van hypothesen vooraf een significantieniveau vastgelegd (bijvoorbeeld 5%). Dit leidde dan tot een kritisch punt (of kritische punten) en de nulhypothese werd verworpen zodra de geobserveerde waarde van de gekozen toetsingsgrootte voldoende extreem was om in het verwerpingsgebied terecht te komen. De uitspraak "de nulhypothese kan verworpen worden op het 5% significantieniveau" geeft enerzijds heel wat informatie, maar vertelt tegelijkertijd toch ook niet het hele verhaal. We zullen nu een grootte definiëren die ons een antwoord geeft op de vraag: hebben we deze nulhypothese maar niet verworpen (de geobserveerde waarde lag dicht bij het kritisch punt) of zouden we ze ook verworpen op een veel lager significantieniveau (de geobserveerde waarde ligt ver van het kritisch punt)?

In ons voorbeeld van de eenzijdige toets verwijderen we de nulhypothese vermits $\frac{118-113}{3} = 1.67$ groter is dan het kritisch punt 1.645 of het steekproefgemiddelde 118 groter dan $117.9 = 113 + 3 \times 1.645$. Onderstel nu eens dat de IQ-scores van de 9 studenten een steekproefgemiddelde van 121 hadden opgeleverd. De nulhypothese zou dan ook verworpen worden ($\frac{121-113}{3} = 2.67 > 1.645$). Als we beide situaties vergelijken, dan geeft het toch wel een onbevredigend gevoel dat we onze conclusie niet kunnen nuanceren. Eigenlijk zouden we voor de tweede situatie een veel sterkere "ja" willen uitspreken omdat onze steekproefresultaten daar, meer dan in de eerste situatie, de nulhypothese onwaarschijnlijk maken.

Een manier om hieraan te verhelpen, en waarbij tegelijkertijd veel preciezere informatie wordt verkregen, is het aangeven van de p-waarde (prob-value) of overschrijdingskans (in SPSS vind je Sig). Dit is niets anders dan de kans om, als de nulhypothese waar is, uitkomsten te zien die "even extreem zijn als of nog extremer dan" de observatie die we hebben waargenomen. Dit levert :

$$\text{Eerste situatie } P\left(\frac{\bar{X}_9 - 113}{3} \geq 1.67\right) = 1 - \Phi(1.67) = 1 - 0.9525 = 0.0475$$

$$\text{Tweede situatie } P\left(\frac{\bar{X}_9 - 113}{3} \geq 2.67\right) = 1 - \Phi(2.67) = 1 - 0.9962 = 0.0038$$

Voor de eerste situatie is de p-waarde 0.0475 en voor de tweede situatie is de p-waarde 0.0038. Bemerkt dat meer extreme situaties corresponderen met kleinere p-waarden. Op een significantieniveau van 4.75% in het eerste geval ligt ons steekproefgemiddelde op de grens tussen het aanvaardings- en verwerpingsgebied. In het tweede geval is dit significantieniveau meer dan een factor 10 kleiner en zouden we de nulhypothese ook nog verwijderen op het 0.5% significantieniveau.

Indien een statistisch pakket ons enkel p-waarden geeft, dan valt de geobserveerde grootheid in het kritisch gebied (verwerpingsgebied; de nulhypothese wordt verwijderen), indien de p-waarde kleiner is dan het significantieniveau. In ons voorbeeld zijn zowel 4.75% als 0.38% kleiner dan 5%. In de tweede situatie, maar niet in de eerste, zouden we de nulhypothese ook verwijderen op het 0.5% significantieniveau..

Voor het berekenen van de overschrijdingskans bij een tweezijdige toets moet je de overschrijdingskans berekenen alsof je met een eenzijdige toets werkte en het verkregen resultaat verdubbelen.

We vatten deze paragraaf als volgt samen:

Samenvatting van de toets voor het gemiddelde als de variantie gekend is

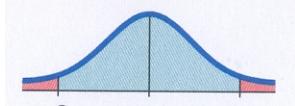
- De tweezijdige toets

$$H_0 : \mu = \mu_0$$

$$H_A : \mu \neq \mu_0$$

We verwijderen de nulhypothese op het niveau α niet, als de waarde z van de toetsingsgrootte Z berekend met de metingen in het aanvaardingsgebied

$$\left[\Phi^{-1}\left(\frac{\alpha}{2}\right), \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \right]$$



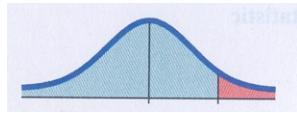
De overschrijdingskans p wordt in dit geval gegeven door $p = 2 P(Z \geq |z|)$

- De eenzijdige toets

Hier onderscheiden we 2 gevallen

- i) $H_0 : \mu = \mu_0$
- $H_A : \mu > \mu_0$

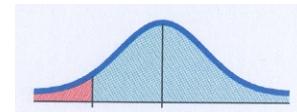
Het aanvaardingsgebied voor z is dan $]-\infty, \Phi^{-1}(1-\alpha)]$ en we verwerpen de nulhypothese als



$z > \Phi^{-1}(1-\alpha)$. De overschrijdingskans is dan $p = P(Z \geq z)$

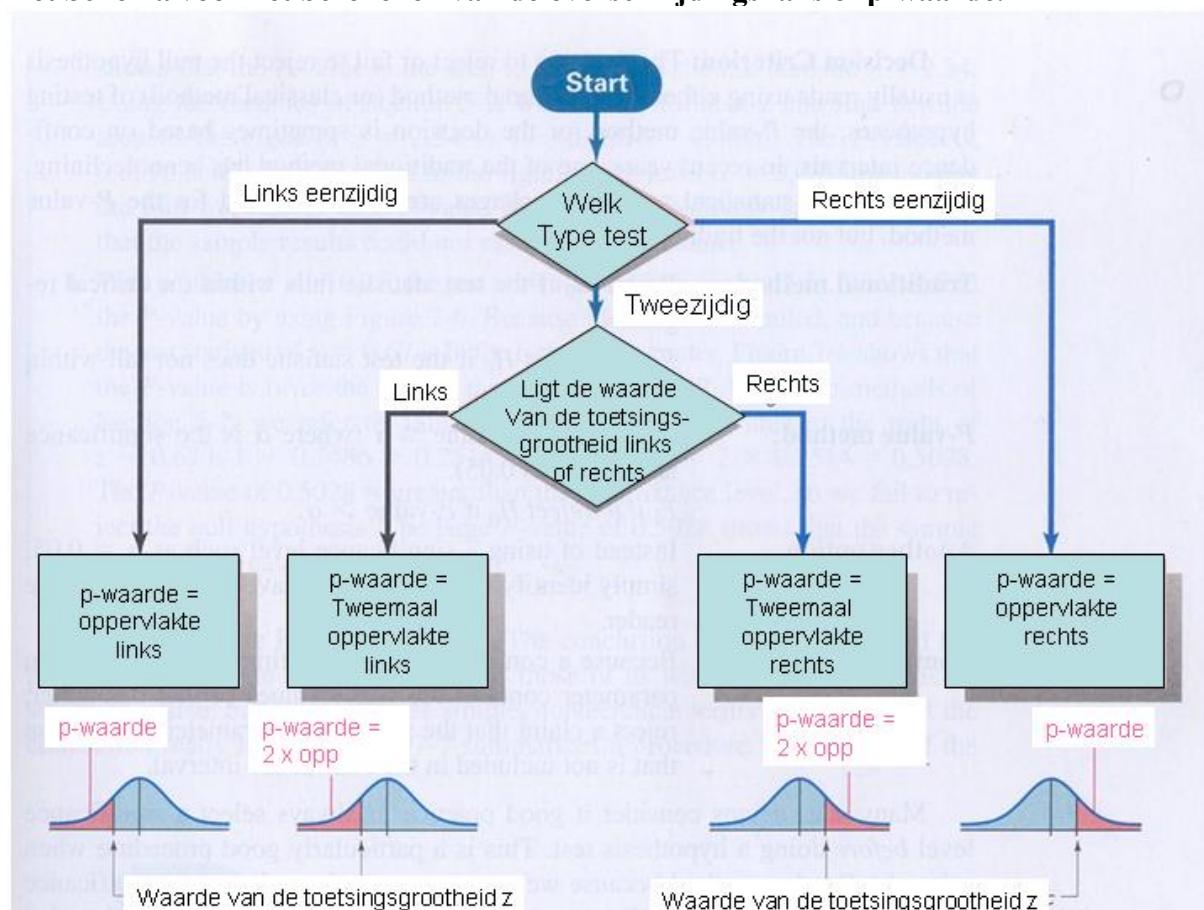
- ii) $H_0 : \mu = \mu_0$
- $H_A : \mu < \mu_0$

Het aanvaardingsgebied voor z is dan $[\Phi^{-1}(\alpha), +\infty[$ en we verwerpen de nulhypothese als



$z < \Phi^{-1}(\alpha)$. De overschrijdingskans is dan $p = P(Z \leq z)$

Het Schema voor het berekenen van de overschrijdingskans of p-waarde.



6.5.5 Het toetsen van het gemiddelde als de variantie onbekend is (de t-toets)

Analoog als bij het opstellen van de betrouwbaarheidsintervallen, zullen we nu de standaardfout van het steekproefgemiddelde ($\frac{\sigma}{\sqrt{n}}$) benaderen door $\frac{s_n}{\sqrt{n}}$. Deze benadering heeft tot gevolg dat we nu met de t-verdeling moeten werken in plaats van de normale verdeling. Dit is de reden waarom deze toets de t-toets wordt genoemd. De populatie moet **normaal verdeeld** zijn ofwel moet de steekproef groot genoeg zijn ($n > 30$).

We bespreken de eenzijdige toets aan de hand van een voorbeeld. De uitbreiding naar een tweezijdige toets is analoog als hiervoor.

Onderstel dat de IQ-scores van eerstejaarsstudenten normaal verdeeld zijn. Men neemt aan dat het gemiddelde gelijk is aan 113. Het onderwijskundig studiebureau beweert dat dit academiejaar het gemiddelde IQ zeker groter is dan 113. We beschikken echter niet over de variantie van de IQ-score. Dezelfde steekproef met steekproefgemiddelde 118 wordt getrokken. Toets op het 5% significantieniveau of het studiebureau gelijk heeft.

Opstellen van het theoretisch model

X = de IQ-score van nieuwe eerstejaarsstudenten

$\mu = E[X]$ = het gemiddelde van deze IQ-scores

De hypotheses

$H_0: \mu = 113$

$H_A: \mu > 113$ (de bewering van het studiebureau)

Keuze van de toetsingsgroothed

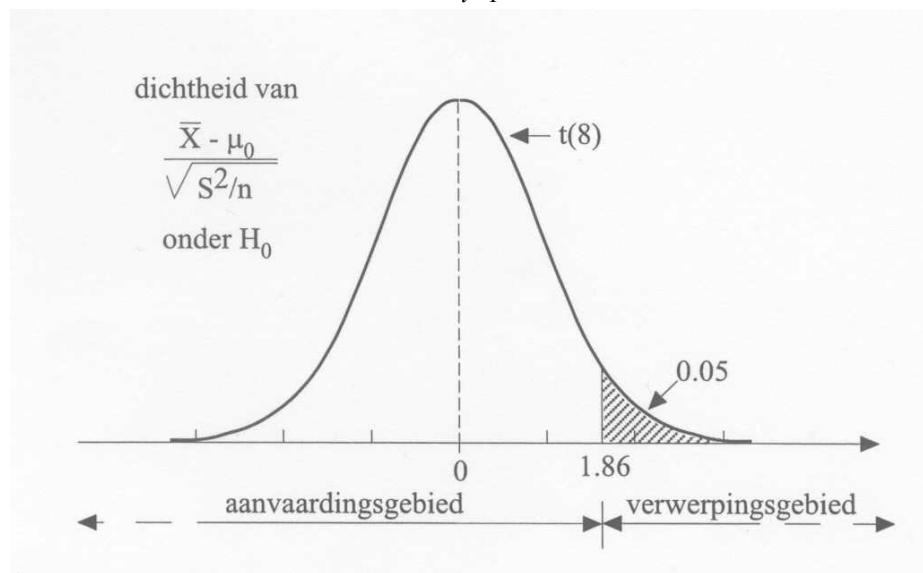
De steekproef is klein

De populatie is normaal verdeeld

De populatievariantie is niet bekend

→ de toetsingsgroothed $T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \sim t(n-1)$ onder H_0

Hier is $n = 9$; $\alpha = 0.05$; $\bar{x}_9 = 118$; $s_9^2 = \frac{1}{8} \sum_{i=1}^9 (x_i - 118)^2 = \frac{648}{8} = 81$



NB: In vorige figuur wordt met \bar{X} , \bar{X}_n bedoeld en niet μ_0 , μ .

$$\Rightarrow P\left(\frac{\bar{X}_9 - 113}{3} < t_{n-1,1-\alpha}\right) = P\left(\frac{\bar{X}_9 - 113}{3} < t_{8,0.95} = 1.86\right) = 1 - \alpha = 0.95$$

Vermits $\frac{\bar{x}_9 - 113}{\sqrt{s_9^2/9}} = \frac{118 - 113}{3} = 1.67$ in het aanvaardingsgebied valt ($1.67 < 1.86$) kunnen we de nulhypothese niet verwerpen op het 5% significantieniveau.

Vergelijk dit met het resultaat dat je kreeg als je de variantie wel kende. Toen kon je de nulhypothese wel verwerpen, ondanks het feit dat je dezelfde steekproef nam met een steekproefvariantie die gelijk was aan de populatievariantie. We zullen later zien dat de "power" van de laatste test niet zo groot is als deze van de eerste, waarbij we de variantie wel kenden. Deze toets houdt dus langer vast aan de nulhypothese. De t-verdeling heeft in de staarten immers meer kans dan een normale verdeling.

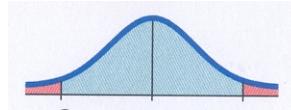
Samenvatting van de t-toets (populatie normaal verdeeld of $n>30$):

- De tweeziijdige t-toets

$$H_0 : \mu = \mu_0$$

$$H_A : \mu \neq \mu_0$$

We verwerpen de nulhypothese op het niveau α niet, als de waarde t van de toetsingsgrootheid T in het aanvaardingsgebied $\left[-t_{n-1,1-\frac{\alpha}{2}}, t_{n-1,1-\frac{\alpha}{2}}\right]$ ligt en we verwerpen H_0



als t er buiten ligt.

De overschrijdingskans p wordt in dit geval gegeven door $p = P(T \geq |t|)$

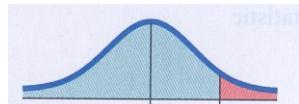
- De eenzijdige t-toets

Hier onderscheiden we weer 2 gevallen

i) $H_0 : \mu = \mu_0$

$$H_A : \mu > \mu_0$$

Het aanvaardingsgebied voor t is dan $[-\infty, t_{n-1,1-\alpha}]$ en we verwerpen de nulhypothese als

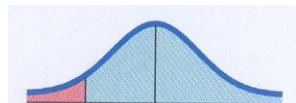


$z > t_{n-1,1-\alpha}$. De overschrijdingskans is dan $p = P(T \geq t)$

ii) $H_0 : \mu = \mu_0$

$$H_A : \mu < \mu_0$$

Het aanvaardingsgebied voor t is dan $[t_{n-1,\alpha}, +\infty[$ en we verwerpen de nulhypothese als



$t < t_{n-1,\alpha}$. De overschrijdingskans is dan $p = P(T \leq t)$

6.5.6 Het toetsen van een proportie

We gaan onze populatie beschrijven door een stochastische veranderlijke X . Zoals bij het opstellen van een BI voor een populatieproportie spreken we af dat het getal 0 overeenstemt met een van beide categorieën en het getal 1 met de andere. Deze stochastische veranderlijke X is Bernoulli verdeeld met parameter p ; $X \sim B(1,p)$

We zagen reeds bij de behandeling van BI's dat $n\bar{P} = n\bar{X}_n \sim B(n,p)$.

