Hogeschool van Amsterdam

WISKUNDE A

Studentenzaken

Gerard Verhoef

1. Voorwoord

Deze syllabus is bedoeld als aanvulling op het boek Wiswijs van Fred Pach en Hans Wisbrun dat we gebruiken voor de module wiskunde A als voorbereiding op een studie aan de Hogeschool van Amsterdam.

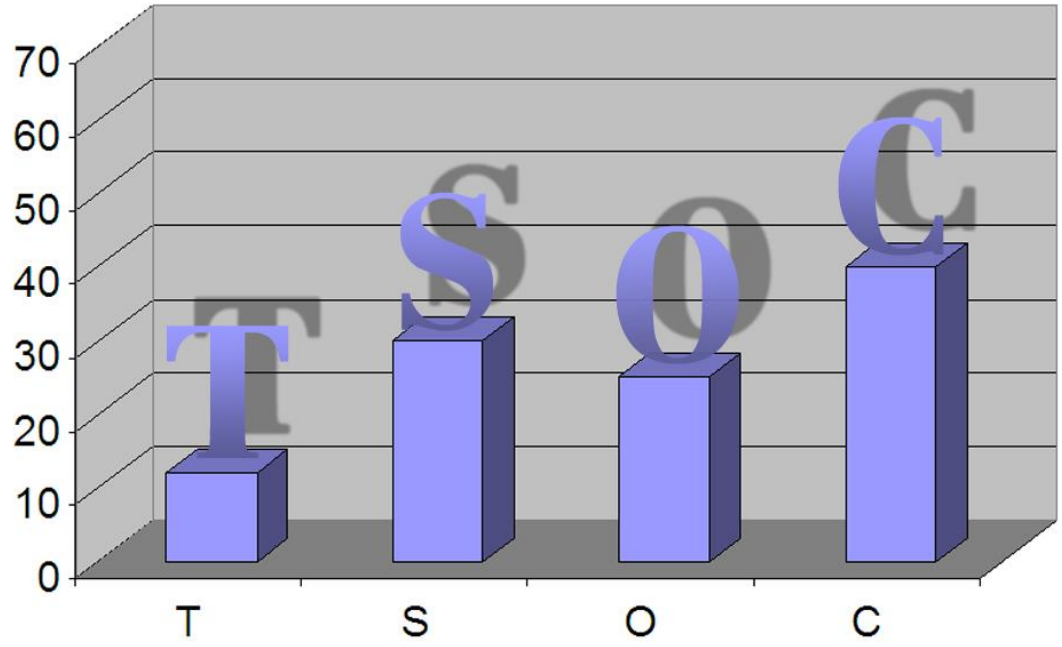
Ik wens alle studenten vel succes met deze cursus en met hun vervolgopleiding op de Hogeschool van Amsterdam

Gerard Verhoef,

November 2014

2. Inhoudsopgave

3. Introductie in de statistiek.



3.1. Inleiding.

Het woord statistiek is via omwegen vrijwel zeker afkomstig van het Italiaanse woord statistica dat staatsman betekent. Staatsmannen maakten al enkele eeuwen geleden gebruik van volkstellingen om meer over de bevolking te weten te komen en daar kwam veel tellen en interpreteren van getallen bij kijken.

Het woord statistica werd het eerst gebruikt door Gottfried Achenwall (1719 — 1772) , een professor die werkte aan de universiteiten van Marlborough en Gottingen. Een andere geleerde , professor Dr. E. Zimmerman , introduceerde het woord statistics voor het eerst in Engeland . En van dat woord statistics is de Nederlandse vertaling statistiek afgeleid.

3.2. The 0. J. Simpson murder case.

De bekendste uitspraak op het gebied van de statistiek is vrijwel zeker de uitspraak van de Engelse staatsman Benjamin Disraeli:

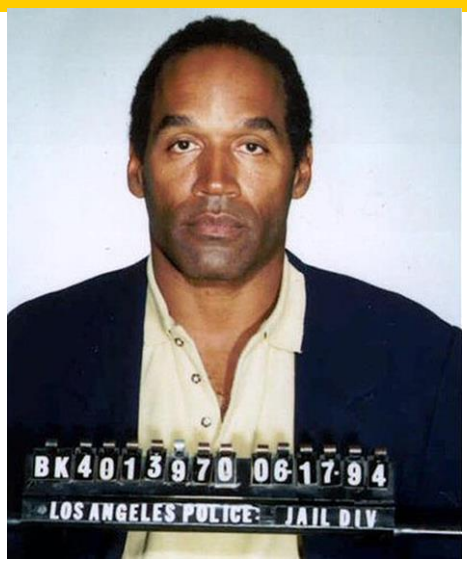
There are three kinds of lies: lies , damned lies and statistics.

waarmee hij zoiets zegt als: mensen, je hebt drie soorten leugens: gewone leugens , je hebt verduiveld grote leugens, maar je hebt nog veel grotere leugens die praktisch niet te ontdekken zijn en die leugens zijn te vinden in de statistiek.

Er zit absoluut veel waarheid in deze uitspraak: statistiek kan een hulpmiddel zijn om buitengewoon subtiel te liegen en waarschijnlijk het mooiste voorbeeld uit het afgelopen decennium zal nu gegeven worden. Enkele jaren geleden speelde er een rechtszaak in de Verenigde Staten die wereldbekendheid verwierf en die dagelijks zeer uitgebreid te zien was op de televisie bij de Amerikaanse nieuwszender CNN.

Deze rechtszaak betrof 0. J. Simpson , een zwarte Amerikaan die landelijk bekend was vanwege zijn sportieve prestaties op het gebied van American football, maar hij was ook bekend vanwege zijn optreden in films.

Deze 0. J. Simpson werd ervan beschuldigd zijn vrouw en tevens de minnaar van zijn vrouw vermoord te hebben en er was buitengewoon veel belastend materiaal tegen hem verzameld. Zo werden bijvoorbeeld bloedsporen met DNA van zijn vrouw op zijn schoenen en in zijn auto gevonden.



De verdediging van 0.J. Simpson werd uiterst geraffineerd gevoerd door de twee Amerikaanse topadvocaten Johnny Crochan en Alan Dershowitch ; de laatste was een blanke man van zo rond de 50 jaar. Zo werd er uit strategische overwegingen door deze advocaten ontzettend veel aandacht besteed aan vormfouten, gemaakt tijdens het onderzoek in deze dubbele moordzaak, met name tijdens de laboratoriumonderzoeken en het onderzoek van de directe omgeving van het huis van Simpson. Een ander punt dat opviel was de manipulatie van de jury door Alan Dershowitch. Met name de nu volgende uitspraak van Dershowitch verdient onze volle aandacht:

Aangezien minder dan 1 op de 1000 vrouwen in de Verenigde Staten , die door hun partner worden mishandeld , ook daadwerkelijk door hun partner worden vermoord , is het feit dat er in het huwelijk van de Simpsons sprake was van ernstige mishandeling niet belastend voor O.J. Simpson. Het bewezen feit dat O. J. Simpson zijn vrouw mishandelde kon niet als bewijsmateriaal gebruikt worden.

Alan Dershowitch herhaalde deze uitspraak keer op keer tegen de jury, waarbij hij waarschijnlijk bewust een techniek hanteerde die reeds door de Duitse propagandist Josef Goebbels tijdens de Tweede Wereldoorlog toegepast en verwoord was n.l. dat als je een leugen maar vele malen herhaalde, de mensen er dan vanzelf in gingen geloven.

Laten we eens zorgvuldig naar de uitspraak van Dershowitch kijken. Hij zegt:

kijk nu eens naar: alle vrouwen in de V.S. die door hun partner mishandeld zijn, dit is (helaas) een hele grote groep vrouwen in de Verenigde Staten en deze groep kan je een populatie noemen (een basisbegrip uit de statistiek) , d.w.z. een grote groep die onderzocht gaat worden.

(Even terzijde: meestal wordt bij het onderzoek van een populatie slechts een klein gedeelte van de populatie onderzocht, de steekproef geheten).

Gaan we weer verder met de hoofdlijn. Dershowitch zegt: kijk nu eens naar die grote groep mishandelde vrouwen. Ze worden allen door hun partner mishandeld, maar in feite wordt in die groep minder dan 1 op de 1000, dus minder dan 0.1 %, door hun partner vermoord. Dus: Een vrouw die mishandeld wordt door haar partner heeft slechts een kans <0.1 % om door haar partner vermoord te worden, dus verwaarloosbaar klein! Het feit dat O.J. Simpson zijn vrouw mishandelde wil helemaal nog niet zeggen dat hij haar vermoord zou moeten hebben. We kunnen stellen dat Dershowitch hier werkelijk op een schitterende manier de zaken verkeerd voorstelde. Hij vertelde wel de waarheid, maar .... slechts een gedeelte van de waarheid en niet de gehele waarheid

De jonge vrouw Marcia Meines , de openbare aanklaagster, liet hier een prachtige kans voorbijgaan om dit argument van Dershowitch om te draaien in haar voordeel. Maar dat was haar nauwelijks kwalijk te nemen: wie de uitzendingen van CNN volgde ontkwam niet aan de indruk dat zij toch enigszins bevangen was door het optreden van deze twee topadvocaten en bovendien was, om het gat in de redenering van Dershowitch te ontdekken, een zekere training in statistisch denken absoluut noodzakelijk.