We construeren een toetsingsgroothed met behulp van $\bar{P} = \bar{X}_n$

Voor $n \geq 30$, $np \geq 5$ en $nq \geq 5$ mogen we $B(n,p)$ benaderen door een normale verdeling:

$$n\bar{X}_n \sim \mathcal{N}(np, npq) \quad \text{en dus} \quad \bar{P} = \bar{X}_n \sim \mathcal{N}\left(p, \frac{pq}{n}\right) \quad \text{of} \quad \frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \sim \mathcal{N}(0,1)$$

We kunnen dan $\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}}$ als toetsingsgroothed nemen.

We illustreren de toets weer aan de hand van 2 voorbeelden, één waarbij de normale benadering kan gebruikt worden en één waarin deze benadering niet kan gebruikt worden.

Voorbeeld 1

Overdracht van aangeleerd gedrag door kannibalisme is een controversiële theorie. Een bepaalde proef, die deze theorie statistisch probeerde te ondersteunen, bestond erin dat een aantal planariën (trilwormen) geleerd werd om elektrische shocks te vermijden. Deze planariën werden dan vermalen en als voedsel toegediend aan 100 niet-getrainde planariën, die normaal gesproken met kans 0.5 de elektrische shocks in het experiment vermijden. Men observeerde dat 57 planariën de shocks vermeden. Ondersteunt dit de theorie op het 10% significantieniveau? Wat is de p -waarde?

Oplossing

De populatie bestaat uit niet-getrainde planariën, die getrainde hebben opgegeten.

Zij $X=1$ als hij een shock ontwijkt.

$X=0$ anders

p = populatieproportie (% dat in staat is een electrische shock te vermijden)

$H_0 : p = 0.5$ (weerspiegelt dat kannibalisme niet helpt)

$H_A : p > 0.5$ (weerspiegelt de controversiële theorie)

Aangezien de steekproef groot is en p niet te groot of te klein is ($np > 5$ en $nq > 5$ onder nulhypothese) hebben we bij benadering onder H_0 :

$$\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} = \frac{\bar{X}_{100} - 0.5}{\sqrt{0.5 \cdot 0.5/100}} = \frac{\bar{X}_{100} - 0.5}{0.05} \sim \mathcal{N}(0,1)$$

Het significantieniveau is gelijk aan 10% en de alternatieve hypothese specificeert "groter dan" zodat we rechts eenzijdig toetsen. Dit leidt tot het volgende verwerpingsgebied: $[1.28, \infty)$.

We berekenen nu de waarde van de toetsingsgroothed voor onze steekproef ($\bar{p} = \bar{x}_{100} = 0.57$).

$$\frac{\bar{x}_{100} - 0.5}{0.05} = \frac{0.57 - 0.5}{0.05} = 1.4$$

Aangezien 1.4 in het verwerpingsgebied ligt bevestigt de proef op 10% significantieniveau dat kannibalisme helpt bij het overdragen van aangeleerd gedrag (tenminste bij trilwormen).

De p-waarde is gelijk aan: $p\text{-waarde} = P\left(\frac{\bar{X}_{100} - 0.5}{\sqrt{0.5 \cdot 0.5/100}} \geq 1.4\right) = 1 - \Phi(1.4) = 0.0808$

wat tevens aantoont dat de nulhypothese niet kan verworpen worden op een significantieniveau dat kleiner is dan 8% (bijvoorbeeld 5%).

Voorbeeld 2

Onderstel dat je wil toetsen of 65% van de Vlaamse bevolking akkoord gaat met legalisatie van drugs, op basis van volgende beweringen:

- georganiseerde criminale benden zullen hun miljarden winst, afkomstig uit illegale drugshandel, zien verdwijnen
- de kleine criminaliteit, gepleegd om aan geld voor drugs te geraken, zal afnemen
- de aantrekkingskracht van "doen wat verboden is" zal wegvalLEN

Bij een steekproef van grootte 14 waren er 7 personen die akkoord gingen met legalisatie. Bevestigt dit je veronderstelling? Op welk niveau? Bereken de p-waarde.

Oplossing

Weze $X=1$ als iemand uit de bevolking akkoord gaat met legalisatie

$X=0$ anders

p = populatieproportie, die akkoord gaat met legalisatie

$$H_0 : p = 0.65$$

$$H_A : p \neq 0.65$$

Aangezien de steekproef klein is, kan de normale benadering niet gebruikt worden. We moeten dus de exacte verdeling $n\bar{P} = n\bar{X}_n \sim B(n,p)$ gebruiken

Uit de steekproef blijkt dat er 7 successen zijn. Tevens is onder H_0 $p=0.65$ en $np=9.1$ en dus:

$$\begin{aligned} P(|n\bar{X}_n - 9.1| \geq |7 - 9.1|) &= P(n\bar{X}_n - 9.1 \geq 2.1 \text{ of } n\bar{X}_n - 9.1 \leq -2.1) \\ &= P(n\bar{X}_n \geq 11.2) + P(n\bar{X}_n \leq 7) \quad \text{met } n\bar{X}_n \sim B(14, 0.65) \end{aligned}$$

Uit de tabel van de binomiale verdeling halen we dan dat :

$$P(n\bar{X}_n \geq 11.2) = P(n\bar{X}_n \geq 12) = P(Y \leq 2) = 0.0839 \quad \text{met } Y \sim B(14, 0.35)$$

$$P(n\bar{X}_n \leq 7) = P(Y \geq 7) = 1 - P(Y \leq 6) = 0.1836$$

zodat de p-waarde gelijk is aan $0.0839 + 0.1836 = 0.2675$

Op het klassieke significantieniveau van 5% kan de nulhypothese dus niet verworpen worden.

6.5.7 Toets voor de variantie: de χ^2 -toets

We hebben n onafhankelijke metingen $\{x_1, x_2, \dots, x_n\}$ van een normaal verdeelde grootheid $X \sim \mathcal{N}(\mu, \sigma^2)$. Beweerd wordt dat de variantie gelijk is aan een bepaalde concrete waarde σ_0^2 .

We willen de waarheid van deze bewering toetsen aan de hand van de metingen. We veronderstellen weer dat de metingen trekkingen zijn uit n onafhankelijke stochastieken $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, waarbij zowel μ als σ^2 onbekend zijn. We kiezen als nulhypothese $\sigma^2 = \sigma_0^2$. We weten dat onder de nulhypothese geldt:

$$Y := \frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi_{n-1}^2 \quad \text{zodat}$$

$$P(Y < \chi_{n-1, \alpha}^2) = \alpha ; P(Y > \chi_{n-1, 1-\alpha}^2) = \alpha ;$$

$$P\left(\chi_{n-1, \frac{\alpha}{2}}^2 < Y\right) + P\left(Y > \chi_{n-1, 1-\frac{\alpha}{2}}^2\right) = 1 - \alpha$$

We nemen Y als toetsingsgrootte en berekenen zijn waarde uit de metingen $\chi := \frac{(n-1)s_n^2}{\sigma_0^2}$.

Afhankelijk van de alternatieve hypothese maken we een keuze tussen een eenzijdige of tweezijdige toets en beslissen we dus of we alleen grote waarden van Y of alleen kleine waarden of beiden onverenigbaar vinden met de nulhypothese.

Samenvatting van de χ^2 -toets

- De tweezijdige χ^2 -toets

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_A : \sigma^2 \neq \sigma_0^2$$

We verwerpen de nulhypothese op het niveau α niet als de χ -waarde van de metingen in het

aanvaardingsgebied $\left[\chi_{n-1, \frac{\alpha}{2}}^2, \chi_{n-1, 1-\frac{\alpha}{2}}^2 \right]$ ligt en we verwerpen H_0 als χ er buiten ligt.

De overschrijdingskans p wordt in dit geval gegeven door:

$$p = 2 \min\{P(Y \geq \chi), P(Y \leq \chi)\}$$

- De eenzijdige χ^2 -toets

Hier onderscheiden we 2 gevallen

$$\text{i)} \quad H_0 : \sigma^2 = \sigma_0^2$$

$$H_A : \sigma^2 > \sigma_0^2$$

Het aanvaardingsgebied voor χ is dan $\left[0, \chi_{n-1, 1-\alpha}^2 \right]$ en we verwerpen de nulhypothese als $\chi > \chi_{n-1, 1-\alpha}^2$. De overschrijdingskans is dan $p = P(Y \geq \chi)$

$$\text{ii) } H_0 : \sigma^2 = \sigma_0^2$$

$$H_A : \sigma^2 < \sigma_0^2$$

Het aanvaardingsgebied voor χ^2 is dan $[\chi_{n-1,\alpha}^2, +\infty]$ en we verwerpen de nulhypothese als $\chi^2 < \chi_{n-1,\alpha}^2$. De overschrijdingskans is dan $p = P(Y \leq \chi^2)$

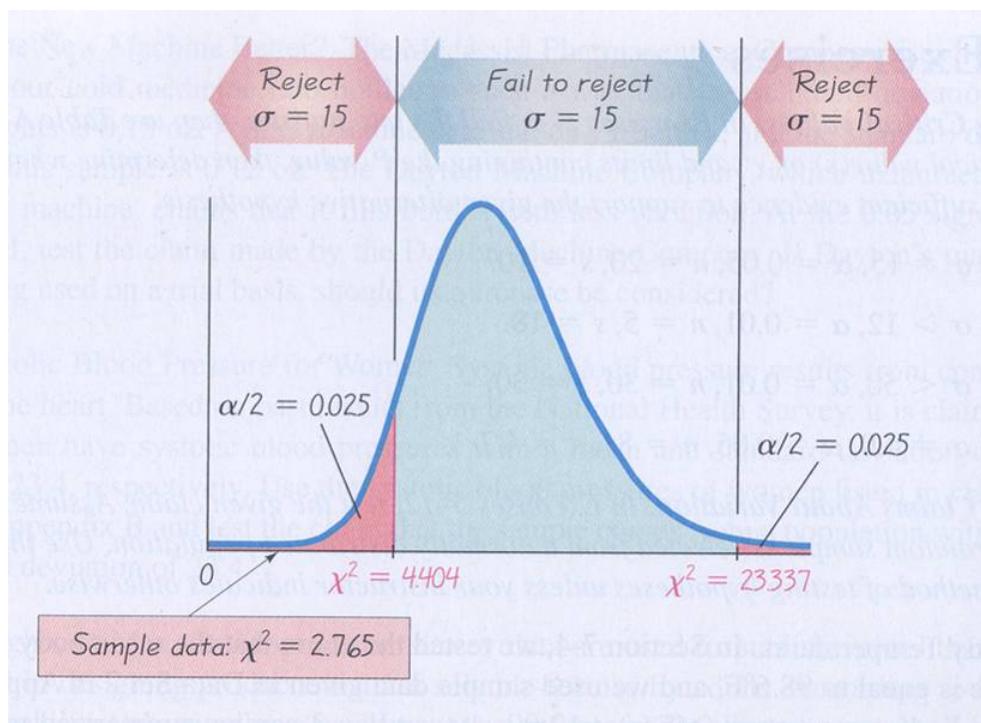
Voorbeeld: De IQ-scores van professoren statistiek

IQ-scores van volwassenen zijn normaal verdeeld met gemiddelde 100 en standaardafwijking 15. Een steekproef van 13 statistiek professoren levert een standaardafwijking $s=7.2$ voor hun IQ-scores. Een psycholoog is ervan overtuigd dat de IQ-scores van de professoren statistiek hoger liggen dan 100. In de veronderstelling dat de IQ-scores van professoren statistiek ook normaal verdeeld zijn toetst hij de hypothese dat de standaardafwijking van de IQ-scores van de populatie van alle statistiek professoren gelijk is aan 15 of de variantie gelijk is aan 225 met een significantie van 5%.

$$H_0 : \sigma^2 = 225$$

$$H_A : \sigma^2 \neq 225$$

$$\text{Toetsingsgrootheid: } Y := \frac{(n-1)s_n^2}{\sigma_0^2} \sim \chi_{n-1}^2 \quad \text{en dus } \chi^2 = \frac{(13-1)(7.2)^2}{(15)^2} = 2.765$$



In de tabel van de χ^2 -verdeling met 12 vrijheidsgraden vinden we de kritische punten 4.404 en 23.337. Vermits de waarde van onze toetsingsgrootheid (2.765) in het verwerpingsgebied ligt, besluiten we met een betrouwbaarheid van 0.95 dat de variantie (of standaardafwijking) van de IQ-scores van de populatie van statistiek professoren niet gelijk is aan 225 (of 15).

Voorbeeld

De standaardafwijking van het gewicht van blikjes groenten van 500g is altijd 2.5g geweest, maar men vreest dat de variabiliteit groter zou kunnen geworden zijn omdat de machines oud zijn. Een steekproef van 20 blikjes levert $s_{20} = 3.2\text{g}$. Is deze stijging significant op niveau $\alpha = 5\%$? En op niveau $\alpha = 1\%$?

Eenzijdige toets

$$H_0 : \sigma^2 = (2.5\text{g})^2$$

$$H_A : \sigma^2 > (2.5\text{g})^2$$

Uit de metingen volgt $\chi^2 = \frac{(n-1)s_n^2}{\sigma_0^2} = 31.13$. Aangezien $\chi^2 > \chi^2_{19,0.95} = 30.144$ wordt H_0 verworpen op het 5%-niveau. Anderzijds geldt $\chi^2 < \chi^2_{19,0.99} = 36.191$ zodat H_0 niet verworpen wordt op het 1%-niveau. De overschrijdingskans $p = P(Y \geq \chi^2) = 1 - F_{\chi^2_{19}}(31.13) = 0.039$, zodat de nulhypothese vanaf het 3.9%-niveau aanvaard wordt. $F_{\chi^2_{19}}$ is de verdelingsfunctie van de χ^2_{19} verdeling.

N.B.: Als we ons enkel de vraag hadden gesteld of de variabiliteit veranderd was, hadden we een tweezijdige toets uitgevoerd. Het aanvaardingsgebied was dan:

$[\chi^2_{19,0.025}, \chi^2_{19,0.975}] = [8.907, 32.852]$, zodat de nulhypothese niet verworpen werd en we zouden besloten hebben dat de variabiliteit niet veranderd was.

6.5.8 Toets voor het vergelijken van varianties in 2 groepen: de F-toets

Neem de onafhankelijke steekproeven $\{X_1, X_2, \dots, X_m\}$ en $\{Y_1, Y_2, \dots, Y_n\}$; beide normaal verdeeld met onbekende parameters. We hebben dus $m+n$ onafhankelijke normaal verdeelde stochastieken $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ met $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ voor $i=1, \dots, m$ en $Y_j \sim \mathcal{N}(\mu_2, \sigma_2^2)$ voor $j=1, \dots, n$. De parameters $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ zijn onbekend. We willen nu toetsen of $\sigma_1 = \sigma_2$. We schatten daarvoor eerst σ_1^2 en σ_2^2 met behulp van de steekproefvarianties :

$$S_1^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X}_m)^2 \quad \text{en} \quad S_2^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$$

We weten dat:

$$\frac{(m-1)S_1^2}{\sigma_1^2} \sim \chi^2_{m-1} \quad \text{en} \quad \frac{(n-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n-1}$$

zodat het quotiënt $\frac{S_1^2}{\sigma_1^2} \frac{\sigma_2^2}{S_2^2} \sim F_{m-1, n-1}$

Voor de toets op de gelijkheid van σ_1^2 en σ_2^2 nemen we als nulhypothese $\sigma_1^2 = \sigma_2^2$, zodat onder de nulhypothese geldt dat de toetsingsgrootheid $F = \frac{S_1^2}{S_2^2} \sim F_{m-1, n-1}$

zodat $P(F \leq F_{m-1,n-1,\alpha}) = \alpha$

Hierop baseren we de F-toets voor het vergelijken van twee varianties. We onderscheiden weer de twee- en eenzijdige toets (keuze afhankelijk van de alternatieve hypothese).

- De tweezijdige F-toets

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_A: \sigma_1^2 \neq \sigma_2^2$$

We berekenen de f-waarde van de steekproef: $f = \frac{s_1^2}{s_2^2}$

We aanvaarden de nulhypothese op het niveau α , als de f-waarde van de metingen in het aanvaardingsgebied $[F_{m-1,n-1}, \frac{\alpha}{2}, F_{m-1,n-1,1-\alpha}]$ ligt en we verwerpen H_0 als f er buiten ligt.

De overschrijdingskans of p-waarde wordt in dit geval gegeven door:

$$p = 2 \min\{P(F < f), P(F > f)\} = 2 \min\{F_{m-1,n-1}(f), 1 - F_{m-1,n-1}(f)\}$$

- De eenzijdige F-toets

We berekenen de f-waarde van de steekproef: $f = \frac{s_1^2}{s_2^2}$

i) $H_0: \sigma_1^2 = \sigma_2^2$

$$H_A: \sigma_1^2 > \sigma_2^2$$

Met aanvaardingsgebied: $[0, F_{m-1,n-1,1-\alpha}]$

En overschrijdingskans: $p = 1 - F_{m-1,n-1}(f)$

i) $H_0: \sigma_1^2 = \sigma_2^2$

$$H_A: \sigma_1^2 < \sigma_2^2$$

Met aanvaardingsgebied: $[F_{m-1,n-1,\alpha}, \infty]$

En overschrijdingskans: $p = F_{m-1,n-1}(f)$

Opmerking: Uit de bovenstaande formules volgt dat je voor de tweezijdige toets steeds twee F-waarden moet bepalen. Dit is echter niet nodig. Als je X en Y zodanig kiest dat $\sigma_1^2 \geq \sigma_2^2$, dan geldt automatisch dat de toetsingsgroothed $f \geq 1$. Omdat de mediaan niet teveel van 1 verschilt, zal de linkergrens veel kleiner dan 1 zijn en ligt f automatisch aan de rechterzijde van $F_{m-1,n-1,\frac{\alpha}{2}}$, zodat we alleen hoeven na te gaan of f kleiner is dan de rechtergrens. Het is

dan alsof je een eenzijdige toets doet met alternatief $\sigma_1^2 > \sigma_2^2$ en met de **helft** van het niveau. Voor de eenzijdige toets kun je X en Y ook steeds zodanig kiezen dat de alternatieve hypothese $\sigma_1^2 > \sigma_2^2$ is.

Voorbeeld

We willen nagaan of twee voltmeters dezelfde nauwkeurigheid bezitten. De variantie is hier een maat voor de nauwkeurigheid. Met elk toestel wordt een bepaalde meting een aantal malen uitgevoerd met resultaat:

toestel 1: $s_1 = 4\mu V$, $m=16$
toestel 2: $s_2 = 3\mu V$, $n=21$

De f-waarde: $f = 16/9 = 1.778$

$$\text{De hypothesen zijn: } H_0: \sigma_1^2 = \sigma_2^2 \\ H_A: \sigma_1^2 \neq \sigma_2^2$$

Voor een toets op 5%-niveau geldt: $F_{m-1, n-1, \frac{\alpha}{2}} = 1 / 2.76 = 0.362$ en $F_{m-1, n-1, 1-\frac{\alpha}{2}} = 2.57$.

$f = 1.778$ ligt in het aanvaardingsgebied $[0.362, 2.57]$ zodat de nulhypothese op het 5%-niveau aanvaard wordt.

N.B.: Wegens voorgaande opmerking had je $F_{m-1, n-1, \frac{\alpha}{2}}$ eigenlijk niet nodig.

6.5.9 Het toetsen van 2 gemiddeldes

6.5.9.a) De ongepaarde toets: de t-toets

Zoals bij de F-toets beschikken we over de onafhankelijke steekproeven $\{X_1, X_2, \dots, X_m\}$ en $\{Y_1, Y_2, \dots, Y_n\}$; beide normaal verdeeld met onbekende parameters. We hebben dus $m+n$ onafhankelijke normaal verdeelde stochastieken $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ met $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ voor $i=1, \dots, m$ en $Y_j \sim \mathcal{N}(\mu_2, \sigma_2^2)$ voor $j=1, \dots, n$. We zullen nu echter veronderstellen dat we mogen aannemen dat de variantie van beide steekproeven dezelfde is (eventueel na het uitvoeren van een F-toets), dus $\sigma_1^2 = \sigma_2^2 = \sigma^2$. De parameters μ_1, μ_2 en σ^2 zijn onbekend. We willen nu toetsen of $\mu_1 = \mu_2$.

We schatten daarvoor eerst σ^2 met behulp van de steekproefvarianties s_1^2 en s_2^2 .

$$\text{Er geldt dat: } \bar{s}^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$$

een zuivere schatter voor σ^2 is. (Bedenk dat zowel s_1^2 als s_2^2 zuivere schatters voor σ^2 zijn en dus zeker elk gewogen gemiddelde.)

\bar{s}^2 wordt ook wel eens de gewogen of gemengde variantie genoemd.

$$\text{Onder de nulhypothese } (\mu_1 = \mu_2) \text{ geldt: } \bar{X}_m - \bar{Y}_n \sim \mathcal{N}\left(0; \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\right)$$

$$\text{ofwel: } U = \frac{(\bar{X}_m - \bar{Y}_n)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \mathcal{N}(0, 1)$$

We weten dat:

$$\frac{(m-1)s_1^2}{\sigma^2} \sim \chi_{m-1}^2 \quad \text{en} \quad \frac{(n-1)s_2^2}{\sigma^2} \sim \chi_{n-1}^2$$

zodat de gemengde variantie ook een chi-kwadraat verdeling volgt:

$$V = \frac{(m-1)s_1^2 + (n-1)s_2^2}{\sigma^2} \sim \chi_{m+n-2}^2$$

De statistiek $T = \frac{U}{\sqrt{\frac{V}{m+n-2}}}$ volgt dus een t-verdeling met $m+n-2$ vrijheidsgraden.

Nu we over deze informatie beschikken kunnen we overgaan tot het formuleren van de t-toets. We berekenen de toetsingsgroothed T voor de steekproef en verkrijgen zo de t-waarde:

$$t = \frac{(\bar{x}_m - \bar{y}_n)}{\bar{s} \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

- De tweezijdige t-toets

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

We aanvaarden de nulhypothese op het niveau α , als de t-waarde in het aanvaardingsgebied $\left[-t_{m+n-2,1-\frac{\alpha}{2}}, t_{m+n-2,1-\frac{\alpha}{2}}\right]$ ligt en we verwerpen H_0 als t er buiten ligt.

De overschrijdingskans of p-waarde wordt in dit geval gegeven door:

$$p = 2F_{t_{m+n-2}}(-|t|) \quad F_{t_{m+n-2}} \text{ is de verdelingsfunctie van de } t_{m+n-2} \text{ verdeling}$$

- De eenzijdige t-toets

i) $H_0 : \mu_1 = \mu_2$

$$H_A : \mu_1 > \mu_2$$

aanvaardingsgebied $]-\infty, t_{m+n-2,1-\alpha}]$

overschrijdingskans $p = F_{t_{m+n-2}}(-t)$

ii) $H_0 : \mu_1 = \mu_2$

$$H_A : \mu_1 < \mu_2$$

aanvaardingsgebied $[-t_{m+n-2,1-\alpha}, +\infty[$

overschrijdingskans $p = F_{t_{m+n-2}}(t)$

Voorbeeld

Een boer wil nagaan of het gebruik van een bepaalde soort kunstmest een verbetering van de graanoogst geeft. Daartoe kiest hij 15 stroken akker met dezelfde oppervlakte, waarvan er 8 worden behandeld met de meststof en de overige 7 niet (de controlegroep). De gemiddelde graanopbrengst \bar{x}_8 op de behandelde akkers is 5.8 ton met een standaardafwijking van 0.36 ton. Voor de controlegroep is de gemiddelde opbrengst $\bar{x}_7 = 4.9$ ton met een standaardafwijking van 0.4 ton. Is de produktie op het 1%-niveau significant hoger op de behandelde akkers?

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 > \mu_2$$

Ga zelf eerst m.b.v. een F-toets na dat de steekproefvarianties niet significant verschillen.

Voor de gemengde variantie vinden we: $\bar{s}^2 = \frac{1}{13}(7s_1^2 + 6s_2^2) = 0.1436 \text{ ton}^2$ zodat $\bar{s} = 0.379$ ton.

$$\text{Bijgevolg is } t = \frac{\bar{x}_8 - \bar{x}_7}{\bar{s} \sqrt{\frac{1}{8} + \frac{1}{7}}} = 4.59.$$

Het aanvaardingsgebied voor t is: $(-\infty, t_{13,0.99}] = (-\infty, 2.6503]$. De nulhypothese wordt verworpen op het 1%-niveau. De verhoging van de opbrengst is significant.

NB: Als we een t-toets willen doen voor de vergelijking van de gemiddeldes van twee groepen, moeten we altijd eerst een F-toets doen om te testen of de varianties gelijk mogen worden verondersteld. In de literatuur wordt het gebruik van robuustere alternatieven voor deze test aanbevolen, met name de varianten van Bartlett en van Levene. In SPSS wordt de F-toets van Levene uitgevoerd. We gaan er hier niet verder op in.

Een benaderende toets voor twee gemiddelden

Bij de afleiding van deze vergelijkende t-toets hebben we aangenomen dat $\sigma_1^2 = \sigma_2^2$. In het meest algemene geval is dit natuurlijk niet waar. Voor m en n zeer groot (>30) kunnen we dan een benaderende toets opstellen: we hebben dan dat $\sigma_1^2 \approx s_1^2$ en $\sigma_2^2 \approx s_2^2$ en we kunnen aannemen dat σ_1^2 en σ_2^2 bekend zijn. Onder de nulhypothese dat $\mu_1 = \mu_2$ geldt dan in goede benadering:

$$Z = \frac{(\bar{X}_m - \bar{Y}_n)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \mathcal{N}(0,1)$$

$$\text{We berekenen weer de z-waarde van de steekproef: } z = \frac{(\bar{x}_m - \bar{y}_n)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

en krijgen dan als aanvaardingsgebied voor z op niveau α :

$$\begin{array}{ll} \text{Tweezijdig} & \left[-\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \right] \\ \text{Eenzijdig} & \left[-\infty, \Phi^{-1}(1 - \alpha) \right] \quad \text{of} \quad \left[-\Phi^{-1}(1 - \alpha), +\infty \right] \end{array}$$

N.B.1: Het bovenstaande kun je natuurlijk ook gebruiken indien je populatievarianties σ_1^2 en σ_2^2 gekend zijn en de populaties normaal verdeeld zijn. Je berekent dan z met de exacte σ_1^2 en σ_2^2 .

N.B.2: In een situatie waarbij σ_1^2 en σ_2^2 niet gekend zijn en de F-toets zegt dat σ_1^2 en σ_2^2 significant verschillen zul je een niet-parametrische toets moeten gebruiken (zelfs indien beide populaties normaal verdeeld zijn). We gaan er hier niet verder op in.

N.B.3: Bij kleine steekproeven (n of $m < 30$) en populaties die niet normaal verdeeld zijn, zul je ook gebruik moeten maken van een niet-parametrische toets.

6.5.9.b) De gepaarde toets voor twee gemiddelden (verbonden steekproeven)

In de voorgaande paragrafen waren de steekproeven $\{X_1, X_2, \dots, X_m\}$ en $\{Y_1, Y_2, \dots, Y_n\}$ onafhankelijk. Dit is niet altijd het geval. Indien er een relatie is tussen elk element van de

eerste steekproef met juist één element van de tweede steekproef, moeten we anders te werk gaan. Om dit te verduidelijken beschouwen we het volgende voorbeeld. We willen 2 benzinemerken met elkaar vergelijken. We laten 10 auto's rijden met benzinemerk X en meten het verbruik. Vervolgens laten we bijvoorbeeld 12 auto's rijden met benzinemerk Y en meten weer het verbruik. Op de resultaten passen we de technieken toe uit de vorige paragraaf. Een probleem hierbij is dat het verbruik van 2 auto's, zelfs van hetzelfde merk en hetzelfde type, aanzienlijk kan verschillen en dat deze verschillen waarschijnlijk veel groter zijn dan die ten gevolge van kwaliteitsverschillen in de benzine. (We hebben dus eigenlijk een verdoken factor.) Een betere strategie is de volgende: eerst meten we het verbruik van 10 auto's, allemaal met benzinemerk X, en dan allemaal met benzinemerk Y (zie tabel onderstaande tabel). Met elke X_i komt dan een Y_i overeen; als er geen kwaliteitsverschil is, zullen de verschillen $X_i - Y_i$ een verwachtingswaarde nul hebben.

We spreken hier van verbonden waarnemingen. Andere voorbeelden van gelijksoortige experimenten zijn: metingen in linker- en rechteroor, eigenschappen van eeneiige tweelingen,...

In het algemeen hebben we dus n koppels statistieken (X_i, Y_i) met

$X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ en $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Onze nulhypothese zal zijn $H_0 : \mu_1 = \mu_2$

Stel nu $D_i = X_i - Y_i$ dan is: $D_i \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$

($\sigma_1^2 + \sigma_2^2$ is de variantie van het verschil van twee stochastieken met σ_1^2 en σ_2^2 als variantie.)

We kunnen nu de t-toets toepassen op D_i met nulhypothese $\mu = 0$.

De toetsingsgrootheid is dus: $T = \frac{\bar{D}_n}{S/\sqrt{n}} \sim t_{n-1}$ waarbij \bar{D}_n en S respectievelijk het

steekproefgemiddelde en de steekproefstandaardafwijking van D_i zijn.

Voorbeeld

Vergelijking van benzinemerken. Men vraagt zich af of het merk Y beter is dan het merk X. De resultaten van de metingen bij de 10 auto's geven:

I	1	2	3	4	5	6	7	8	9	10
X_i	99	110	105	101	90	92	104	100	101	100
Y_i	100	110	111	110	91	97	106	100	104	96
D_i	-1	0	-6	-9	-1	-5	-2	0	-3	4

We rekenen nu gemakkelijk uit, dat

$$\bar{d}_{10} = -2.3 \text{ km} \quad s_{10}^2 = 13.344 \text{ km}^2 \quad s_{10} = 3.6530 \text{ km} \quad \text{zodat} \quad t = \frac{\bar{d}_{10} \sqrt{10}}{s_{10}} = -1.99$$

De hypothesen zijn: $H_0 : \mu = 0$ en $H_A : \mu < 0$

Het aanvaardingsgebied voor t is $[-t_{n-1, 1-\alpha}, +\infty)$

Voor $\alpha = 5\%$ wordt dit $[-1.833, +\infty)$, zodat het verschil significant is op niveau 5%.

Voor $\alpha = 1\%$ wordt dit $[-2.821, +\infty)$, zodat het verschil niet significant is op niveau 1%.

N.B.: Indien onze steekproef groot genoeg is ($n > 30$) kunnen we weer de normale benadering gebruiken.

6.5.10 De Chi-kwadraat toets

Inleiding

De chi-kwadraat toets wordt gebruikt om geobserveerde frequenties te vergelijken met verwachte frequenties.

We kunnen dus vragen beantwoorden zoals:

- Hoe goed passen de geobserveerde gegevens bij een gepostuleerde verdelingsfunctie F? Zijn de gegevens normaal verdeeld, uniform verdeeld, Poisson verdeeld, ...?
- Zijn twee eigenschappen in een populatie onafhankelijk? Is er een wanverhouding tussen de aantallen mannelijke en vrouwelijke studenten en AP-leden?

De werkwijze

Alvorens de verschillende stappen in de toets te bespreken, formuleren we eerst de voorwaarden waaronder we deze toets kunnen toepassen.

- De data moeten frequenties zijn, geteld in een verzameling disjuncte categorieën/klassen. Percentages kunnen niet gebruikt worden.
- De som van de frequenties moet groter zijn dan 20.
- De verwachte frequentie in elk van de categorieën mag normaal gezien niet kleiner zijn dan 5. Indien dit niet het geval is kun je klassen samenvoegen.
- Alle waarnemingen die bijdragen tot de geobserveerde frequenties, moeten onafhankelijk zijn.

De verschillende stappen

1. **De nulhypothese:** er is geen significant verschil tussen de geobserveerde en de verwachte frequenties.