Want wat deed Dershowitch? Hij vestigde de aandacht van de jury op alle Amerikaanse vrouwen die mishandeld worden door hun partner. Hij vestigde de aandacht op de verkeerde groep : de verkeerde populatie. Want de vrouw van O.J. Simpson was niet alleen mishandeld, maar ook vermoord! Dat nare feit werd door Dershowitch even gemakshalve over het hoofd gezien.

Marcia Meines had tegen de jury moeten zeggen: Leden van de jury, U moet niet kijken naar alle vrouwen in de Verenigde Staten, die door hun partner mishandeld zijn maar U moet kijken naar alle vrouwen in de Verenigde Staten die mishandeld zijn door hun partner en vermoord zijn. En dan leden van de jury, als U kijkt naar deze groep vrouwen, dan blijkt dat in meer dan 80 % van de gevallen de vrouwen vermoord zijn door hun partner!

Dus als je kijkt naar de groep vrouwen in de Verenigde Staten, die mishandeld werden door hun partner en ook nog vermoord werden, dan kijk je naar een groep vrouwen die veel en veel kleiner is dan de groep vrouwen in de VS. die door hun partner mishandeld worden. En dat is de reden dat de kans explosief stijgt van minder dan 0.1 % naar meer dan 80 %. Dus alleen al op puur statistische gronden (meer dan 80 % kans !) was er al aanleiding om O.J. Simpson als serieuze potentiële verdachte te beschouwen. Hier liet Marcia Meines een werkelijk unieke kans liggen om het argument van Dershowitch over te nemen, te corrigeren en in haar voordeel te doen ombuigen.

Alan Dershowitch vestigde de aandacht van de jury dus op de verkeerde populatie, maar het is de jury nauwelijks kwalijk te nemen dat hen dit ontging.

3.3. Lucia de B.

Een ander bekend voorbeeld is dat van de verpleegster Lucia de B. Hieronder een artikel hierover van Joep Engels uit Trouw van 12 Maart 2007.

Hij schaamde zich voor zijn beroepsgroep. Allerlei collega's hadden zich gemengd in het statistische debat rond Lucia de B., de Haagse verpleegkundige die een levenslange gevangenisstraf uitzit omdat zij zeven patiënten om het leven gebracht zou hebben.

Tijdens het proces heeft de statistiek moeten uitwijzen of het toeval kon zijn dat tijdens haar dienst zo vaak patiënten overleden. Toxicologisch onderzoek bracht onvoldoende duidelijkheid.

De statistici steggelden over de wetenschappelijke methode, over rekenfouten en verkeerde conclusies.

Maar ze zagen één ding over het hoofd.

"De data deugden niet", zegt Richard Gill, hoogleraar statistiek aan de Rijksuniversiteit van Leiden. "Niemand heeft nagegaan hoe de gegevens zijn verzameld. Ikzelf ook niet, tot mijn grote schaamte. Het is alsof iemand vraagt hoe groot de kans is dat een muntstuk bij tachtig van de honderd worpen op kop valt, en dat we er later achterkomen dat hij het muntstuk regelmatig zelf op kop heeft gelegd. Het hele debat over de gebruikte rekenmethode en de beste schatting voor de kans werd in dit licht zinloos."

Toch is het die kansberekening die het lot van Lucia de B. bezegelde. Die beruchte kans van één op 342 miljoen. Het was deze onwaarschijnlijkheid die zich volgens Gill in vele hoofden nestelde en liet vertalen als 'dat kan geen toeval zijn'. (Ter vergelijking: de kans om 'alle zes goed' te hebben bij de Lotto is een op 4,5 miljoen. De kans om 'alle dertien goed' te hebben bij de toto is een op 1,5 miljoen).

Zo had de statisticus van de rechtbank, de Leidse hoogleraar Henk Elffers, het ook gebracht. Hij had berekend dat dit de kans was dat de Haagse verpleegkundige bij toeval aanwezig was bij verdachte sterfgevallen en reanimaties in het Juliana Kinderziekenhuis en het Rode Kruis Ziekenhuis. En omdat Elffers vooraf een kans van één op tienduizend als ondergrens had geponeerd, concludeerde hij dat Lucia's aanwezigheid bij al deze gevallen niet door louter toeval kon worden verklaard.

De berekening van Elffers, die verbonden is aan het Nederlands Studiecentrum Criminaliteit en Rechtshandhaving in Leiden, is rechttoe-rechtaan statistiek (zie kader). Maar wat zegt dat getal eigenlijk? Elffers heeft de kans berekend dat een onschuldige bij toeval tijdens de verdachte incidenten dienst heeft. Maar dat doet er niet toe, zeggen critici. Interessant is de vraag hoe groot de kans is dat iemand die bij die incidenten aanwezig is, onschuldig is. Dat lijkt taalkundige muggenzifterij, maar wiskundig maakt het wel degelijk uit.

De Nijmeegse filosoof Ton Derksen vergelijkt het verschil met de zoektocht naar een bonte kraai. Vogelaar A heeft een bonte vogel gevangen en vraagt zich af hoe groot de kans is dat dit een kraai is. Die kans ligt anders bij vogelaar B die een kraai in zijn netten heeft en wil weten of deze bont is.

Statistici hebben gedebatteerd over de vraag hoe ze die andere, reëlere kans voor Lucia de B. moeten berekenen. Elffers had het beschouwde jaar met de acht incidenten moeten vergelijken met andere jaren, zeggen sommigen. Dan had hij in elk geval gezien dat de afdeling drie jaar eerder, toen Lucia er nog niet werkte, meer incidenten telde.

Anderen stellen dat je de kans dat een onschuldige Lucia een ongelukkig lot heeft getrokken, moet wegen tegenover de kans dat een Nederlandse verpleegkundige een seriemoordenares is.

Intussen wordt het gesteggel door nieuwe feiten ingehaald. Men ontdekt dat Elffers de kansen van de drie ziekenhuizen met elkaar heeft vermenigvuldigd. Een blunder, zegt Gill. Elffers vermengt de kans dat Lucia de B. bij de incidenten aanwezig was, met de kans dat ze op bepaalde afdelingen werkte. Volgens die rekenwijze maakt iemand die regelmatig van baan wisselt, zich vanzelf verdacht. Het scheelt een factor duizend in de kansberekening.

Maar de grote klapper moet nog komen. De data deugen niet.

Rechercheurs blijken de dossiers te hebben doorzocht terwijl ze in hun achterhoofd Lucia de B. al hadden veroordeeld. Daardoor zijn gebeurtenissen ineens verdacht geworden omdat zij erbij betrokken is, zijn andere verdachte gebeurtenissen waarbij ze niet betrokken is, in haar schoenen geschoven of over het hoofd gezien, en zijn verdachte gevallen weer onverdacht geworden omdat ze er niets mee te maken had.

Nu het lijstje met verdachte incidenten zo onzorgvuldig blijkt te zijn opgesteld, doet de statisticistrijd er nauwelijks meer toe. Alle berekeningen komen op losse schroeven te staan. De filosoof Derksen reconstrueert zo goed en zo kwaad als het kan de feitelijke gang van zaken. Met die gegevens levert de berekeningswijze van Elffers - mits correct uitgevoerd - een heel ander kans: één op vijftig.

Ook daar is Gill niet tevreden mee. Dit soort kansverdelingen laat zich niet herleiden tot bakken met witte en zwarte ballen, zegt hij. De diensten in de ziekenhuizen zijn niet allemaal op briefjes gezet en in een grote schaal gelegd waarna Lucia er 124 mocht uitvissen. Verpleegkundigen draaien een tijdje dezelfde (dag- of nacht-)diensten, gaan op vakantie of zijn ziek.

Er zitten patronen in die dienstroosters. Gill heeft die patronen proberen te vangen in een standaardverdeling uit de kanstheorie (de zogeheten gammaverdeling).

Daarbij gebruikt hij één aanname: 30 procent kans dat een willekeurige verpleegster twee keer zoveel meemaakt als een willekeurige andere.

Gill: "Zo gek is dat niet. De ervaring leert dat de ene verpleegster veel meer meemaakt dan de andere en dat gebeurtenissen zich veelal geclusterd voordoen."

De uitkomst van die aanname is opmerkelijk: jaarlijks overkomt één op de negen verpleegsters dezelfde reeks incidenten als Lucia de B. "Dat zou dus normaal nooit opvallen. Tenzij het om een verpleegkundige gaat die iedereen al verdacht vindt vanwege haar kleding of haar verleden."