De alternatieve hypothese zegt dan dat er wel een significant verschil is.

2. Definitie van een toetsingsgrootheid

Veronderstel dat je n waarnemingen hebt, onderverdeeld in k disjuncte klassen. Je beschikt dan over k waargenomen frequenties O_i . Bij elke klasse hoort ook een verwachte frequentie E_i . Er geldt dat

$$n = \sum_i O_i = \sum_i E_i$$

We definiëren nu de volgende toetsingsgrootheid T:

$$T = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

De som loopt over alle klassen.

Deze toetsingsgrootheid is dus de som van de kwadraten van het verschil tussen de geobserveerde en de verwachte frequentie gedeeld door de verwachte frequentie.

3. De verdeling van de toetsingsgrootheid

Voor $i=1, \dots, k$ definiëren we de volgende stochastieken en kansen:

$X_i = \#$ waarnemingen in klasse i

$p_i =$ kans dat een willekeurige waarneming in klasse i ligt onder de gepostuleerde verdeling
dus $p_i = E_i/n$

We kunnen T dan herschrijven als:

$$T = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i}$$

waarbij de stochastieken X_i binomaal verdeeld zijn: $X_i \sim B(n, p_i)$

Men kan bewijzen dat in de limiet voor n gaande naar oneindig T chi-kwadraat verdeeld is met k-1 vrijheidsgraden:

$$\lim_{n \rightarrow \infty} T = X \text{ met } X \sim \chi^2_{k-1} \text{ en dus } \lim_{n \rightarrow \infty} F_T = F_{\chi^2_{k-1}}$$

Bewijs: we zullen dit enkel bewijzen in het geval van 2 klassen.

In dit geval geldt dat $X_1 + X_2 = n$ en $p_1 + p_2 = 1$. We kunnen T dan schrijven als:

$$\begin{aligned} T &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_1 - n(1 - p_1))^2}{n(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2}{n} \left(\frac{1}{p_1} + \frac{1}{1 - p_1} \right) = \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} \end{aligned}$$

Vermits $\sigma^2 = np_1(1 - p_1)$ (Bedenk dat $X_1 \sim B(n, p_1)$) vertelt de centrale limietstelling ons dat:

$$\sqrt{T} \xrightarrow{n \rightarrow \infty} Y \text{ met } Y \sim N(0, 1) \text{ en dus } \lim_{n \rightarrow \infty} F_T = F_{\chi^2_1}$$

4. Het beslissingscriterium

Voor een significantieniveau α zal het aanvaardingsgebied bestaan uit het interval

$$\left[0, \chi^2_{k-1, 1-\alpha} \right]$$

We geven hier eerst een voorbeeld uit de geografie. Nadien zullen we de chi-kwadraat toets toepassen voor het nagaan van een kansverdeling en voor het toetsen van de onafhankelijkheid in kruistabellen.

Voorbeeld

We willen nagaan of het aantal landbouwbedrijven, die graangewassen telen, afhangt van de bodemsamenstelling. (Opgelet voor verdoken factoren zoals subsidies, grootte van het beschouwde gebied,...) De volgende data werden verzameld:

Gebied	bodemtype	Aantal bedrijven
1	Mergel	20
2	Krijt	100
3	Zandsteen	80
4	Klei	20
5	Kalksteen	40

$$n = 260$$

De nulhypothese: De bodemsamenstelling in een bepaald gebied heeft geen effect op het aantal landbouwbedrijven in dat gebied.

Onder de nulhypothese zou de frequentie in elk gebied $n/5$ dus 52 moeten zijn. We kunnen dus de volgende tabel construeren:

Bodemtype	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
Mergel	20	52	19.7
Krijt	100	52	44.3
Zandsteen	80	52	15.1
Klei	20	52	19.7
Kalksteen	40	52	2.8

De toetsingsgroothed is de som van de getallen in de laatste kolom van bovenstaande tabel: $t = 101.6$

In de tabel van de chi-kwadraat verdeling met 4 vrijheidsgraden vinden we voor $\alpha = 0.05$ dat $\chi^2_{k-1,1-\alpha} = 9.49$. Vermits t groter is, kunnen we de nulhypothese op het 5%-significantie niveau verwijderen en besluiten dat we, met een betrouwbaarheid van 95%, mogen stellen dat het bodemtype wel een invloed heeft op het aantal landbouwbedrijven.

6.5.10 a) Toepassing: Toets op een verdeling met geschatte parameters

We willen nagaan of een aantal meetwaarden voldoen aan een bepaalde verdeling. Dit betekent eigenlijk dat we willen nagaan of de getallen een steekproef kunnen zijn uit een populatie met de bewuste verdeling. We illustreren de werkwijze aan de hand van een voorbeeld.

Voorbeeld

We meten de snelheid van 256 auto's op een weg en wensen na te gaan of de snelheid normaal verdeeld is. De metingen zijn samengevat in onderstaande tabel.

snelheidsklasse	aantal
<35	2
35-40	22
40-45	48
45-50	65
50-55	72
55-60	29
60-65	15
>65	3

Omdat er slechts 2 auto's met een snelheid lager dan 35 km/u passeerden hebben we de eerste twee klassen samengevoegd. Hetzelfde geldt voor de laatste twee klassen. We krijgen dan de volgende tabel.

snelheidsklasse	klassemidden	Aantal
<40	35	24
40-45	42.5	48
45-50	47.5	65
50-55	52.5	72
55-60	57.5	29
>60	65	18

Uit de tabel kunnen we een gemiddelde snelheid berekenen van 49.16 km/u en een standaarddeviatie van 7.53 km/u.

We willen nu verifiëren of de gemeten snelheden normaal verdeeld zijn. We zullen de gemeten waarden voor gemiddelde en standaardafwijking als parameters van onze verdeling gebruiken. Met deze waarden tabelleren we de gepostuleerde verdelingsfunctie F in de grenspunten van onze intervallen:

$$F(40) = \Phi\left(\frac{40 - 49.16}{7.53}\right) = 0.1119$$

$$F(45) = \Phi\left(\frac{45 - 49.16}{7.53}\right) = 0.2903$$

$$F(50) = \Phi\left(\frac{50 - 49.16}{7.53}\right) = 0.5444$$

$$F(55) = \Phi\left(\frac{55 - 49.16}{7.53}\right) = 0.7810$$

$$F(60) = \Phi\left(\frac{60 - 49.16}{7.53}\right) = 0.9250$$

Met deze waarden kun je de verwachte frequenties voor de verschillende klassen berekenen. We krijgen:

snelheidsklasse	klassemidden	O_i	p_i	E_i
<40	35	24	0.1119	28.6
40-45	42.5	48	0.1784	45.7
45-50	47.5	65	0.2541	65.0
50-55	52.5	72	0.2366	60.6
55-60	57.5	29	0.1440	36.9
>60	65	18	0.0750	19.2

We kunnen weer de toetsingsgrootheid T uitrekenen en bekomen 4.7813. Dit getal moeten we vergelijken met $\chi^2_{3,0.95} = 7.815$. We aanvaarden de nulhypothese op niveau 5%. De overschrijdingskans of p-waarde voor dit voorbeeld is: $p = P(T > 4.7813) = 0.1885$ voor een chi-kwadraat verdeling met 3 vrijheidsgraden. Bemerk dat we hier een chi-kwadraat verdeling met drie vrijheidsgraden gebruikten. We hebben ons totale gebied in 6 deelintervallen opgedeeld en we hebben twee verdelingsparameters uit de data geschat. Indien we de parameters van onze verdeling niet hadden moeten schatten hadden we 2 vrijheidsgraden meer gehad. We zouden de nulhypothese dan ook nog aanvaarden indien T een waarde kleiner dan 11.070 aanneemt.

6.5.10 b) Toepassing: Kruistabellen (Engels: contingency tables or crosstabs)

Indien we over telgegevens beschikken, die volgens twee (kwalitatieve) variabelen zijn geklassificeerd, kunnen we deze gemakkelijk in een tabel weergeven. We noemen deze tabel een kruistabel.

We vragen ons nu af of er een relatie bestaat tussen de kolomvariabele en de rijvariabele. We illustreren de werkwijze voor het geven van een antwoord op deze vraag met behulp van het onderstaande voorbeeld.

Voorbeeld

Doen mannen en vrouwen om dezelfde reden aan sport ? Eén motivatie voor sportbeoefening is competitiegeest – het verlangen om te winnen of beter te zijn dan de andere. Een andere is meesterschap – het verlangen zijn vaardigheden te verbeteren of zich tot het uiterste in te spannen. Bij een studie naar redenen waarom studenten aan sport doen, werden data verzameld van 69 mannelijke en 65 vrouwelijke studenten. Elke student werd, op basis van zijn/haar reacties op een vragenlijst over sportmotivaties, ondergebracht in één van de vier categorieën: hoge competitiegeest – hoog meesterschap (HC – HM); hoge competitiegeest – laag meesterschap (HC – LM); lage competitiegeest – hoog meesterschap (LC – HM); lage competitiegeest – laag meesterschap (LC – LM). De data van de steekproef zijn voorgesteld in onderstaande tabel.

Motivatie	Vrouw	Man	Rijsom
HC – HM	14	31	45
HC – LM	7	18	25
LC – HM	21	5	26
LC – LM	23	15	38
Kolomsom	65	69	Tot=134

In de tabel hebben we ook de totalen uitgerekend.

Elke combinatie van waarden voor de twee variabelen definieert een cel. Een tabel met r rijen en c kolommen bevat $r \times c$ cellen. We noemen deze tabel dan een $r \times c$ kruistabel. Wij hebben hier dus te maken met een 4×2 tabel met 8 cellen. In dit voorbeeld is het de bedoeling mannen en vrouwen met elkaar te vergelijken. De kolomvariabele beschrijft uit welke populatie de waarneming komt. De rijvariabele is een kwalitatieve variabele, het type sportmotivatie. Het is niet altijd het geval dat één richting van de tabel de te vergelijken populaties identificeert. Kruistabellen kunnen voor een willekeurig tweetal kwalitatieve variabelen de waarnemingen weergeven.

De nulhypothese

H_0 : Er is geen afhankelijkheid tussen de rij- en de kolomvariabele of er bestaat geen afhankelijkheid tussen geslacht en sportmotivatie.

H_A : Er is wel een afhankelijkheid tussen de 2 variabelen motivatie en geslacht

De verwachte frequenties onder de nulhypothese:

Met behulp van de tabel kunnen we nu de fracties mannen ($p_M = 69/134$) en vrouwen ($p_V = 65/134$) en de fracties HC-HM ($p_{HH} = 45/134$), HC-LM ($p_{HL} = 25/134$), LC-HM ($p_{LH} = 26/134$), LC-LM ($p_{LL} = 38/134$) uitrekenen.

Onder de nulhypothese moet voor de frequenties in de verschillende groepen gelden dat ($n=134$):

Motivatie	Vrouw	Man	Rijsom
HC – HM	$n p_V p_{HH} = 21.83$	$n p_M p_{HH} = 23.17$	45
HC – LM	$n p_V p_{HL} = 12.13$	$n p_M p_{HL} = 12.87$	25
LC – HM	$n p_V p_{LH} = 12.61$	$n p_M p_{LH} = 13.39$	26
LC – LM	$n p_V p_{LL} = 18.43$	$n p_M p_{LL} = 19.57$	38
Kolomsom	65	69	Tot=134

$$\text{De toetsingsgrootte } T \text{ wordt dan } T = \sum_i \frac{(O_i - E_i)^2}{E_i} = 22.69$$

Met de chi-kwadraat toets kunnen we nagaan of de geobserveerde frequenties “voldoende” overeenstemmen met de verwachte frequenties. Bemerk dat we hier 4 parameters hebben moeten schatten: $p_V, p_{HH}, p_{HL}, p_{LH}$ (NB met de kennis van deze parameters kunnen we p_M en p_{LL} berekenen). Vermits de som van alle frequenties n moet zijn houden we dus nog $7-4=3$ vrijheidsgraden over. In het algemeen hebben we voor een $r \times c$ kruistabel $(r-1)*(c-1)$ vrijheidsgraden.

Het aanvaardingsgebied op het 5%-niveau is [0,7.815] en op het 1%-niveau [0,11.345]. We verwerpen de nulhypothese dus met een betrouwbaarheid van meer dan 99%; om precies te zijn: de overschrijdingskans of significantie is kleiner dan 0.0005.

N.B.: Afrondingen hebben een groot effect op deze toetsingsgrootte. Neem dus genoeg cijfers na de komma mee, zeker indien de berekende waarde van de toetsingsgrootte op de grens tussen het aanvaardings- en het verwerpingsgebied ligt.

6.5.11 De Kolmogorov-Smirnov toets

We willen aan de hand van metingen $\{x_1, x_2, \dots, x_n\}$, onafhankelijke trekkingen uit een stochastische variabele X , toetsen of de theoretische verdelingsfunctie van X gelijk is aan F . We vragen ons bijvoorbeeld af of de getallen $\{x_1, x_2, \dots, x_n\}$ trekkingen uit een normale verdeling zijn. Een mogelijke werkwijze is: deel de data in klassen en voer vervolgens een chi-kwadraat toets uit. Een andere, meerelegante methode, die rechtstreeks de grafieken van de empirische verdelingsfunctie F_n met de theoretische verdelingsfunctie F vergelijkt, stamt van Kolmogorov en Smirnov. We zullen ze hier illustreren. Deze toets is vooral handig bij kleine steekproeven, omdat in dit geval de chi-kwadraat toets minder aangewezen is (ihb het indelen in klassen met voldoende elementen x_i is praktisch onmogelijk indien men toch een voldoende aantal klassen wil behouden).

Zij $\{x_1, x_2, \dots, x_n\}$ onafhankelijke trekkingen uit een stochastische variabele X , dan is de empirische verdelingsfunctie:

$$F_n(x) = \frac{\text{aantal}\{x_i \leq x\}}{n} = \begin{cases} 0 & \text{als } x < y_1 \\ \frac{k}{n} & \text{als } y_k \leq x < y_{k+1} \\ 1 & \text{als } x \geq y_n \end{cases}$$

waarbij $\{y_1, y_2, \dots, y_n\}$ de gesorteerde data $\{x_1, x_2, \dots, x_n\}$ zijn.