Intussen had het Haagse gerechtshof al ingezien dat statistiek zeer controversieel kon zijn en distantieerde het zich ervan. Terwijl de rechtbank in Den Haag de waarschijnlijkheidsberekeningen van Elffers in 2003 nog zwaar liet meewegen en het 'uitermate onwaarschijnlijk achtte dat de verdachte de incidenten bij toeval zou hebben meegemaakt', veegde het hof in zijn arrest op 18 juni 2004 alle statistische bewijzen van tafel: "Er is geen statistisch bewijs in de vorm van toevalsberekeningen gebruikt."

Elffers zelf denkt daar inmiddels ook zo over. Desgevraagd zegt hij nu:

"Het Hof heeft het statistische argument als onvoldoende terzijde geschoven. En dat is goed, de rechter verdient hier eigenlijk een pluim voor. De statistiek kan in zijn eentje nooit voldoende bewijs leveren. En als het statistische bewijs wel voldoende zou zijn, is de statistiek niet meer nodig."

"Dan weten we het namelijk zeker. Hoe spannend deze discussies ook mogen zijn, ik denk niet dat de statistiek ooit nog een rol zal spelen in de rechtszaal."

Als je het zo bekijkt, lijkt het erop dat de hele statisticistrijd een achterhoedegevecht is geweest. Derksen ziet dat anders. Hij wijst erop dat het hof weliswaar geen termen als toeval of kans gebruikt, maar dat het arrest doordrenkt is van de statistiek.

Het meest opvallend is dat in de redenering die wel het schakelbewijs wordt genoemd. Dat gaat als volgt. Het hof acht bij twee gevallen - een overlijden en een reanimatie - bewezen dat er giftige stoffen zijn toegediend, en wel door Lucia de B.

Bij andere incidenten is er geen enkel bewijs, maar het hof vindt ze zo lijken op de twee genoemde gevallen, dat aangenomen moet worden dat dit ook moorden zijn.

In alle gevallen was Lucia de B. in de buurt en ze "bestreken een betrekkelijk korte periode". Deze geciteerde passage uit het arrest heeft alleen betekenis in de statistische redenering dat 'dat geen toeval kan zijn'.

NOTES: De chronologie van de zaak Lucia de B.; "Er is geen statistisch bewijs in de vorm van toevalsberekeningen gebruikt", stelt het Haagse gerechtshof in zijn arrest.; Toch gaat het hof in op de waarschuwing van Elffers dat zijn statistische conclusie dat de samenloop geen toeval kon zijn, niet hetzelfde was als een bewijs voor moord. Er konden immers andere factoren in het spel zijn geweest. Lucia's diensten vielen vaak samen met die van een ander. Of: ze verzorgde altijd de ernstig zieke kinderen en daarom was ze zo vaak bij sterfgevallen betrokken.; Het hof acht het nodig om alle alternatieve verklaringen die Elffers noemt, te weerleggen. Nee, staat er: de verdachte was niet meer dan haar collega's belast met de zorg voor ernstig zieke kinderen. Of: nee, een ander dienstrooster is geen aannemelijke verklaring. Zo werkt de rechter het rijtje van Elffers af en zorgt het ervoor dat de conclusie van de Haagse rechtbank overeind kan blijven: de enige verklaring voor die onwaarschijnlijke samenloop van gebeurtenissen is dat Lucia de B. de dader is.; Voor de leek biedt dit overigens een interessant inkijkje in het juridisch denken. Elffers had enkele alternatieve verklaringen gegeven, zijn lijstje bevatte slechts, bij wijze van voorbeeld, vijf hypotheses. Maar in een rechtszaak doet alleen ter zake wat ter tafel komt. De rechter heeft met zijn vijf tegenargumenten alle hypotheses weerlegd. Omdat die alle vijf weerlegd zijn, is er volgens de rechter dus sprake van moord.; Toxicologie; Er zijn twee toxicologische bewijzen. Op 25 januari 2001 moet de zesjarige lichamelijk en geestelijk gehandicapte Ahmad worden gereanimeerd: zijn bloed blijkt een hoge concentratie van een slaapmiddel te bevatten. Dat betwist verder niemand. De vraag is waarom dit middel in zo'n hoge dosis is toegediend, en of het om een fout of een bewuste daad ging.; Bij baby Amber ligt het ingewikkelder. Op 4 september 2001 overlijdt de zes maanden oude Amber, volgens de rechters aan digoxine. Dit medicijn, dat wordt gebruikt om de hartfunctie te reguleren, vereist een nauwkeurige dosering. Een beetje te veel digoxine is al gauw dodelijk.; Maar Ambers overlijden wordt aanvankelijk als natuurlijk beoordeeld. De eerste sectie geeft geen aanleiding om er anders over te denken. Pas als de commotie zo hoog is opgelopen en er een tweede sectie wordt verricht, vindt de patholoog gaasjes in Ambers lichaam en die gaasjes bevatten volgens het Nederlands Forensisch Instituut een dodelijke concentratie digoxine.; Maar de deskundigen verklaren tegenover de rechtbank dat daaruit geen conclusies kunnen worden getrokken. De digoxine kan zich in de gaasjes hebben opgehoopt: de gevonden concentraties zeggen niets over de concentraties in het bloed van Amber. De rechter neemt deze nuancering voor kennisgeving aan, maar houdt uiteindelijk toch vast aan de theorie dat het eigenlijk niet anders kan dan dat de baby is overleden aan een overdosis digoxine.; Het hof is stelliger. Alle wetenschappelijke twijfels zijn volgens de rechter door nieuwe inzichten verdwenen. Maar daarbij gaat het hof aan andere tegenstrijdigheden voorbij. Zo is het hart bij een acute dioxinevergiftiging samengetrokken; Ambers hart was dat niet. Maar de opmerkelijkste tegenstrijdigheid schuilt in de uitslagen van een laboratorium in Straatsburg dat de test van het NFI herhaalde. Straatsburg vond digoxineconcentraties waarmee de dood van Amber niet te verklaren is. Het rapport verscheen te laat voor het hoger beroep en lag twee jaar in een bureaula van het NFI. Pas als de Hoge Raad de zaak terugverwijst naar het Gerechtshof van Amsterdam, komt het Straatsburg-rapport boven tafel. Maar in dit nieuwe hoger beroep speelt het geen rol, omdat het hof van de Hoge Raad de bewijsvoering niet opnieuw mag onderzoeken.; Wel of geen statistiek; Op 4 september 2001 overlijdt in het Haagse Juliana Kinderziekenhuis (JKZ) de zes maanden oude Amber. Artsen stellen een natuurlijke dood vast, maar als daags erna een verpleegkundige meldt dat dit al het zoveelste verdachte sterfgeval is tijdens een dienst van Lucia de B., stelt het ziekenhuis een onderzoek in.; Meer mensen op de afdeling blijken zich te hebben verwonderd over de incidenten tijdens Lucia's diensten. De volgende ochtend ligt er al een lijstje met 'verdachte incidenten'.; In 2003 veroordeelt de rechtbank van Den Haag haar tot levenslang voor de moord op vier patiënten en poging tot moord op drie patiënten. Ruim een jaar later voegt het Haagse gerechtshof aan het levenslang nog tbs toe, nu vanwege zeven moorden en drie pogingen. In 2006 verwerpt de Hoge Raad klachten over de bewijsvoering,

maar beslist wel dat levenslang met tbs niet mogelijk is. In juli 2006 veroordeelt het gerechtshof in Amsterdam Lucia de B. opnieuw tot levenslang, nu zonder tbs.; Tijdens geen enkele rechtszitting is een direct bewijs voor de moorden geleverd. Niemand heeft het haar zien doen, en veel deskundigen hebben verklaard dat het gaat om ernstig zieke patiënten die onverwacht en door een natuurlijke oorzaak kunnen overlijden.; De rechtbank had alleen een statistisch bewijs. Het Haagse hof achtte bewezen dat Amber was vermoord en zag in de dagboekaantekeningen van Lucia dat ze weer had toegegeven aan haar 'compulsie', het bewijs dat zij een 'dwangmatige drang had om patiënten om het leven te brengen'. Daarmee was voor het hof bewezen dat zij de moordenares was. En dit bewijs vormde weer de basis voor het bewijs van de andere moorden; het zogeheten schakelbewijs. Volgens Lucia zelf sloeg die compulsie op haar drang om Tarotkaarten te leggen.; Terwijl wetenschappers over de statistiek twistten, dook de Nijmeegse filosoof Ton Derksen in de dossiers en ontdekte dat het OM en de rechters wel zeer selectief met de bewijzen en verklaringen waren omgegaan.; Zijn boek 'Lucia de B., reconstructie van een gerechtelijke dwaling' leidde ertoe dat de zaak opnieuw werd bestudeerd. Op 19 oktober 2006 adviseerde de Toegangscommissie van de Commissie evaluatie afgesloten zaken (ingesteld naar aanleiding van de Schiedammer Parkmoord) om te onderzoeken of er fouten zijn gemaakt bij de opsporing of tijdens de rechtszittingen.; De verwachting was dat het onderzoek rond deze tijd zou zijn afgerond waarna het College van procureurs-generaal eventueel tot heropening zou kunnen besluiten. Het onderzoek laat echter nog enige maanden op zich wachten, meldt het Openbaar Ministerie.; Rekenwerk; Tussen 1 oktober 2000 en 9 september 2001 hebben 27 verpleegkundigen in het Juliana Kinderziekenhuis 1029 diensten gedraaid. De 8 incidenten, zoals ze in het jargon worden genoemd, vinden allemaal plaats tijdens de 142 diensten van Lucia de B. De kans dat dit toevallig gebeurt, is volgens Elffers gelijk aan de kans om in een bak met 1029 ballen - 1021 witte en 8 zwarte - al na 142 keer grijpen alle acht zwarte ballen te pakken te hebben.; Die kans is kleiner dan één op 9 miljoen. Maar omdat dit lot alle 27 verpleegkundigen van de afdeling had kunnen treffen, vermenigvuldigt Elffers de kans met 27 en komt hij uit op een kans van één op 300.000.; Vergelijkbare berekeningen bij twee afdelingen van het Rode Kruis Ziekenhuis geven hem ook twee kansen. Elffers vermenigvuldigt de drie kansen met elkaar en bepaalt zo de uitkomst van één op 342 miljoen.; Na de 'correctie' door Derksen gaat het in het JKZ om vijf incidenten tijdens 1029 diensten, waarvan vier tijdens de 142 diensten van Lucia de B. In de ballenbakstatistiek is dat een kans van ongeveer één op 25. Een juiste combinatie met de gecorrigeerde kansen voor het RKZ levert een totaalkans van één op 50.; Elffers erkent dat er met de cijfers van Derksen een andere kans uitrolt, maar doet geen uitspraak over de juistheid van deze nieuwe cijfers. Ook over de kritiek op zijn methode is hij terughoudend. "Ik ben het met Gill eens dat er aan mijn methode bezwaren kleven, maar over de vraag wat dan wel de juiste methode is, bakkeleien we nog."

3.4. Onzuivere steekproeven

Zoals reeds opgemerkt werd in het verhaal van de 0. J. Simpson murder case , wordt een grote populatie vaak onderzocht door slechts een gedeelte van die populatie , de steekproef geheten, te onderzoeken.

Het is een van de grote problemen in de statistiek om een goede steekproef te nemen, d.w.z. een steekproef die een goede afspiegeling is van de populatie. Zo'n steekproef wordt een zuivere steekproef genoemd. Om dit te illustreren, worden nu enkele voorbeelden gegeven van onzuivere steekproeven.

Eerst een extreem voorbeeld.

Stel je stuurt naar een aantal mensen, die in dezelfde stad als jij wonen, een vragenlijst waarin o.a. de volgende vraag is opgenomen:

Houdt U ervan vragenlijsten in te vullen?

Tel de antwoorden bij elkaar op en je zou waarschijnlijk tot de conclusie komen dat een overweldigende meerderheid (waarbij je een percentage geeft tot op een decimaal achter de komma om het overtuigender te laten lijken, bv. 87.3 %!) van deze "representatieve steekproef uit de populatie" aangeeft er geen enkel bezwaar tegen te hebben vragenlijsten in te vullen.

Maar wat waarschijnlijk gebeurd is, is dat de meeste mensen , die de vragenlijst ontvingen en er totaal niet van houden om deze lijsten in te vullen , er meteen een papierprop van maakten richting de vuilnisemmer. Zodat het onderzoek waardeloos werd.

Een volgende voorbeeld:

een psychiater rapporteerde eens dat bijna iedereen neurotisch is. Nog afgezien van het feit dat zo'n bewering elke betekenis van het woord "neurotisch" bij voorbaat vernietigt, is het natuurlijk wel even zaak om te kijken naar de steekproef die deze psychiater heeft genomen. En dan zien we dat de psychiater tot zijn conclusie gekomen is op basis van het onderzoek van zijn patiënten en het is maar helemaal de vraag of deze patiëntenverzameling wel een goede afspiegeling is van de totale bevolking. Zeer waarschijnlijk niet!

Een laatste voorbeeld :

Amerikaanse kranten beweerden een tijdje geleden dat ongeveer 4 miljoen Amerikaanse katholieken in de laatste 10 jaar protestant waren geworden. De bron van deze bewering was een enquête, afgenomen door de geestelijke Daniel A. Poling, uitgever van de confessionele krant Christian Herald. De Herald was tot deze conclusie gekomen door een enquête sturen naar 25000 protestantse predikanten. De 2219 protestantse predikanten, die reageerden, rapporteerden dat zij in totaal 51361 vroegere katholieken in hun kerk als nieuwe leden hadden mogen ontvangen in de laatste 10 jaar. Omdat er indertijd zo'n 181000 protestantse predikanten in de Verenigde Staten waren, redeneerde

de Herald dat er dus landelijk het onderstaande aantal "overlopers" waren van de katholieke naar de protestantse kerk

$$\frac{181000}{2219} \times 51361 = 4189428$$

Will Oursler , een hoge geestelijke in de protestantse kerk, schreef daarop

Zelfs indien we rekening houden met mogelijke foutenmarges , dan nog is het zo dat nationaal gezien er nauwelijks minder dan 2 a 3 miljoen "overlopers" zijn en naar alle waarschijnlijkheid zal het getal wel dichter bij de 5 miljoen liggen.

In dit onderzoekje reageerde dus meer dan 90 % van de ondervraagde predikanten uit op de enquête.

Teneinde dit landelijke resultaat van 4189428 overlopers onderuit te halen, hoeven we alleen maar op te merken dat deze 90 % predikanten niet antwoordden omdat ze geen "overlopers ' te melden hadden.

Een betere schatting zou misschien zijn dat, omdat hij bij 25000 predikanten 51361 overlopers vond, dat voor 181000 predikanten zou leiden tot:

$$\frac{181000}{25000} \times 51361 = 371854$$

Het laatste antwoord (zo'n 370 000) steekt schril af bij het eerder gevonden antwoord van zo'n 4 miljoen 200 000. Het ene getal is maar liefst 10 maal zo groot als het andere getal!

3.5. Statistische basisbegrippen.

Na de introductie van de begrippen populatie en steekproef gaan we nu verder met de introductie van enkele elementaire begrippen uit de statistiek. We gaan dit doen door te kijken naar de onderstaande tabel , waarin de lengte vermeld staat van 104591 Nederlandse jonge mannen (dienstplichtige militairen).

Klasse	Lengte		Lengte op 1 mm		Frequentie	Cumulatieve frequentie	Klassemidden	Relatieve frequentie	Cumulatieve relatieve frequentie
	van	tot	van	t/m					
1	155	159	154,5	159,4	105	105	156,95	0,1%	0,1%
2	160	164	159,5	164,4	1046	1151	161,95	1,0%	1,1%
3	165	169	164,5	169,4	4393	5544	166,95	4,2%	5,3%
4	170	174	169,5	174,4	14224	19768	171,95	13,6%	18,9%
5	175	179	174,5	179,4	26880	46648	176,95	25,7%	44,6%
6	180	184	179,5	184,4	29495	76143	181,95	28,2%	72,8%
7	185	189	184,5	189,4	18931	95074	186,95	18,1%	90,9%
8	190	194	189,5	194,4	7321	102395	191,95	7,0%	97,9%
9	195	199	194,5	199,4	1778	104173	196,95	1,7%	99,6%
10	200	204	199,5	204,4	418	104591	201,95	0,4%	100,0%
Totalen					104591			100,0%	

Tabel: Lengte militairen 3-1

In de eerste kolom staan de nummers van de klassen. Zo bestaat klasse 1 uit dienstplichtigen met een lengte van 155 cm t/m 159 cm.

In de tweede kolom staat het-lengte-interval (in cm), behorend bij die klasse, zoals door het CBS (= het Centraal Bureau voor de Statistiek) is opgegeven.

De derde kolom is ontstaan door de volgende redenering: meestal wordt de lengte opgegeven tot op 1 mm nauwkeurig, bv. 186.3 cm. Dat impliceert dat het CBS de lengten, die in de tabel staan, heeft afgerond op hele centimeters. Hoe zal die afronding in zijn werk zijn gegaan? Wel, je mag logischerwijze veronderstellen dat iemand met een lengte van 154.4 cm afgerond zal worden tot 154 cm en waarschijnlijk is iemand met een lengte van 154.5 cm afgerond tot 155 cm.

Deze methode is in elk geval onder statistici volkomen gebruikelijk. Dus zullen mensen met een lengte van 154.5 cm t/m 159.4 cm behoren tot klasse 1.