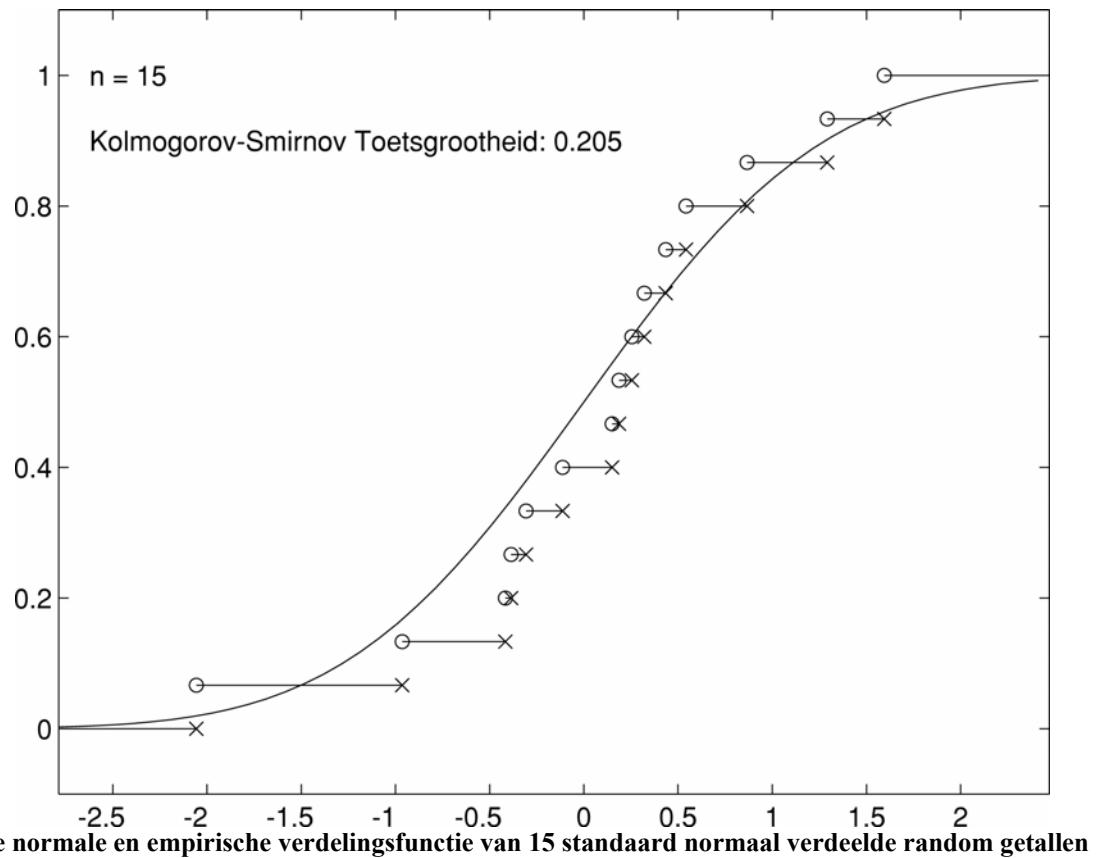
We kijken nu naar het verschil tussen deze functie en de gepostuleerde theoretische verdelingsfunctie F . De maximale waarde van dit verschil wordt gegeven door:

$$d_n = \max_x |F_n(x) - F(x)| = \max \left(\max_k \left| \frac{k}{n} - F(y_k) \right|, \max_k \left| \frac{k-1}{n} - F(y_k) \right| \right)$$

H_0 : De verdelingsfunctie van $X = F$

H_A : De verdelingsfunctie van $X \neq F$

Toetsingsgrootte: d_n



We zullen de verdeling van deze toetsingsgrootte hier niet afleiden. Intuïtief: d_n mag niet te groot zijn. We verwijzen de nulhypothese als d_n groter is dan de kritische waarde volgens onderstaande tabel, waarbij n de steekproefgrootte is.

Kritische grenzen voor de Kolmogorov-Smirnov toets

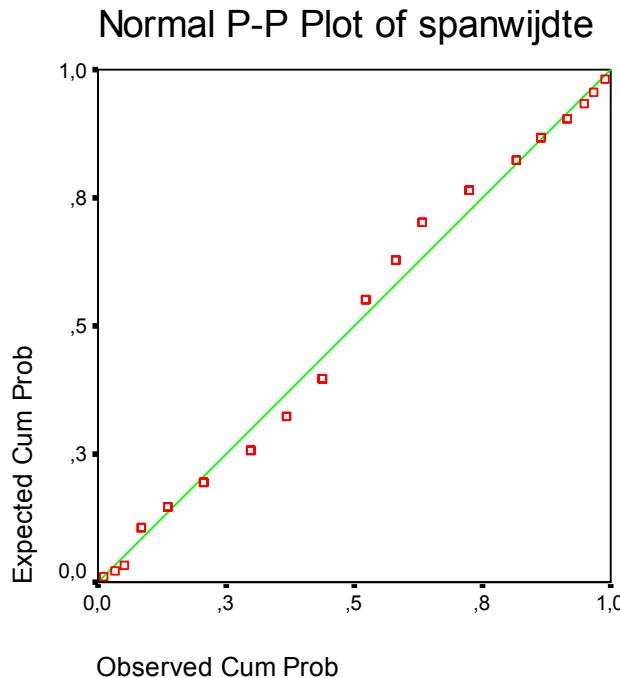
n	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	0.90	0.95	0.98	0.99
2	0.68	0.78	0.84	0.93
3	0.56	0.64	0.71	0.83
4	0.49	0.56	0.62	0.73
5	0.45	0.51	0.56	0.67
6	0.41	0.47	0.52	0.62
7	0.38	0.44	0.49	0.58
8	0.36	0.41	0.46	0.54
9	0.34	0.39	0.43	0.51
10	0.32	0.37	0.41	0.49
12	0.30	0.34	0.38	0.45
15	0.27	0.30	0.34	0.40
20	0.23	0.26	0.29	0.35
25	0.21	0.24	0.26	0.32
30	0.19	0.22	0.24	0.29
35	0.18	0.21	0.23	0.27
40	0.17	0.19	0.21	0.25
45	0.16	0.18	0.20	0.24
	1.07	1.22	1.36	1.63
	\sqrt{n}	\sqrt{n}	\sqrt{n}	\sqrt{n}

De laatste lijn geeft een benadering voor grote waarden van n

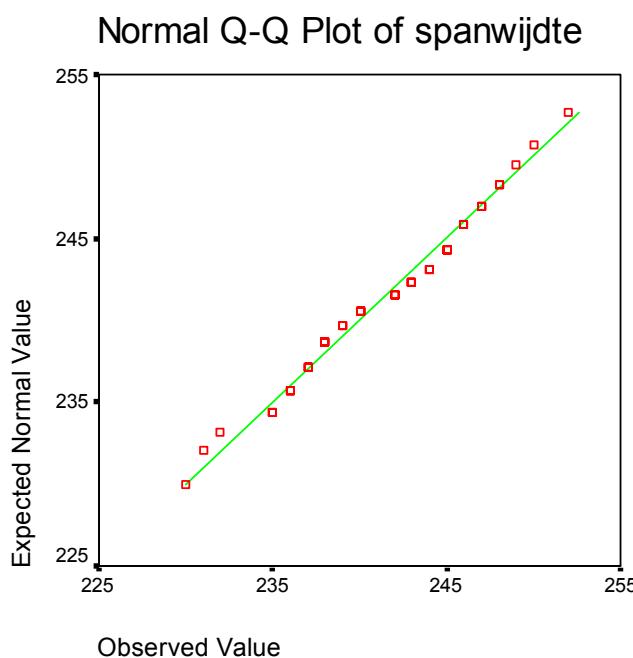
Een alternatieve grafische methode is het maken van een P-P-plot of Q-Q-plot..

In een P-P-plot worden de theoretische kansen $F(y_i)$ uitgezet tegen de empirische $F_n(y_i)$. Indien beide verdelingsfuncties gelijk zijn, zullen de punten dicht bij de rechte $y = x$ (bissectrice van het eerste kwadrant) liggen. De toetsingsgrootheid zal dan klein zijn en we mogen de nulhypothese aanvaarden.

Hieronder wordt de P-P-plot gegeven om de normale verdeling van de spanwijdte in de Bumpus dataset na te gaan.



In een Q-Q-plot daarentegen worden de theoretische kwantielen $F^{-1}\left(\frac{i}{n}\right)$ uitgezet tegen de empirische waarden y_i . Om de nulhypothese te aanvaarden moeten ook deze punten dicht bij de rechte $y = x$ liggen.



Bemerk dat de P-P-plot naar de verticale afwijkingen tussen de verdelingsfuncties F_n en F kijkt terwijl de Q-Q-plot de horizontale afwijkingen weergeeft. Beide plots zijn dan ook op een andere schaal: de P-P-plot varieert (op de X- en Y-as) tussen 0 en 1; de Q-Q-plot heeft dezelfde variatie als de oorspronkelijke data.

N.B.: Bij het maken van deze plots zal een empirische verdelingsfunctie F_n gebruikt worden waarbij men deelt door $n+1$ in plaats van n . We gaan er hier niet verder op in.

6.5.12 De macht van een toets

Bij het toetsen zijn er twee types fouten die we kunnen maken. Ze worden in het onderstaande diagram nog eens samengevat.

	Werkelijkheid H_0 is waar	H_A is waar
H_0 wordt verworpen	Type I fout kans α	Correcte uitspraak kans $1-\beta$
H_0 wordt niet verworpen	Correcte uitspraak kans $1-\alpha$	Type II fout kans β

Een type I fout, het verwerpen van de nulhypothese terwijl deze waar is, gebeurt met kans $\alpha = P(H_0 \text{ wordt verworpen} | H_0 \text{ is waar})$.

Een type II fout, het aanvaarden van de nulhypothese terwijl deze onwaar is, gebeurt met kans $\beta = P(H_0 \text{ wordt niet verworpen} | H_0 \text{ is onwaar})$.

De kans op het maken van een fout van de tweede soort is afhankelijk van de werkelijke waarde van de onbekende parameter.

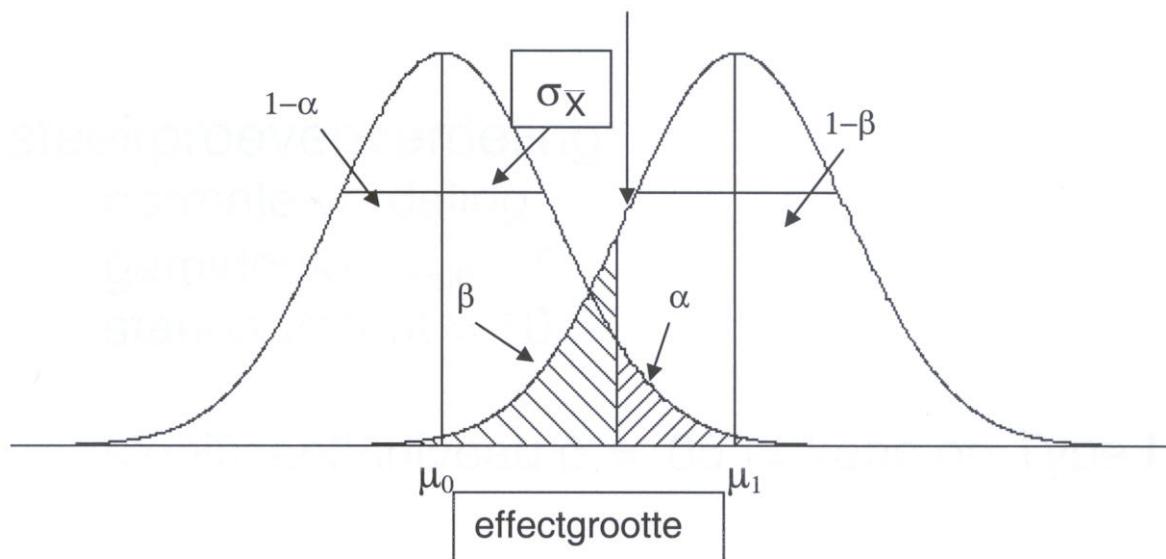
We noemen $1-\beta$ de macht of het onderscheidingsvermogen van een toets (Eng: power). Dit getal geeft de kans weer op het ontdekken van een echt alternatief.

We zullen nu de macht van een eenzijdige t-toets berekenen.

De hypotheses

$$H_0: \mu = \mu_0$$

$$H_A: \mu = \mu_1 > \mu_0$$



De linkerkromme schetst de verdeling onder H_0 en de rechter de verdeling onder H_A . De scheidingslijn tussen de twee gearceerde gebieden is getrokken bij het kritisch punt, bepaald door α . $\sigma_{\bar{X}}$ staat voor s_n .

De toetsingsgroothed is $T = \frac{\bar{X}_n - \mu_o}{S_n / \sqrt{n}}$ met verdeling t_{n-1} onder de nulhypothese.

Indien de schatting t van T , berekend met de waarden van de steekproef, in het verwijdingsgebied $[t_{n-1,1-\alpha}, +\infty)$ ligt, wordt de nulhypothese verworpen op significantieniveau α .

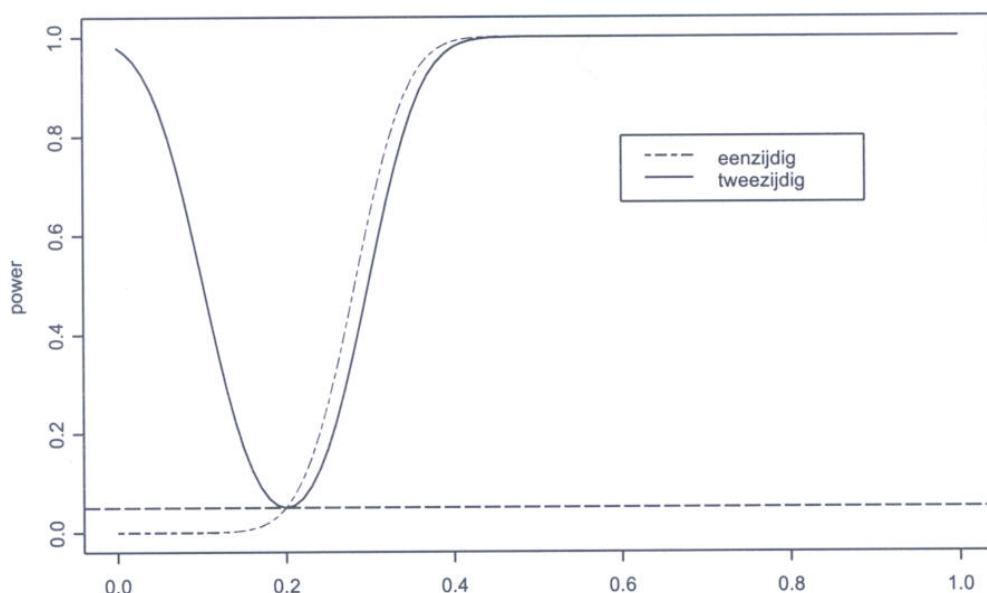
H_0 wordt dus verworpen als het steekproefgemiddelde $\bar{x}_n > \mu_o + t_{n-1,1-\alpha} \frac{s_n}{\sqrt{n}}$.

De macht van de toets = $1 - \beta = P(H_0 \text{ wordt verworpen} | H_A)$

$$= P\left(\bar{x}_n > \mu_o + t_{n-1,1-\alpha} \frac{s_n}{\sqrt{n}} | H_A\right)$$

Onder H_A geldt dat : $\frac{\bar{X}_n - \mu_1}{S_n / \sqrt{n}} \sim t_{n-1}$. We kunnen de bovenstaande kans dus ook berekenen met de verdelingsfunctie van de t-verdeling met $n-1$ vrijheidsgraden $(F_{t_{n-1}})$.

$$\begin{aligned} \text{De macht van de toets} &= 1 - \beta = 1 - F_{t_{n-1}}\left(\frac{\mu_o - \mu_1 + t_{n-1,1-\alpha} \frac{s_n}{\sqrt{n}}}{\frac{s_n}{\sqrt{n}}}\right). \\ &= F_{t_{n-1}}\left(-t_{n-1,1-\alpha} + \frac{\mu_1 - \mu_o}{\frac{s_n}{\sqrt{n}}}\right) \end{aligned}$$



Macht van de t-toets (zowel eenzijdig als tweezijdig) afgebeeld in functie van μ_1 (bemerk dat $\mu_o=0.2$)

De macht van de toets stijgt dus met stijgende n , met stijgende α en met dalende σ (geschat door s_n). We zien tevens dat de macht afhankelijk is van de alternatieve waarde μ_1 ; hoe groter het verschil $\mu_1 - \mu_0$, hoe groter de macht van de toets.