3.5.1. De absolute frequentie.

In de vierde kolom staat het aantal dienstplichtigen met die betreffende lengte. Zo zijn er 105 militairen die een lengte hebben zich bevindend in het interval 154.5 t/m 159.4 cm , en er zijn 1047 militairen met een lengte van 159.5 cm t/m 164.4 cm, enzovoorts. Het valt op dat de aantallen regelmatig opklimmen totdat een maximum bereikt wordt voor klasse 6 en daarna nemen de aantallen weer af. Het ziet er een beetje symmetrisch uit. Deze absolute frequenties (absoluut, omdat deze aantallen op zichzelf staan en niet vergeleken worden met een ander getal, dus niet vergeleken worden met bv. het totale aantal van 104591) worden vaak in statistiekboeken aangeduid met het symbool

$$f_i \text{ de frequentie van klasse i}$$

Dit symbool is dan een overkoepelend symbool voor:

$$f_1 = 105 \quad (i = 1)$$
$$f_2 = 1046 \quad (i = 2)$$
$$f_3 = 4393 \quad (i = 3)$$

enzovoorts. Alle absolute frequenties bij elkaar opgeteld resulteert in 104591 dienstplichtigen.

3.5.2. De cumulatieve absolute frequentie.

In de vijfde kolom staan de aantallen militairen cumulatief vermeld: er wordt uitgegaan van 105 militairen in de eerste klasse, vervolgens wordt het aantal van de tweede klasse (1046) erbij opgeteld en dan vind je 1151 militairen. Van deze 1151 militairen kan gezegd worden dat hun lengte minimaal 154.5 cm is en maximaal 164.4 cm. Daarna worden 4393 militairen opgeteld bij het totaal van de eerste twee klassen, enzovoorts. Het overkoepelende symbool voor de cumulatieve frequentie is:

$$f_{i,c}$$

Dus overkoepelend voor:

$$f_{1,c} = 105 \quad (i = 1)$$
$$f_{2.c} = 1151 \quad (i = 2)$$
$$f_{3,c} = 5544 \quad (i = 3)$$

enzovoorts. Uiteindelijk kom je weer uit bij een totaal van 104591.

3.5.3. Het klassemidden

Vervolgens wordt in de zesde kolom het begrip klassemidden geïntroduceerd. Voor verschillende berekeningen blijkt het handig te zijn om gebruik te maken van dit begrip, zoals verderop uitgelegd zal worden.

$$\text{Klassemidden} = \frac{(\text{kleinste waarde} + \text{grootste waarde})}{2}$$

Laten we naar de eerste klasse kijken. De kleinste waarde is 154.5 cm en de grootste waarde is 159.4 cm. Het klassemidden van de eerste klasse zal dus zijn:

$$\frac{(154{,}5 + 159{,}4)}{2} = 156{,}95$$

Voor het klassemidden gebruiken we het algemene symbool: k_i .

De letter i geeft hier aan om welke klasse het gaat. Spreken we over het klassemidden van de eerste klasse, dan moeten we i 1 nemen, dus:

$$k_1 = 156{,}95\,\text{cm}$$

Spreken we over het klassemidden van de tweede klasse, dan is i = 2 en krijgen we:

$$k_2 = \frac{(159{,}5 + 164{,}4)}{2} = 161{,}95\,\text{cm}$$

Op analoge wijze vinden we de klassemiddens van alle andere klassen. De afstand tussen de klassemidden is 5 cm en dat is niet zo verwonderlijk want de klassen zijn onderling telkens 5 cm ten opzichte van elkaar verschoven.

3.5.4. De modale klasse

Een volgend bekend basisbegrip uit de statistiek is het begrip modale klasse.

De modale klasse is de klasse met de hoogste absolute frequentie

Het kost niet veel moeite om in te zien dat de modale klasse bij de dienstplichtigen gelijk is aan klasse 6 , d.w.z. de klasse die loopt van 179.5 tot en met 184.4 cm. Deze klasse bevat namelijk 29495 dienstplichtigen en dat is het grootste aantal dienstplichtigen in een klasse.

Het is overigens wel zo dat de klassen bij deze definitie van de modale klasse wel even breed moeten zijn. In ons geval is aan deze voorwaarde voldaan want alle klassen zijn precies 5 cm breed. Zijn de klassen echter niet even breed dan verstaat men onder de modale klasse die klasse welke de grootste frequentiedichtheid vertoont. Op het begrip frequentiedichtheid zal in het kader van de beperkte doelstelling van deze reader niet verder worden ingegaan.

3.5.5. De modus (symbool : m).

Een volgend basisbegrip uit de statistiek is de modus:

De modus m is het klassemidden van de modale klasse

De modus m is dus gelijk aan het klassemidden van de zesde klasse:

$$m = k_6 = \frac{179{,}5 + 184{,}4}{2} = 181{,}95\,\text{cm}$$

Onder de modus wordt in het geval van de dienstplichtigen verstaan de modale lengte, omdat we te maken hebben met de lengten van mannen. Een ander voorbeeld van een modus is het begrip modale inkomen. Dan heeft het begrip modus betrekking op een inkomensverdeling. Zoals bekend verdienen de meeste werkende mensen een inkomen dat gelijk is of Vrij dicht in de buurt ligt van het modale inkomen.

Zouden we de inkomensverdeling, net als bij de lengteverdeling van de militairen, in klassen verdelen dan zou de modale klasse van de inkomensverdeling de meeste inkomens bevatten.

Verdere voorbeelden van de modus zijn de modale levensduur van mensen , het modale gewicht, enz.

3.5.6. Het rekenkundig gemiddelde.

Het rekenkundig gemiddelde is ook weer een veel voorkomend basisbegrip in de statistiek.

Hebben we te maken met drie dienstplichtigen met lengten van respectievelijk 168.5 cm, 173.0 cm en 180.5 cm , dan is het rekenkundig gemiddelde van deze drie lengten gelijk aan

$$\frac{168,5+176,0+180,5}{3}=174,0\,\text{cm}$$

Dus je telt de waarden op en daarna deel je door het aantal waarden dat je hebt opgeteld.

Voor het rekenkundig gemiddelde worden verschillende symbolen gebruikt. Hebben we te maken met een steekproef, dan wordt gebruikt het symbool

$$\bar{x}$$

Berekenen we het rekenkundig gemiddelde van een populatie dan wordt gebruik gemaakt van het symbool

$$\mu\,(\text{de Griekse letter mu})$$

We zullen in deze cursus nog enkele Griekse letters tegen komen. Hier alvast de volledig lijst, al is het alleen maar omdat ze er zo mooi uit zien.

α	A	*alfa*	η	nu	M
β	B	*beta*	ξ	xi	Ξ
γ	Γ	*gamma*	o	omikron	O
δ	Δ	*delta*	π	*pi*	Π
ε	E	*epsilon*	ρ	*rho*	P
ζ	Z	*zèta*	σ of ς	*sigma*	Σ
η	H	*èta*	τ	*tau*	T
θ of ϑ	Θ	*thèta*	υ	*ypsilon*	Y
ι	I	*iota*	ϕ	*phi*	Φ
κ	K	*kappa*	χ	*chi*	X
λ	Λ	*lambda*	ψ	*psi*	Ψ
μ	M	*mu*	ω	*omega*	Ω

Grieks alfabet

Overigens worden in sommige statistiekboeken ook andere symbolen gebruikt voor de begrippen rekenkundig gemiddelde van een steekproef en rekenkundig gemiddelde van een populatie, maar dat is eigenlijk helemaal niet zo belangrijk en zorgt eigenlijk maar voor verwarring. In het vervolg worden de bovenstaande symbolen gebruikt omdat die het meest gangbaar zijn.

De 104591 dienstplichtigen vormen een populatie. Stel dat men om een of andere reden het rekenkundig gemiddelde van deze populatie wil berekenen. Het berekenen van het precieze rekenkundig gemiddelde zou dan inhouden dat van alle individuele dienstplichtigen de lengte opgeteld zou moeten worden en daarna gedeeld zou moeten worden door 104591. Dat is natuurlijk een Sisyfusarbeid , een ondoenlijke taak, en daarom stellen we ons nu ten doel uit de gegeven tabel van de lengten van de dienstplichtigen een zo goed mogelijke schatting van het rekenkundig gemiddelde te maken.

Het probleem dat nu opdoemt is : welke lengte moeten we nu gaan toekennen aan dienstplichtigen in een bepaalde klasse? Laten we eens gaan kijken naar de vijfde klasse. Er zijn dan 26880 militairen met een lengte die kan variëren van 174.5 cm tot en met 179.4 cm. Welke lengte deze 26880 militairen individueel hebben weten we niet, maar we zullen ze toch een lengte moeten toekennen, want anders kunnen we geen rekenkundig gemiddelde berekenen. Wat moeten we nu doen?

Om verder te komen moet er een aanname worden gemaakt. We veronderstellen dat de militairen qua lengte homogeen verdeeld liggen in het interval 174.5 cm t/m 179.4 cm. Met homogeen verdeeld wordt bedoeld dat er evenveel militairen zijn met een lengte van 174.5 cm, 174.6 cm , enz. tot en met 179.4 cm.