We kunnen nu de vereiste steekproefgrootte n berekenen, waarmee we een alternatieve waarde μ_1 met een bepaalde macht $1 - \beta$ kunnen detecteren bij een significantieniveau α .

$$1 - \beta = F_{t_{n-1}} \left(-t_{n-1,1-\alpha} + \frac{\mu_1 - \mu_0}{\frac{s_n}{\sqrt{n}}} \right)$$

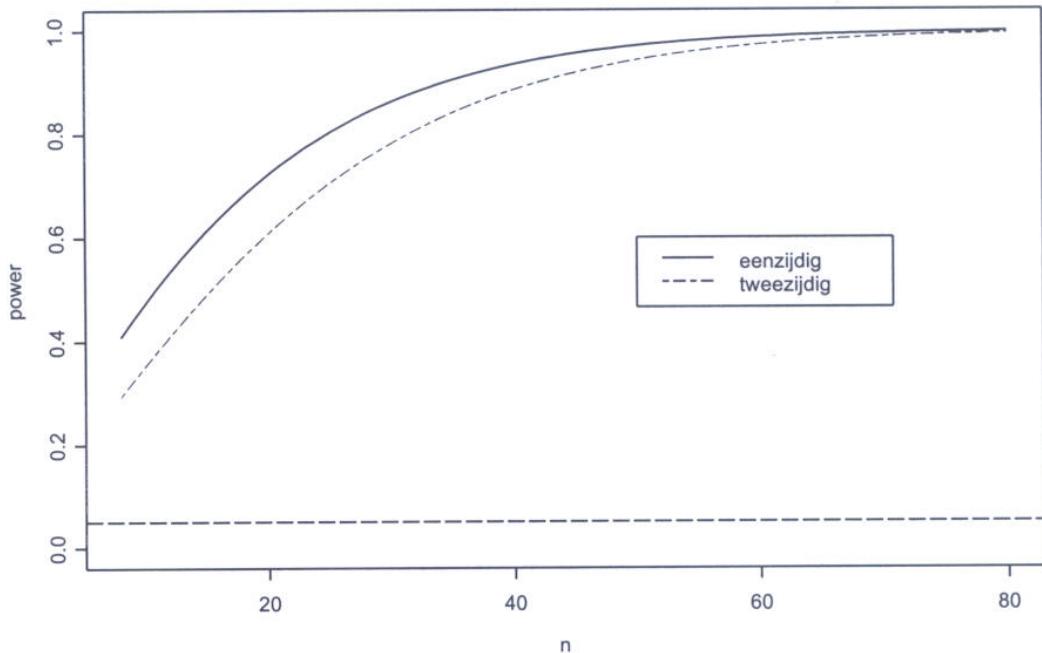
of

$$t_{n-1,1-\beta} = -t_{n-1,1-\alpha} + \frac{\mu_1 - \mu_0}{\frac{s_n}{\sqrt{n}}}$$

of

$$n = \frac{(t_{n-1,1-\beta} + t_{n-1,1-\alpha})^2 s_n^2}{(\mu_1 - \mu_0)^2}$$

Om een kans op een type II fout te hebben die kleiner is dan β , moet de grootte van onze steekproef minstens de bovenstaande waarde van n zijn.



Macht van een t-toets (zowel eenzijdig als tweezijdig) afgebeeld in functie van n
voor $\mu_0=25$, $\mu_1=20$ en $s_n=10$.

Een analoge redenering kan gevuld worden voor het berekenen van de macht van de tweezijdige t-toets en voor de andere toetsen. We gaan er hier niet verder op in.

Hoofstuk 7: Correlatie en regressie

7.1 Correlatie

We herhalen enkele begrippen.

Zij X en Y twee stochastieken

De covariantie van twee stochastische variabelen X en Y wordt gegeven door:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

De correlatiecoëfficiënt van twee stochastische variabelen X en Y wordt gegeven door:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

De correlatiecoëfficiënt is begrensd: $-1 \leq \rho \leq 1$

De waarde van ρ zal dicht bij 1 liggen indien er een sterk positief lineair verband is tussen X en Y. De waarde van ρ zal dicht bij -1 liggen indien er een sterk negatief lineair verband is tussen X en Y.

Zij $\{(x_i, y_i) | i = 1, \dots, n\}$ een verzameling van geordende paren waarbij $\{x_i | i = 1, \dots, n\}$ en $\{y_i | i = 1, \dots, n\}$ twee series van n metingen zijn (steekproeven uit X en Y) met gemiddelden \bar{x} respectievelijk \bar{y} en standaarddeviaties s_x respectievelijk s_y , dan zijn:

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{en} \quad r(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{s_x s_y}$$

de steekproefcovariantie en steekproefcorrelatiecoëfficiënt.

Een grote correlatiecoëfficiënt wil nog niet zeggen dat er een causaal verband is tussen beide variabelen. We kunnen ons voorstellen dat de correlatiecoëfficiënt van het aantal verkochte liter frisdrank en het aantal verdrinkingsdoden positief is. Toch heeft het drinken van een frisdrank geen invloed op de kans om te verdrinken. Er is een verdoken factor, namelijk de temperatuur, waar beide variabelen wel een verband mee hebben.

Om te testen of we op basis van onze steekproef mogen besluiten, met een bepaald significantieniveau, dat er geen correlatie is tussen X en Y voeren we een t-toets uit. Deze toets kan enkel uitgevoerd worden indien X en Y normaal verdeeld zijn.

$$H_0: \rho = 0$$

$$H_A: \rho \neq 0: \text{er is een significant lineair verband tussen X en Y in de populatie}$$

$$\text{Toetsingsgrootte: } T = \rho \sqrt{\frac{n-2}{1-\rho^2}}$$

De toetsingsgrootte T heeft een t-verdeling met $n-2$ vrijheidsgraden. We bewijzen dit hier niet.

We berekenen eerst de correlatiecoëfficiënt r en nadien de toetsingsgrootte met de waarden van onze steekproef $\rightarrow t = r \sqrt{\frac{n-2}{1-r^2}}$

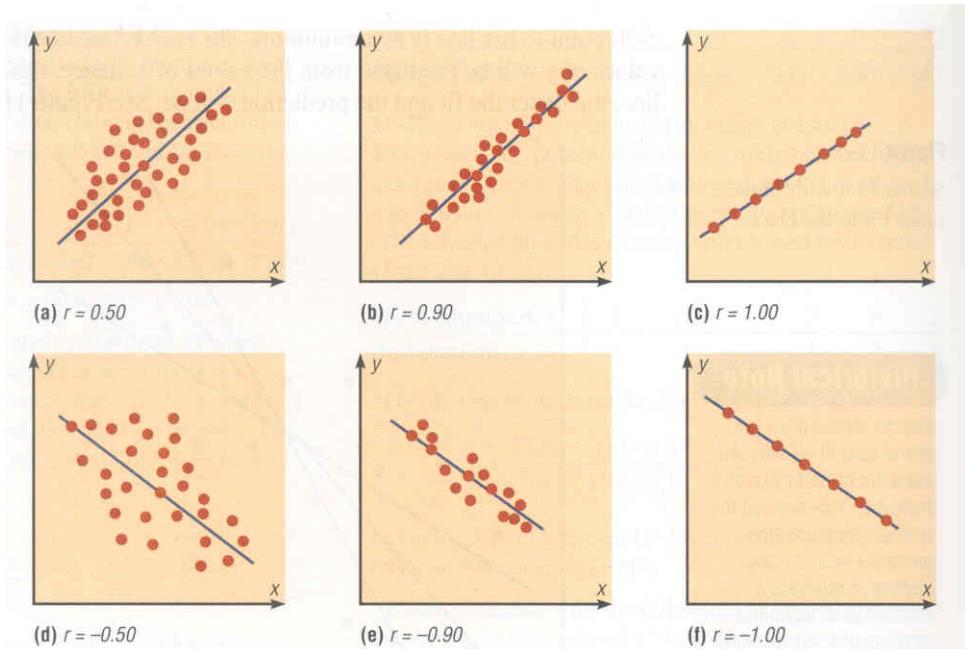
Het aanvaardingsgebied is $\left[t_{n-2, \frac{\alpha}{2}}, t_{n-2, 1-\frac{\alpha}{2}} \right]$

De kritische waarde of p-waarde is $p = 2F_{t_{n-2}}^{-1}(-|t|)$

7.2 Enkelvoudige lineaire regressie

7.2.1 Inleiding tot lineaire regressie en de kleinste kwadraten methode

Bij het bestuderen van verbanden tussen twee variabelen zullen we eerst metingen uitvoeren. Dit levert ons $\{(x_i, y_i) | i = 1, \dots, n\}$; een verzameling van geordende paren. We kunnen deze datapunten voorstellen in een scatterplot. Elk punt in deze plot stelt dan een datakoppel voor. Het doel van deze scatterplot bestaat erin grafisch een idee te krijgen van de aard van het verband tussen X en Y. Is dit verband lineair, kwadratisch, exponentieel,...?



Nadien bepalen we best de correlatiecoëfficiënt. Indien de correlatiecoëfficiënt significant is, kunnen we een lineair model opstellen. We gaan door onze meetpunten (x_i, y_i) een rechte fitten. Indien we de parameter Y gemeten hebben voor verschillende waarden van X, noemen we X de onafhankelijke variabele en Y de afhankelijke of respons variabele. Natuurlijk kunnen beide variabelen gemeten waardes zijn. In dat geval is het best een goede keuze te

maken van de X en de Y variabele. Het fitten van een rechte voor y in functie van x zal een iets andere rechte opleveren dan het fitten van een rechte voor x in functie van y.
De regressierechte zal ons toelaten voorspellingen te maken en een eventuele trend in de data te observeren. Ook voor calibratiedoelen zal men een regressierechte opstellen.

Indien we veronderstellen dat de echte relatie tussen de respons en de onafhankelijke variabele een rechte is krijgen we het volgende model:

$$\eta = \beta_0 + \beta_1 x$$

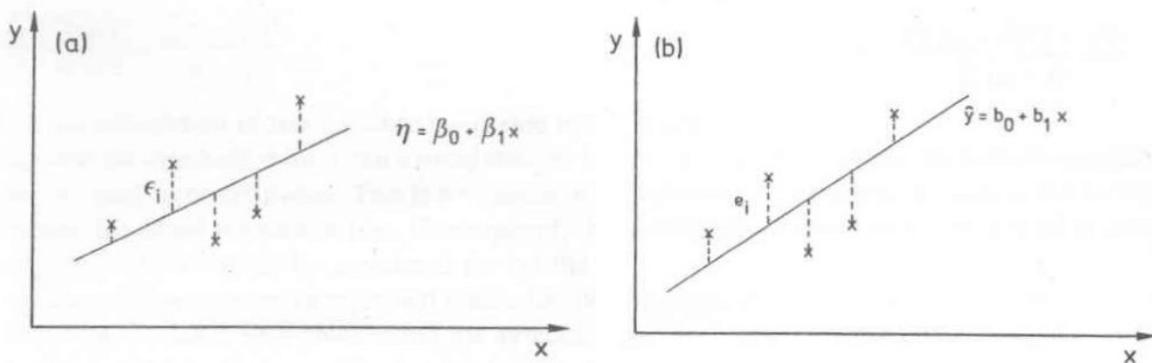
waarbij η de echte respons voorstelt bij een waarde x van X. β_0 en β_1 zijn de modelparameters. Zij zijn respectievelijk het snijpunt met de Y-as en de richtingscoëfficiënt van de regressierechte. Bovenstaande rechte is de regressierechte van de populatie.
Vermits metingen onderhevig zijn aan fouten zullen de metingen bijna nooit op de rechte liggen. Elk meetpunt (x_i, y_i) voldoet aan:

$$y_i = \eta_i + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Elke meting y_i is dus opgebouwd uit een component bepaald door het lineair model en een component ϵ_i , het verschil tussen de geobserveerde respons y_i en de ware respons η_i . De modelparameters β_0 en β_1 zijn onbekend. Met behulp van de meetpunten zullen deze parameters geschat worden door b_0 en b_1 . Deze schattingen zullen zodanig berekend worden dat de rechte

$$\hat{y} = b_0 + b_1 x$$

de n experimentele punten zo goed mogelijk fit.



Hoe gaan we nu de beste rechte door de meetpunten bepalen? Welk criterium gaan we gebruiken ? Hiervoor bestaan natuurlijk verschillende mogelijkheden. Wij gebruiken hier de kleinste kwadraten methode, die erin bestaat de som van de kwadraten van de verticale verschillen e_i tussen de punten en de rechte te minimaliseren (zie bovenstaande figuur). Het residu e_i is dus het verschil tussen het meetpunt y_i en de waarde \hat{y}_i , voorgespeld door de regressierechte:

$$e_i = y_i - \hat{y}_i$$

7.2.2 De methode

De getallen b_0 en b_1 worden bekomen door het minimiseren van de volgende som:

$$R = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_o - b_1 x_i)^2$$

waarbij de som over alle meetpunten ($i=1,\dots,n$) loopt (we zullen in het vervolg de indexen bij het somteken enkel vermelden indien de som voor i niet loopt van 1 tot n).

R is een som van kwadraten en dus altijd positief. Minimalisatie van R gebeurt door het afleiden van de som enerzijds naar b_o en anderzijds naar b_1 en de 2 verkregen uitdrukkingen gelijk te stellen aan 0. Dit levert het volgende stelsel van 2 vergelijkingen in de onbekende parameters b_o en b_1 :

$$\begin{cases} \frac{\partial R}{\partial b_o} = \sum 2(y_i - b_o - b_1 x_i)(-1) = 0 \\ \frac{\partial R}{\partial b_1} = \sum 2(y_i - b_o - b_1 x_i)(-x_i) = 0 \end{cases}$$

We kunnen dit laatste herschrijven als:

$$\begin{cases} \sum y_i - nb_o - b_1 \sum x_i = 0 \\ \sum x_i y_i - b_o \sum x_i - b_1 \sum x_i^2 = 0 \end{cases}$$

Deze vergelijkingen worden de **normaal vergelijkingen** genoemd, die resulteren in de volgende uitdrukkingen voor de kleinste kwadraten schatters b_o en b_1 .

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_o = \bar{y} - b_1 \bar{x}$$

waarbij \bar{x} en \bar{y} respectievelijk het gemiddelde van alle x_i 's en het gemiddelde van alle y_i 's zijn: $\bar{x} = \frac{\sum x_i}{n}$ en $\bar{y} = \frac{\sum y_i}{n}$

Bemerk dat b_1 gegeven wordt door: $b_1 = \frac{Cov(x, y)}{s_x^2}$. Indien de mate van het lineair verband klein is zal de covariantie klein zijn en zal de regressierechte praktisch horizontaal zijn; hoe kleiner de covariantie, hoe kleiner de richtingscoëfficiënt en hoe minder relevant de regressie.

Een belangrijke statistische grootheid in de regressieanalyse is de residuele variantie:

$$s_e^2 = \frac{\sum (e_i)^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} \quad (s_e = \text{standard error of the estimate})$$

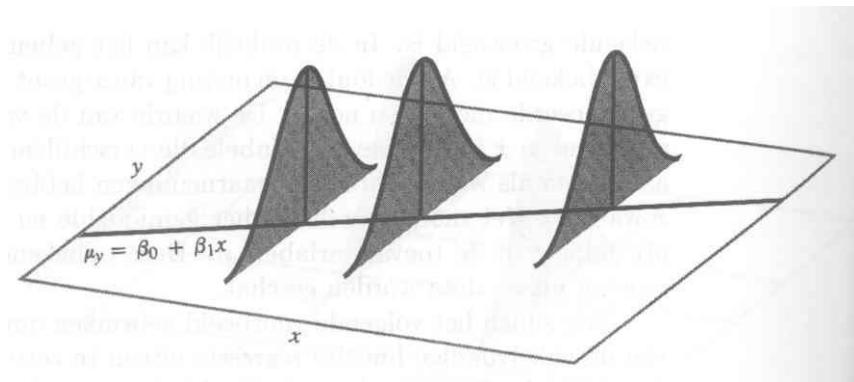
Voor het berekenen van deze variantie delen we door $n-2$, omdat we 2 parameters hebben moeten schatten voor het berekenen van de rechte, namelijk b_o en b_1 . Deze grootheid geeft een maat voor de spreiding van de punten rond de rechte. Het stelt dus dat deel van de variantie voor, dat niet door onze regressierechte wordt beschreven. Indien het lineaire model correct is, is s_e^2 een schatting van de variantie op de metingen, ook wel de pure experimentele fout genoemd. Deze grootheid zal gebruikt worden bij de constructie van de betrouwbaarheidsintervallen voor de parameters en de predicties (zie verder).