Als hiervan uitgegaan wordt dan mag de lengte van elke militair voor wat betreft de berekening van het rekenkundig gemiddelde gelijkgesteld worden aan het klassemidden, dus gelijkgesteld worden aan 176.95 cm.

Met behulp van de veronderstelling van een homogene verdeling van de lengten in iedere klasse is het verder vrij gemakkelijk om het rekenkundig gemiddelde van de lengte te berekenen.

Eerst berekenen we telkens per klasse de absolute frequentie maal het klassemidden

105	x	156,95	=	16479,75
1046	x	161,95	=	169399,7
4393	x	166,95	=	733411,4
14224	x	171,95	=	2445817
26880	x	176,95	=	4756416
29495	x	181,95	=	5366615
18931	x	186,95	=	3539150
7321	x	191,95	=	1405266
1778	x	196,95	=	350177,1
418	x	201,95	=	84415,1
Totaal				18867147

Vervolgens worden de uitkomsten bij elkaar opgeteld:

16479.75 + 169399.70 + + 350177.10 + 84415.10 = 18867147

En dit getal moet natuurlijk door 104591 gedeeld worden teneinde de gemiddelde lengte te verkrijgen:

$$\mu = \frac{18867147}{104591} = 180{,}38978\,\text{cm}$$

Bij de interpretatie van het resultaat van onze berekening is het belangrijk zich te realiseren dat de afzonderlijke militairen op 1 mm nauwkeurig gemeten zijn. Dat betekent dat de militairen in principe allemaal 1 mm langer kunnen zijn dan de gemeten waarde, maar ze kunnen ook allemaal 1 mm kleiner zijn. Dat zal in de praktijk natuurlijk niet het geval zijn: sommigen zijn 1 mm langer en anderen 1 mm kleiner. Maar het is belangrijk om zich te realiseren dat het rekenkundig gemiddelde ook slechts op 1 mm nauwkeurig bekend is. Het antwoord 180.38978 cm is daarom een zinloos antwoord: er wordt een veel grotere nauwkeurigheid opgegeven dan waargemaakt kan worden. Het correcte antwoord is derhalve:

$$\mu = 180{,}4$$

Het wonderlijk is nu dat volgens het CBS (Centraal Bureau voor de Statistiek) het echte precieze gemiddelde van alle 104591 dienstplichtigen (dus alle 104591 lengten individueel optellen en door 104591 delen) ook 180.4 cm is. Dat betekent dus dat de veronderstelling van homogeen verdeelde klassen een goed werkende aanname is geweest. Per klasse zullen er waarschijnlijk fouten gemaakt zijn, maar de gemaakte fouten in de verschillende klassen zijn blijkbaar tegen elkaar weggevallen.

3.5.7. De mediaan (symbool m_e).

De formele definitie voor de mediaan is als volgt:

De mediaan is de middelste waarde, indien alle waarden naar opklimmende grootte gerangschikt worden.

Laten we ter illustratie van deze definitie een paar voorbeelden bekijken.

Een eerste voorbeeld: we kijken naar de onderstaande 9 getallen, die al naar opklimmende grootte gerangschikt zijn:

1,1,2,3,5,6,8,14,23

Het zijn 9 waarden, dus een oneven aantal. De middelste waarde is 5 omdat er evenveel waarden onder 5 zitten als dat er boven 5 zitten. Er zijn nl, 4 waarden onder 5 en er bevinden zich 4 waarden boven 5. De mediaan is in dit geval dus 5, want 5 staat precies in het midden , n.l. op de vijfde plaats.

Een vaak gebruikt symbool voor de mediaan is:

$$m_e$$

Dus we kunnen stellen dat in dit eerste voorbeeld geldt:

$$m_e = 5$$

Als we de mediaan zelf even niet meerekenen , dan bevindt zich dus 50 % van de waarden boven de mediaan en de andere 50 % zit onder de mediaan qua grootte.

Nu een tweede voorbeeld. Stel dat het totale aantal waarden nu even is. Bijvoorbeeld de waarden:

1, 10, 20, 22, 48, 133, 219, 220, 410, 411, 412, 470, 480, 500

We hebben hier te maken met 14 waarden , die al naar opklimmende grootte gerangschikt zijn. Maar wat is nu de mediaan? Als je 219 als mediaan neemt, dan zitten er 6 waarden onder en 7 waarden boven en dat is niet gelijk aan elkaar. En als je 220 voor de mediaan neemt, dan zitten er 7 waarden onder en 6 erboven en dat is ook weer niet gelijk aan elkaar.

Statistici hebben in deze situatie van een even aantal waarden voor de volgende oplossing gekozen : als je de mediaan wilt berekenen , neem dan de twee meest centrale waarden , tel ze op en deel dan door 2 . De mediaan wordt dan:

$$m_e = \frac{219 + 220}{2} = 219{,}5$$

Boven 219.5 bevinden zich 7 waarden en onder 219.5 bevinden zich ook 7 waarden Dus 50 % van de waarden bevindt zich boven de mediaan en de andere 50 % bevindt zich onder de mediaan , precies zoals we het zouden willen hebben . Op deze wijze is dus het probleem van de berekening van de mediaan bij een even aantal waarden opgelost.

Laten we nu als derde illustratie van de berekening van de mediaan terugkeren naar de frequentietabel van de militairen en ons afvragen hoe we de mediaan (beter gezegd: de mediane lengte) voor deze frequentietabel zouden kunnen berekenen.

We willen dus die lengte bepalen , die als eigenschap heeft dat 50 % van de militairen een lengte heeft die groter is dan deze mediane lengte en de andere 50 % van de militairen heeft dan een lengte kleiner dan deze mediane lengte. We moeten hier echter wel in het achterhoofd houden dat we slechts een schatting van de mediaan kunnen maken. Voor een precieze bepaling zouden we 104591 individuele lengten moeten gaan rangschikken naar opklimmende grootte en daarna dan de meest centrale waarde eruit pikken als zijnde de mediaan. Ondoenlijk dus. We moeten daarom iets anders bedenken en dat zal uitgelegd worden in de volgende paragrafen, waarbij gebruik gemaakt wordt van de nieuwe begrippen relatieve frequentie, cumulatieve relatieve frequentie en Lineair interpoleren.

3.5.8. De relatieve frequentie

We kijken allereerst naar het begrip relatieve frequentie. Het algemene symbool voor dit begrip is:

$$f_{i,r}$$

Voor de eerste klasse , dus voor i = 1, krijgen we derhalve:

$$f_{1.r}$$

Voor de tweede klasse geldt i = 2 , enzovoorts.

Wat is nu de betekenis van deze relatieve frequentie? Het woord relatief houdt in het algemeen in dat er een vergelijking gemaakt wordt met iets anders. Zo ook hier: kijken we naar de eerste klasse , dan vergelijken we de 105 dienstplichtigen uit deze klasse met het totale aantal van 104591 dienstplichtigen. Dit doen we door deze 105 militairen als percentage van het totale aantal dienstplichtigen op te geven:

$$f_{1.r} = \frac{105}{104591} \times 100\% = 0{,}1\%$$

Je kan ook redeneren: 1 % van 104591 is gelijk aan 1045,91 , zeg 1046. Hiervan weer een tiende deel , dus 104,6 , zeg maar 105 , is gelijk aan 0.1 % . De 105 dienstplichtigen in de eerste klasse vormen dus een dikke 0.1 % van het totale aantal dienstplichtigen.

Evenzo voor de tweede klasse. Dan is i = 2 en krijgen we:

$$f_{2,r} = \frac{1046}{104591} \times 100\% = 1{,}0\%$$

Voor de derde klasse wordt de relatieve frequentie:

$$f_{3,r} = \frac{4393}{104591} \times 100\% = 4{,}2\%$$

Het vorige even resumerend : slechts 0.1 % van alle militairen heeft een lengte variërend van 154.5 cm t/m 159.4 cm , slechts 1 % heeft een lengte variërend van 159.5 cm t/m 164.4 cm en een iets groter aantal , n.l. 4.2 % , heeft een lengte variërend van 164.5 tot en met 169.4 cm . Alle berekende relatieve frequenties voor alle tien klassen staan vermeld in de frequentietabel voor de militairen, evenals de cumulatieve relatieve frequentie , die in de nu volgende paragraaf geïntroduceerd zal worden.

3.5.9. De cumulatieve relatieve frequentie.

Tel je de relatieve frequenties van klasse 1 en klasse 2 bij elkaar op , dan vind je dat 0.1 % + 1.0 % = 1.1 % van alle dienstplichtigen maximaal l64.4 cm lang is. Tel je de derde klasse erbij op dan vind je dat

$$0.1\% + 1.0\% + 4.2\% = 5.3\%$$

van alle militairen een lengte heeft die kan variëren van 154.5 tot en met 169.4 cm, dus maximaal 169.4 cm lang . Op deze wijze ontstaat dan een nieuwe serie getallen, welke vermeld staan in de laatste kolom van de frequentietabel en welke getallen de cumulatie relatieve frequenties genoemd worden.