7.2.3 De voorwaarden

Om betrouwbaarheidsintervallen op te stellen voor onze geschatte parameters en voor de voorspellingen gebaseerd op de regressierechte evenals voor het testen van hypothesen aangaande deze parameters, moeten de residus e_i aan de volgende voorwaarden voldoen:

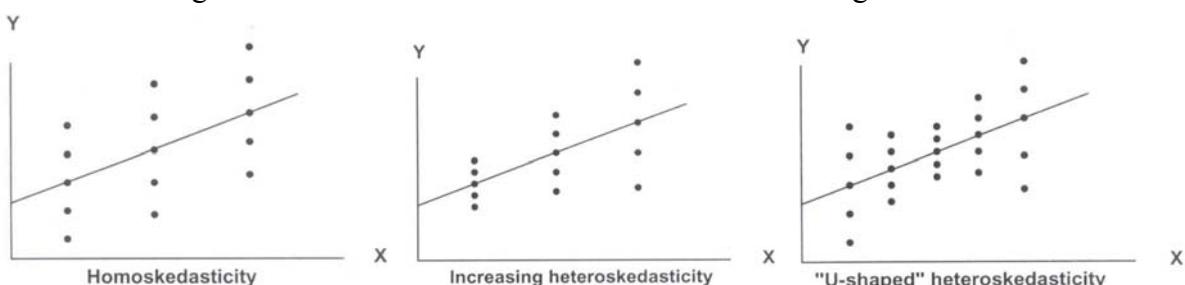
- voor elke x_i is het residu een trekking uit een normaal verdeelde stochastiek met gemiddelde nul
- de residus zijn onafhankelijk
- de residus bij verschillende waarden van x_i hebben dezelfde variantie

Deze voorwaarden zijn grafisch weergegeven in onderstaande grafiek.



In vele experimentele situaties is de experimentele fout een som van vele kleinere onafhankelijke fouten. Het is dan redelijk te veronderstellen dat de experimentele fout normaal verdeeld is (de normaliteit is een gevolg van de centrale limietstelling).

De laatste voorwaarde wordt de homoscedasticiteitsvoorwaarde genoemd. Aan deze voorwaarde is niet altijd voldaan. We spreken dan van heteroscedasticiteit (verschillende varianties bij verschillende x-waarden: zie figuur). Om een idee van de variantie van de metingen bij een vaste x te krijgen zal men meerdere metingen moeten doen bij deze x-waarde. In vele gevallen echter zullen we niet over herhaalde metingen beschikken.



Het gebeurt meer dan eens dat de variantie evenredig is met de x of y (zie tweede figuur); grotere meetwaarden hebben dan een grotere fout dan kleinere meetwaarden.

In gevallen waarbij men een transformatie van de oorspronkelijke variabele(n) moet uitvoeren om een lineair verband te krijgen, zullen de getransformeerde data, zelfs als de oorspronkelijke metingen een gelijke variantie hebben, niet homoscedastisch zijn. (We merken op dat we in dit geval best met de oorspronkelijke homoscedastische data werken, zelfs als het verband niet lineair is. We zullen dan niet-lineaire regressie technieken gebruiken.) In dit geval kent men de variantiefunctie: de manier waarmee de variantie verandert. Indien men bij elke waarde van x de variantie van de meetfout kent, kan men de gewogen kleinste kwadraten methode gebruiken: men gaat aan metingen met een kleinere variantie meer belang (gewicht) hechten dan aan metingen met een grotere variantie. Ook blijkt een transformatie

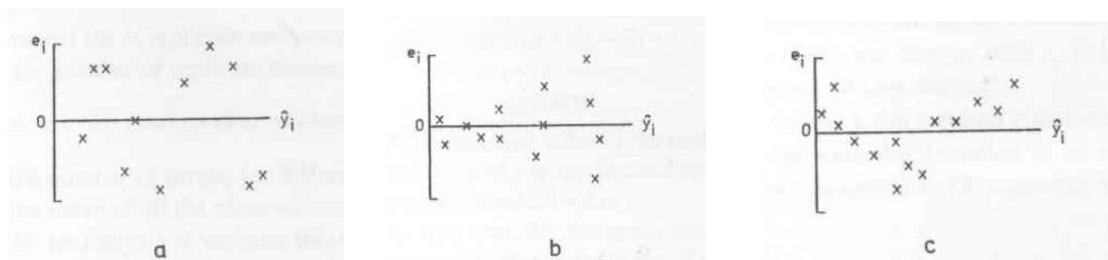
van de data soms een oplossing te bieden (men kan in plaats van y bijvoorbeeld $\ln(y)$ als variabele gebruiken). We gaan er hier niet verder op in.

7.2.4 Validatie van het model

Na het opstellen van de regressierechte is het noodzakelijk het model te verifiëren: is het model een rechte of worden de data beter beschreven door een kromme? Is er "lack of fit"? Tevens moeten we nagaan of de residus voldoen aan de voorwaarden van normaliteit en homoscedasticiteit. De studie van de residus zal ons ook informatie verschaffen aangaande de lack of fit (LOF).

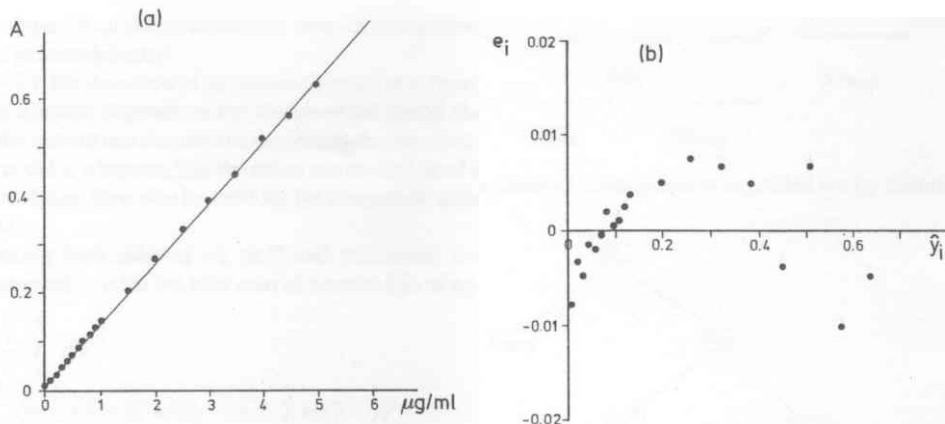
7.2.4.a Scatterplot van gefitte data en de residuele plot

Om de voorwaarde van normaliteit van de residus na te gaan moeten we voor elke x -waarde een KS-toets of een chi-kwadraat toets uitvoeren. Om dit te kunnen doen moeten we echter over een groot aantal herhaalde metingen beschikken. Dit zal zoals reeds eerder vermeld meestal niet het geval zijn. Residuele plots zijn dan zeer informatief. In een residuele plot wordt het residu e_i uitgezet in functie van x_i en/of \hat{y}_i . Deze plot moet je ALTIJD maken. Hij geeft ook nuttige informatie aangaande de LOF. Voorbeelden van residuele plots worden hieronder gegeven.



Voor een goede regressie zullen de residus random verspreid liggen in een horizontale band rond 0 met ongeveer evenveel positieve als negatieve residus (figuur a). In figuur b zie je duidelijk dat de voorwaarde van homoscedasticiteit niet vervuld is. De spreiding van de punten stijgt met een stijgende waarde van \hat{y}_i . Het model daarentegen lijkt op het eerste zicht correct omdat de X-as ongeveer een symmetrieas van de puntenwolk is. De U-vorm van de residus in figuur c duidt op een LOF. De data worden beter voorgesteld door een kromme (bijvoorbeeld door een kwadratische vergelijking).

In onderstaande grafiek wordt een calibratierchte voorgesteld. Op het eerste zicht lijkt de lineaire fit perfect. Indien we echter een residuele plot maken, zien we dat de residus niet random verspreid liggen. Er is een zeker patroon herkenbaar. We kunnen ons model waarschijnlijk verbeteren door er nog hogere orde termen (termen in x^2, x^3, \dots) aan toe te voegen. De lineaire fit heeft reeds een groot gedeelte van de variatie in de data verklaard, maar we kunnen ons model nog verfijnen. Bemerk dat we het voorgaande zonder het maken van een residuele plot niet zouden gemerkt hebben. Vermits de schaal van de residuele plot veel kleiner is dan deze van de oorpronkelijke dataplot, is het gemakkelijker een patroon in de residuele plot op te merken. (Er bestaan natuurlijk ook toetsen om dit na te gaan.)



Indien we willen nagaan of een bijkomende parameter een rol speelt in de respons, maken we een scatterplot van de residus e_i in functie van de bewuste parameter. Het is in die context ook nuttig een scatterplot te maken van e_i in functie van het tijdstip (of de volgorde) waarop de metingen uitgevoerd werden. Indien deze plot niets abnormaals toont, kunnen we besluiten dat de bewuste parameter of in het tweede geval de tijd geen rol speelt. Indien de laatste scatterplot een zekere trend laat zien is er een drift in de data en zijn de metingen waardeloos.

7.2.4.b De variantieanalyse (ANalysis Of VAriance = ANOVA)

In het bovenstaande voorbeeld hadden we een LOF. Toch verklaarde de regressierechte reeds een groot gedeelte van de variatie in onze gegevens. Hoe kunnen we nu nagaan of onze regressierechte reeds een significant gedeelte van de variatie in onze gegevens verklaard heeft?

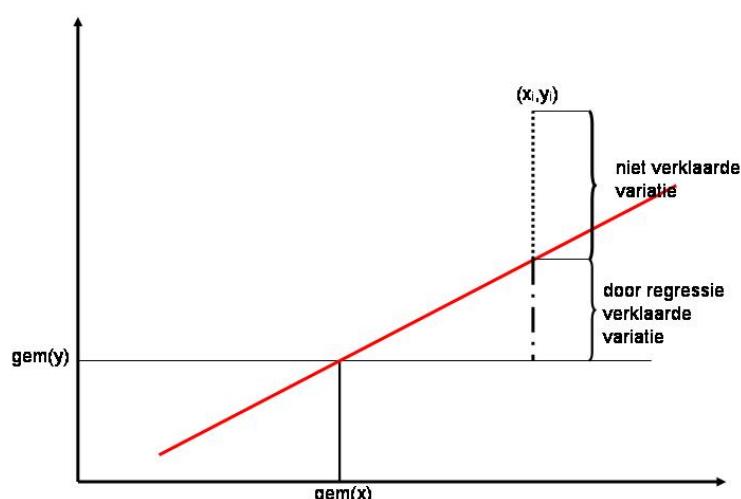
De totale variatie van onze y-waarden wordt gegeven door SS_T (total sum of squares):

$$SS_T = \sum (y_i - \bar{y})^2$$

Het is de som van de kwadratische afwijkingen tot de gemiddelde y-waarde. Deze variatie kan in 2 stukken opgedeeld worden: een stuk SS_{Reg} vanwege de variatie van de rechte tov de gemiddelde y-waarde en een stuk SS_R vanwege de variatie van de punten rond de rechte (de variatie in de residus). Men kan bewijzen dat:

$$SS_T = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 = SS_R + SS_{Reg}$$

Beide variaties worden in de onderstaande grafiek afgebeeld.



$$\text{gem}(x) = \bar{x} \quad \text{en} \quad \text{gem}(y) = \bar{y}$$

Indien we gerepliceerde metingen hebben, kunnen we de variatie in de residus (SS_R) nog opplitsen in een term te wijten aan de zuivere experimentele fout en een term te wijten aan de lack of fit. We gaan er hier niet verder op in.

De **determinatiecoëfficiënt R^2** wordt nu gedefinieerd als de proportie van de totale variatie in de afhankelijke variabele y die door de regressie verklaard wordt:

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Dit getal fluctueert tussen 0 en 1 en is gelijk aan het kwadraat van de correlatiecoëfficiënt tussen y en x (vandaar de notatie R^2).

Om de significantie van de verklaarde variantie na te gaan kunnen we ook een F-toets uitvoeren. Hiervoor zullen we eerst beide variaties SS_{Reg} en SS_R delen door hun respectievelijk aantal vrijheidsgraden 1 en $n-2$. We krijgen dan MS_{Reg} en MS_R (MS: mean square). De nulhypothese stelt dat de regressie geen significante bijdrage leverde en $MS_{Reg} = MS_R$. Het quotiënt MS_{Reg}/MS_R heeft onder H_0 een F-verdeling met 1 vrijheidsgraad in de teller en $n-2$ vrijheidsgraden in de noemer. Indien:

$$p = P\left(x > \frac{MS_{Reg}}{MS_R}\right) = 1 - F_{F_{1,n-2}}\left(\frac{MS_{Reg}}{MS_R}\right) < \alpha$$

dan zeggen we met een betrouwbaarheid van $1-\alpha$ dat de regressie significant is. $F_{F_{1,n-2}}$ is de verdelingsfunctie van de F-verdeling met 1 vrijheidsgraad in de teller en $n-2$ vrijheidsgraden in de noemer.

De meeste statistische softwarepakketten vatten deze F-toets samen in een ANOVA tabel:

	SS	Df	MS	F	Significantie p
Regressie	SS_{Reg}	$2-1=1$	$MS_{Reg}=SS_{Reg}/1$	MS_{Reg}/MS_R	$P(F>MS_{Reg}/MS_R)$
Rest	SS_R	$n-2$	$MS_R=SS_R/n-2$		
Totaal	SS_T	$n-1$			

Deze F-toets is gelijkwaardig met een t-toets voor het testen of de richtingscoëfficient van onze regressierechte (b_1) verschillend is van 0 en is dus gelijkwaardig met het nagaan of 0 in het $100(1-\alpha)\%$ betrouwbaarheidsinterval van b_1 ligt (zie verder).

Indien we over gerepliceerde metingen beschikken, kan men deze ANOVA tabel uitbreiden. Er worden dan 2 F-toetsen uitgevoerd: één zoals hierboven voor het nagaan of de regressie significant is en één waarbij wordt nagegaan of er nog een lack of fit is (dus of er nog een extra term in het model moet toegevoegd worden).

7.2.5 Betrouwbaarheidsintervallen

Nadat men heeft nagegaan dat de berekende regressierechte $\hat{y} = b_0 + b_1x$ de experimentele gegevens goed beschrijft, gaat men na hoe betrouwbaar de geschatte parameters b_0 , b_1 en de voorspellingen \hat{y} zijn. Er worden betrouwbaarheidsintervallen voor het snijpunt β_0 , de richtingscoëfficiënt β_1 en de respons η opgesteld. Deze $100(1-\alpha)\%$ betrouwbaarheidsintervallen hebben de volgende algemene vorm:

100(1- α)% BI voor de echte parameter =

[schatting van de parameter $-t_{n-2,1-\alpha/2} s_{\text{geschatte parameter}}$, schatting van de parameter $+t_{n-2,1-\alpha/2} s_{\text{geschatte parameter}}$]

$s_{\text{geschatte parameter}}$ is de **standaardfout** van de geschatte parameter. Bemerk dat hier vanwege het schatten van de 2 regressieparameters een t verdeling met n-2 parameters gebruikt wordt.

7.2.5.a BI voor het snijpunt en de richtingscoëfficiënt

Men kan aantonen dat:

$$s_{b_0} = s_e \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}} \quad \text{en} \quad s_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- 100(1- α)% BI voor $\beta_0 = [b_0 - t_{n-2,1-\alpha/2} s_{b_0}, b_0 + t_{n-2,1-\alpha/2} s_{b_0}]$
- 100(1- α)% BI voor $\beta_1 = [b_1 - t_{n-2,1-\alpha/2} s_{b_1}, b_1 + t_{n-2,1-\alpha/2} s_{b_1}]$

Als een alternatief voor de F-toets, die de significantie van de regressie weergeeft, kunnen we zoals reeds vermeld ook een t-toets uitvoeren.

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

$$\text{Waarde van de toetsingsgrootte: } t = \frac{b_1}{s_{b_1}}$$

$$\text{Aanvaardingsgebied: } \left[-t_{n-2,1-\alpha/2}, t_{n-2,1-\alpha/2} \right]$$

We kunnen met een analoge t-toets nagaan of we met 100(1- α)% betrouwbaarheid kunnen beweren dat $\beta_0 = 0$ of dat de regressierechte door de oorsprong gaat. (Doe dit zelf).