Als algemeen symbool voor de cumulatieve relatieve frequentie wordt gebruikt:

$$f_{i,r,c}$$

Zo is voor i = 6 de cumulatieve relatieve frequentie gelijk aan

$$f_{6,r,c} = 72{,}8\%$$

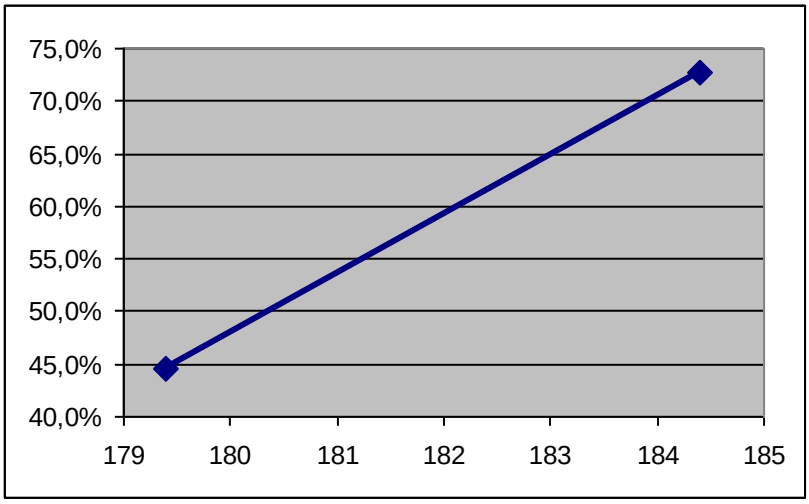
wat inhoudt dat 72.8 % van het totale aantal dienstplichtigen een lengte heeft van maximaal 184.4 cm , dus of 184.4 cm lang is, of kleiner dan 184.4 cm.

Laten we nu weer terugkeren naar het probleem van het berekenen van de mediaan (de mediane lengte) uit de frequentietabel van de lengten der militairen. Het zojuist geïntroduceerde begrip cumulatieve relatieve frequentie kan ons helpen om een schatting te maken van de mediane lengte. De echte mediane lengte is namelijk niet met behulp van de tabel te bepalen. We zouden dan 104591 individuele lengte- waarden moeten gaan rangschikken naar opklimmende grootte en daarna zouden we de “middelste” waarde , de meest centrale waarde, moeten gaan bepalen. Een niet erg realistisch klusje , maar welke lengtewaarde zou dat zijn? De 104591 militairen vormen een oneven aantal . Trekken we er 1 vanaf (de mediaan) dan houden we 104590 over. Delen we 104590 door 2 dan krijgen we het getal 52295.

Je kan je dus voorstellen dat de 104591 lengtewaarden naar opklimmende grootte gerangschikt worden : eerst de 52295 kleinste waarden , dan de 52296-ste waarde (= de mediaan , want deze waarde staat in het midden) en dan de 52295 grootste waarden . In dat geval bevinden zich dan

50 % van de lengtewaarden onder de mediaan en ook 50 % van de waarden bevinden zich boven de mediane lengte. Maar dat is in de praktijk ondoenlijk.

We keren terug naar het probleem : het begrip cumulatieve relatieve frequentie kan ons nu helpen bij het maken van een schatting van de mediane lengte. Kijk je in de tabel naar de kolom met cumulatieve relatieve frequenties dan zie je dat 44.6 % van de militairen een lengte heeft van maximaal 179.4 cm en 72.8 % van de militairen heeft een lengte van maximaal 184.4 cm. In een grafiek weergegeven zie je het volgende:

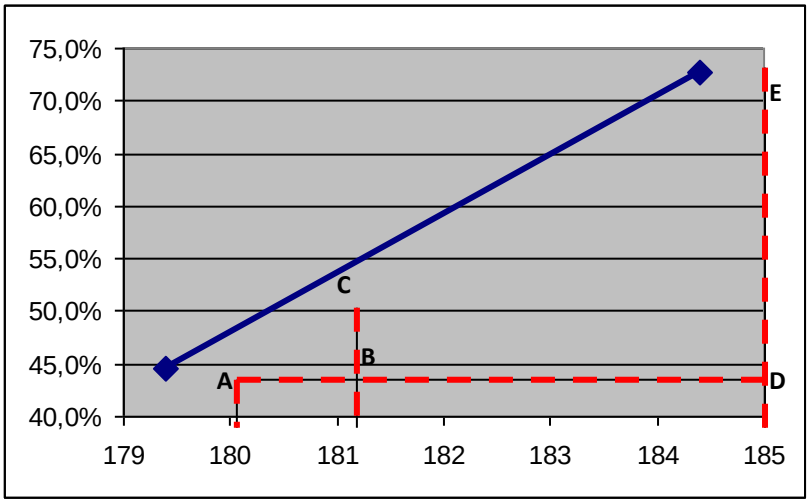


Om de mediaan (dat is de mediane lengte) te vinden, moeten we kijken waar de grafiek de 50% lijn snijdt. Zo op het eerste gezicht zal dat tussen de 180 en de 181 cm zijn.

In de volgende paragraaf zullen we dat precies uitrekenen met behulp van lineaire interpolatie.

3.5.10. Lineair interpoleren.

In bovenstaande grafiek gaan we wat hulplijnen plaatsen:



Je ziet zo twee driehoeken ontstaan: de grote driehoek: ΔADE en de kleinere driehoek ΔABC . Beide driehoeken hebben precies dezelfde vorm, ze verschillen alleen in grootte. Dergelijke driehoeken noemen we gelijkvormig..

In deze tekening is de grote driehoek ongeveer vier keer zo groot als de kleine driehoek. Hiermee bedoelen we dat een zijde van de grote driehoek ongeveer vier keer zo groot is als de overeenkomstige zijde van de kleine driehoek. Uit deze gelijkvormigheid volgt bijvoorbeeld:

$$\frac{DE}{DA} = \frac{BC}{BA}$$

De lengte van DE is:

$$DE = 72{,}8\% - 44{,}6\% = 28{,}2\%$$

Evenzo geldt:

$$DA = 184{,}4cm - 179{,}4cm = 5cm$$

$$BC = 50\% - 44{,}6\% = 6{,}4\%$$

$$BA = M_e - 179{,}4$$

Dus geldt dat:

$$\frac{28{,}2}{5} = \frac{6{,}4}{M_e - 179{,}4}$$

en hieruit volgt dan weer dat:

$$28{,}2 \times \left(M_e - 179{,}4\right) = 5 \times 6{,}4$$

en dus dat:

$$28{,}2 \times M_e = 32 + 28{,}2 \times 179{,}4 = 5091{,}08$$

$$M_e = \frac{5091{,}08}{28{,}2} = 180{,}5$$

3.5.11. Kwantielen: kwartielen, docielen en percentielen.

Zoals in de vorige paragraaf beschreven werd , deelt de mediaan de getalgegevens in twee gelijke delen: 50 % van alle getallen (lengtewaarden waren het in het voorgaande voorbeeld) bevindt zich onder de mediaan en de andere 50 % bevindt zich boven de mediaan.

Er worden in de praktijk van de statistiek ook andere grootheden gebruikt die de getalgegevens in meer dan twee delen verdelen. Zo verdelen kwartielen de getalgegevens in vier gelijke delen.

De kwartielen zijn dus die drie getallen die de populatie of steekproef in vieren delen

25% van de populatie valt onder het eerste kwartiel (Q_1)

25% van de populatie valt tussen Q_1 en Q_2

25% van de populatie valt tussen Q_2 en Q_3

5% van de populatie valt boven het derde kwartiel: Q_3

Als we kijken naar de definitie van de mediaan, dan blijkt dat het tweede kwartiel identiek is aan de mediaan. Tenslotte het derde kwartiel

Op dezelfde manier kun je ook spreken over docielen

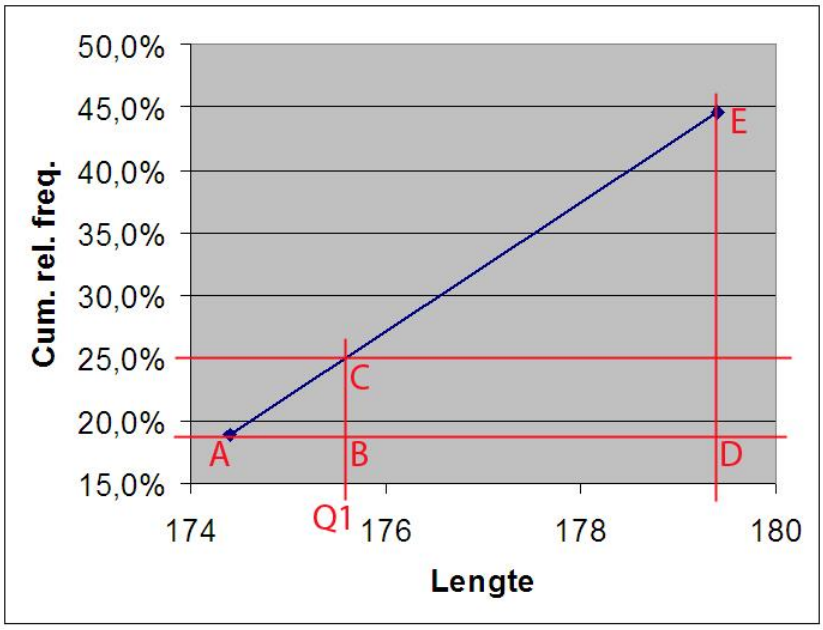
Docielen verdelen de getalgegevens in tien gelijke delen .