7.2.5.b BI voor de echte waarde η_o van de respons bij een gegeven x-waarde x_o

Men kan bewijzen dat

$$s_{\hat{y}_o} = s_e \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

- 100(1- α)% BI voor $\eta_o = [\hat{y}_o - t_{n-2,1-\alpha/2} s_{\hat{y}_o}, \hat{y}_o + t_{n-2,1-\alpha/2} s_{\hat{y}_o}]$

7.2.5.c BI voor predictie van een meetresultaat y_o bij een gegeven x-waarde x_o

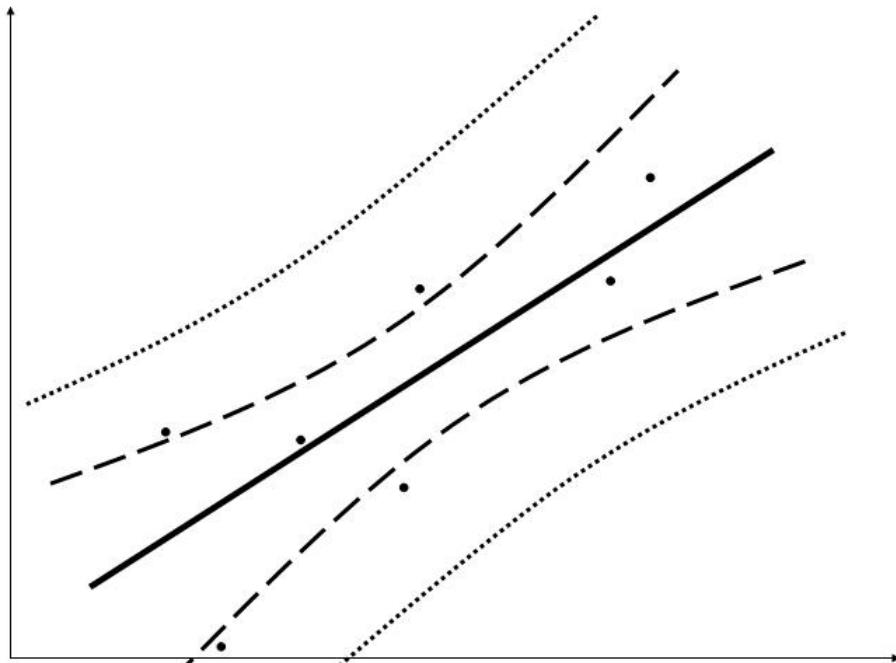
Bemerk dat de onzekerheid van de respons niet alleen te wijten is aan de onzekerheid van de regressierechte, maar ook aan de onzekerheid van de meting.

Men kan bewijzen dat

$$s_{y_o} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$\Rightarrow 100(1-\alpha)\% \text{ BI voor meetwaarde bij } x_o = \left[\hat{y}_o - t_{n-2,1-\alpha/2} s_{y_o}, \hat{y}_o + t_{n-2,1-\alpha/2} s_{y_o} \right]$$

In de onderstaande figuur worden de BI voor zowel de echte responswaarden als voor de meetresultaten voor een hele range van x-waarden voorgesteld.



Een typische grafiek van een regressierechte met haar betrouwbaarheidsintervallen
 (---): BI voor η_o ; (....): BI voor y_o

Bemerk dat de nauwkeurigheid afneemt naarmate we metingen willen voorspellen voor x-waarden die verder van \bar{x} liggen.

7.2.5.d BI voor predicties van x_o bij een gegeven y-waarde y_o

Bij calibratie-experimenten beschikken we over y-waarden en willen we met behulp hiervan de x-waarde voorspellen.

Men kan bewijzen dat

$$s_{x_o} = \frac{s_e}{b_1} \sqrt{1 + \frac{1}{n} + \frac{(y_o - \bar{y})^2}{b_1^2 \sum (x_i - \bar{x})^2}}$$

$$\Rightarrow 100(1-\alpha)\% \text{ BI voor } x_o = \left[\hat{x}_o - t_{n-2,1-\alpha/2} s_{x_o}, \hat{x}_o + t_{n-2,1-\alpha/2} s_{x_o} \right]$$

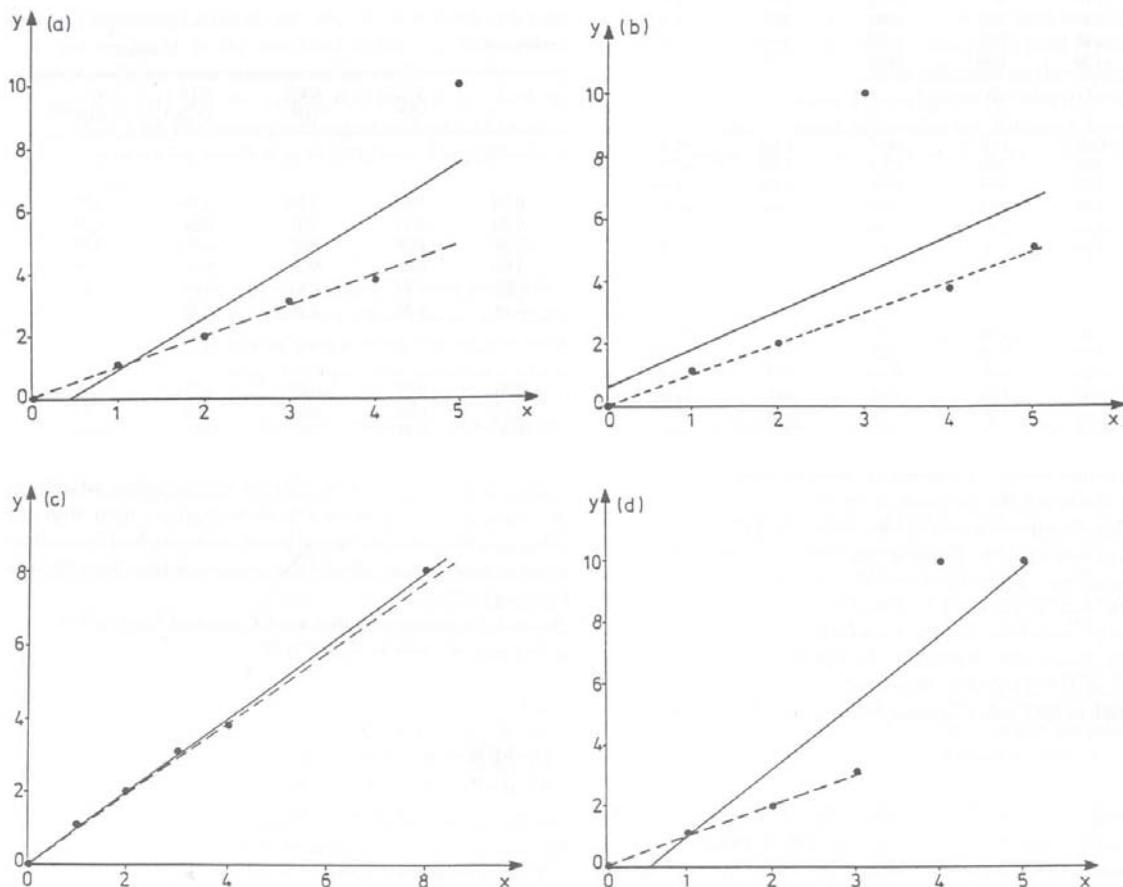
7.2.6 Experimentele design

Uit het bovenstaande volgt dat de breedte van het betrouwbaarheidsinterval evenredig is met $s_{\text{geschatte parameter}}$. Hoe kleiner dit getal hoe kleiner het betrouwbaarheidsinterval en hoe beter onze puntschattingen. Voor alle betrouwbaarheidsintervallen die vermeld werden, vindt men

een term $\sum(x_i - \bar{x})^2$ in de noemer van deze standaardafwijking. We hebben er dus alle voordeel bij de x-waarden waarbij de metingen worden uitgevoerd, zodanig te kiezen dat deze som zo groot mogelijk is. Dit betekent dat we onze x-waarden best zo ver mogelijk van de gemiddelde x-waarde kiezen. Indien we dus 10 metingen kunnen uitvoeren zullen we best 5 metingen doen bij de laagste x-waarde en vijf bij de hoogste x-waarde. Het grote nadeel hiervan is dat we geen enkele controle hebben over het al dan niet correct zijn van het lineaire model. Vandaar dat men meestal de metingen doet bij x-waarden gelijk gespreid tussen dit minimum en maximum.

7.2.7 Outliers; aberrante metingen, uitschieters,

De volgende grafieken illustreren verschillende soorten uitschieters en hun gevolgen op de regressierechte.



Bemerk dat afwijkende meetresultaten bij x-waarden, die dicht bij de maximale of minimale x-waarde liggen, een groter effect hebben dan deze die midden in het x-interval liggen (figuur a en b). Het is ook best geen metingen te doen bij een extreme waarde van x (\rightarrow hefboompunt of leverage point) (figuur c en d). Indien je geconfronteerd wordt met een uitschieter ga dan steeds eens naar de data kijken. Misschien is er een eenvoudige reden voor deze afwijkende meting: een typfout, het gebruik van een ander toestel, meting bij een andere temperatuur, ... Voor het behandelen van outliers bestaat geen eenduidige strategie. Bij het gebruik van robuuste lineaire regressie methoden wordt het resultaat niet zozeer beïnvloed door deze meting (cfr. de mediaan). We raden het gebruik van deze methoden dan ten sterkste aan. We vermelden kort de methode waarbij men tussen elk koppel van meetpunten een rechte trekt; de richtingscoëfficiënt en het snijpunt van deze rechten met de Y-as berekent en nadien als

richtingscoëfficiënt de mediaan neemt van de richtingscoëfficiënten van de verschillende rechten en analoog als snijpunt met de Y-as de mediaan neemt van de berekende snijpunten.

Intermezzo

Faculteiten, Binomiaalcoëfficiënten, Permutaties en Combinaties

In dit hoofdstukje herhalen we enkele definities en begrippen uit de combinatoriek, die van belang zijn in de kansrekening.

A. Het rekenen met faculteiten en binomiaalcoëfficiënten

A.1 Definities

- i) Faculteit: $n! = n(n-1)(n-2)\dots 1$ en $0! = 1$
- ii) Binomiaalcoëfficiënt: $C_n^r = \frac{n!}{r!(n-r)!} = \frac{n(n-1)\dots(n-r+1)}{r\cdot\dots\cdot 2\cdot 1}$

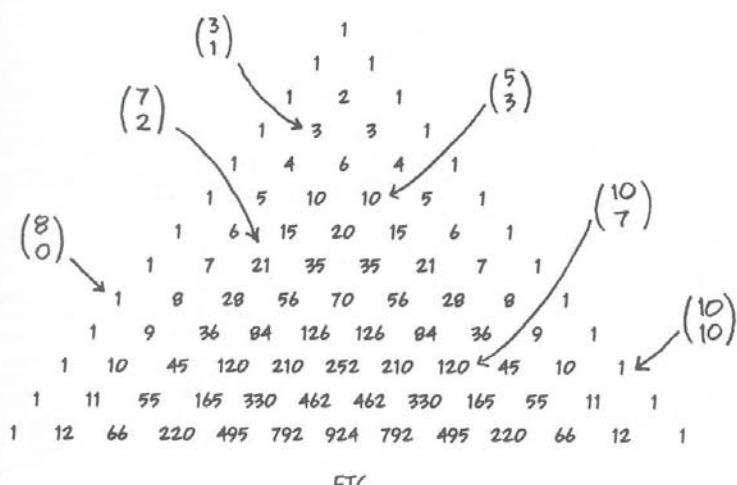
A.2 Eigenschappen

- i) $\binom{n}{r} = \binom{n}{n-r}$ voor elke r
- ii) $\binom{n}{0} = \binom{n}{n} = 1$ en $\binom{n}{1} = \binom{n}{n-1} = n$

A.3 Driehoek van Pascal

$$\binom{n+1}{r+1} = \binom{n}{r} + \binom{n}{r+1}$$

De binomiaalcoëfficiënten kun je ook vinden met **Pascal's driehoek**. Elk cijfer is de som van de twee getallen die er schuin boven staan.



Om $\binom{n}{k}$ te vinden, ga n rijen naar beneden en dan k cijfers naar rechts (vergeet niet te beginnen met tellen bij 0).

$\binom{0}{0}$	$\binom{1}{0}$	$\binom{1}{1}$
$\binom{2}{0}$	$\binom{2}{1}$	$\binom{2}{2}$
$\binom{3}{0}$	$\binom{3}{1}$	$\binom{3}{2}$
$\binom{4}{0}$	$\binom{4}{1}$	$\binom{4}{2}$
$\binom{4}{3}$	$\binom{4}{4}$	$\binom{4}{4}$

De buitenste getallen zijn “1” en de andere getallen zijn de som van de 2 getallen er direct boven.

A.4 Binomium van Newton

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$$

A.5 Formule van Stirling voor het benaderen van $n!$ als n groot is

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e} \right)^n = \sqrt{2\pi} e^{-n} n^{n+1/2}$$

Bedenk dat je rekenmachine reeds bij $20!$ problemen heeft.

B. Permutaties en combinaties

Indien we n experimenten na elkaar uitvoeren willen we een idee hebben van de mogelijke uitkomsten; soms is het voldoende het aantal mogelijke uitkomsten te kennen; soms verlangen we een lijst met alle mogelijke uitkomsten. Men kan hiervoor meerdere telmethodes gebruiken: de produktregel, de permutatieregel en de combinatieregel.
We behandelen deze regels hier in het kort.

B.1 De produktregel

Bij het uitvoeren van een sequentie van n experimenten, waarbij het eerste experiment k_1 mogelijke uitkomsten heeft, het tweede k_2 , enz., is het aantal mogelijke oplossingen gelijk aan het produkt: $k_1 k_2 \dots k_n$

Voorbeeld

In een ziekenhuis heeft men de bloedsamples van donors gelabeld. Er zijn vier bloedtypes: A, B, AB en O. Bloed kan rhesus positief of negatief zijn. Tevens kan een bloeddonor mannelijk of vrouwelijk zijn. Indien men het bloedtype, de rhesusfactor en het geslacht van de donor op het bloedsample noteert, hoeveel mogelijkheden heeft men dan ?

#bloedtype x #rhesusfactor x #sekse = $4 \times 2 \times 2 = 16$ mogelijkheden

N.B.: Let op dat de volgorde van de experimenten vastligt, anders heeft men natuurlijk meer mogelijkheden. Dit wordt duidelijk met de permutatieregel.

B.2 De permutatieregel

Een permutatie is een organisatie van n objecten in een bepaalde volgorde.

Het aantal mogelijke keuzes, in een bepaalde volgorde, van r objecten uit een totaal van n objecten is: $\frac{n!}{(n-r)!} = n(n-1)\dots(n-r+1)$

Bedenk dat je voor je eerste object kunt kiezen uit n objecten, voor het tweede uit n-1 en voor het laatste uit n-r+1.

N.B.: De volgorde is van belang en je trekt zonder terugleggen.

Voorbeeld

Een televisienieuws directeur wil 3 reportages gebruiken voor het avondnieuws. Hij heeft de keuze uit 8 reportages. Eén hieruit wil hij als “hoofdnieuws”, één als “bijkomend nieuws” en één als “afsluiter”. Op hoeveel mogelijke manieren kan hij het avondnieuws opstellen ?

$$\frac{8!}{(8-3)!} = 8 \cdot 7 \cdot 6 = 336$$

B.3 Combinatieregel

In het vorige voorbeeld was de volgorde van de keuzes belangrijk. Indien de volgorde geen rol speelt, spreken we van een combinatie ipv een permutatie. Er geldt dan ook de combinatieregel.

Het aantal mogelijke combinaties van r objecten uit n objecten wordt gegeven door:

$$C_n^r = \binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)...(n-r+1)}{r \cdot ... \cdot 2 \cdot 1}$$

Vermits de volgorde geen rol speelt, moeten we het aantal mogelijke sequenties nog delen door $r!$, het aantal permutaties van r objecten.

N.B.: De volgorde is niet van belang en je trekt zonder terugleggen.

Voorbeeld

In een club zijn er 7 vrouwen en 5 mannen. Hieruit wordt een comité van 3 vrouwen en 2 mannen gekozen. Hoeveel mogelijkheden zijn er ?

We moeten dus 3 vrouwen uit 7 kiezen en 2 mannen uit 5. $\rightarrow \binom{7}{3} \cdot \binom{5}{2} = \frac{7!}{3!4!} \cdot \frac{5!}{2!3!} = 350$