Of kwantielen:

De algemene naam voor de mediaan , kwartielen en docielen is kwantielen.

En percentielen:

Tot de kwantielen rekent men ook de percentielen. Dit zijn grootheden die de getalgegevens in 100 gelijke delen verdelen.



De berekeningen van de mediaan, kwartielen, docielen en percentielen verlopen allemaal op dezelfde wijze, via de techniek van het lineair interpoleren.

Als eerste concrete illustratie nu de berekening van het eerste kwartiel. Laten we eerst maar en schets maken van de grafiek waar het om gaat.:

Het punt A heeft coördinaten: (174.4, 18.9)

Het punt B heeft coördinaten: (Q1, 18.9)

Het punt C heeft coördinaten: (Q1, 25)

Het punt D heeft coördinaten: (179.4, 18.9)

Het punt E heeft coördinaten: (179.4, 44.6)

We gaan weer precies op dezelfde manier rekenen als in het eerdere voorbeeld:

$$\frac{DE}{DA} = \frac{BC}{BA}$$

De lengte van DE is:

$$DE = 44{,}6\% - 18{,}9 = 25{,}7\%$$

Evenzo geldt:

$$DA = 179{,}4cm - 174{,}4cm = 5cm$$

$$BC = 25\% - 18{,}9\% = 6{,}1\%$$

$$BA = Q1 - 174{,}4$$

Dus geldt dat:

$$\frac{25{,}7}{5} = \frac{6{,}1}{Q1 - 174{,}4}$$

en hieruit volgt dan weer dat:

$$25{,}7 \times (Q1 - 174{,}4) = 5 \times 6{,}1$$

en dus dat:

$$25{,}7 \times Q1 = 30{,}5 + 25{,}7 \times 174{,}4 = 4512{,}58$$

$$Q1 = \frac{4512{,}58}{25{,}7} = 175{,}6$$

De precieze waarde van het eerste kwartiel kan er een millimetertje naast zitten, maar dat is niet zo belangrijk. Alleen als er heel speciale redenen zijn , kunnen we extra aandacht gaan besteden aan de kwestie van de nauwkeurigheid.

3.6. De standaarddeviatie ofwel standaardafwijking.

In het kader van de introductie van statistische basisbegrippen mag de standaardafwijking ofwel standaarddeviatie zeker niet overgeslagen worden omdat het een zeer veel gebruikte grootheid is. We zullen de standaardafwijking introduceren door o.a. wederom gebruik te maken van de lengtetabel van de dienstplichtigen.

Uit de frequentietabel van de dienstplichtigen hebben we het rekenkundig gemiddelde, d.w.z. de gemiddelde lengte , berekend. Het resultaat was 180.4 cm. Een logische vraag is nu: hoe zijn de lengten van alle dienstplichtigen nu gerangschikt om deze gemiddelde lengte ? Zitten alle lengtewaarden heel dicht bij de gemiddelde lengte of is het wellicht zo dat de lengtewaarden zeer verspreid zijn rondom de gemiddelde lengte

Men heeft geprobeerd om deze spreiding van de waarden rondom de gemiddelde waarde uit te drukken in een getal. Het bekendste voorbeeld van deze pogingen is de standaardafwijking. Is de standaardafwijking groot, dan is er een grote spreiding van de lengtewaarden om de gemiddelde lengte heen, maar is de standaardafwijking daarentegen klein dan zullen de meeste lengtewaarden dicht in de buurt van de gemiddelde lengte zitten.

De lengten van de Nederlandse jonge mannen vormen een zogenaamde normale verdeling, waarover veel meer in deze reader te vinden is. Deze normale verdeling heeft als eigenschap dat zo'n 68 % van alle lengtewaarden zich binnen een afstand gelijk aan de standaardafwijking van de gemiddelde lengte bevindt. Weet je dus de standaardafwijking dan heb je al een duidelijke indruk van de lengte van bijna 70 % van de dienstplichtigen.

Behalve de lengteverdeling van de militairen (een normale verdeling) heb je echter ook nog verschillende andere typen verdelingen , zoals de inkomensverdeling, de gewichtsverdeling, een levensduurverdeling , enzovoort enzovoort. Al deze verdelingen hebben een min of meer verschillende vorm.

Voor al deze verdelingen geldt het theorema van de Rus Chebyshev

Tenminste 75 % van de waarden van een willekeurige verdeling zal zich bevinden binnen een afstand van tweemaal de standaardafwijking tot de gemiddelde waarde en bijna 90 % zal zich op een maximale afstand van driemaal de standaardafwijking t.o.v. de gemiddelde waarde bevinden.

De stelling van Chebyshev wordt slechts vermeld om het belang van het begrip standaarddeviatie te onderstrepen. Blijkbaar wordt een verdeling overzichtelijker indien je de beschikking hebt over de standaardafwijking van die verdeling.

Er wordt onderscheid gemaakt tussen de standaardafwijking voor een steekproef en de standaardafwijking voor een populatie. De reden van dit onderscheid is een wiskundig technische en het voert te ver dat hier te bespreken.

4De standaardafwijking voor een steekproef (symbool: s).

Eerst geven we een eenvoudig voorbeeld van de techniek van het berekenen van de standaarddeviatie. Ga uit van drie dienstplichtigen met als lengten:

174.0 cm , 176.6 cm en 182.0 cm.

Deze drie lengten vormen een steekproef uit de populatie van dienstplichtigen. Het gemiddelde van deze drie waarden is:

$$\frac{174{,}0+176{,}6+182{,}0}{3}=177{,}5\,\text{cm}$$

We berekenen de standaardafwijking van een steekproef (symbool s) door

Bereken het gemiddelde

Bereken voor iedere individuele waarde het verschil met dat gemiddelde.

Kwadrateer al die verschillen

Tel de kwadraten bij elkaar op

Deel de gekregen uitkomst door ÉÉN MINDER[1] dan het aantal getallen van de steekproef (de steekproefgrootte)

Trek de (vierkants)wortel uit dat getal, dat is de standaardafwijking.

Je merkt het: het is best veel rekenwerk. Vandaar dan ook dat we dat in de praktijk met een computer doen.

Hier doen we het bij ons voorbeeld gewoon met de hand.

1 Het gemiddelde is, zoals eerder uitgerekend: 177,5 cm

2 De verschillen met dat gemiddelde zijn achtereenvolgens: 3,5; 1,1 en 4,5 cm

3 De kwadraten zijn dan achtereenvolgens: 12,25; 1,21 en 20,25

4 Tel de kwadraten op: 12,25+1,21+20,25=33,71

5 Deel dit door 3-1=2: $\frac{33{,}71}{2}=16.9$

6 Neem hier de wortel van: $\sqrt{16{,}9}=4{,}1$

De standaardafwijking voor een populatie (symbool: σ, de griekse letter sigma).

Heb je te maken met een populatie, dan moet je precies hetzelfde doen, alleen moet je niet door (steekproefgrootte — 1) delen, maar slechts delen door de populatie grootte N.

Zouden we dus de drie lengtewaarden uit de vorige paragraaf opvatten als een populatie, dan wordt de standaardafwijking gelijk aan:

[1] Als het niet om een steekproef, maar om een populatie gaat, hoeven we niet één minder te nemen, we delen dan gewoon door de populatiegroott

$$\sigma = \sqrt{\frac{33{,}71}{3}} = 3{,}4$$

3.6.1. Intermezzo: handige schrijfwijze

Bij het berekenen van de standaarddeviatie moeten met heel veel getallen worden opgeteld, vermenigvuldigd en wortel getrokken. In de wiskunde is daar een handige schrijfwijze voor bedacht die we in deze paragraaf uitleggen.

We hebben bij de frequenties van de klassen al gewerkt met $f_1, f_2, f_3, \ldots$ .. Of: meer algemeen de frequentie van de klasse i geven we aan met f_i . Die i noemen we wel de index en is niets anders dan een tellertje om al die verschillende frequentie waarden (eentje voor elke klasse) uit elkaar te houden.

We gebruiken dit idee niet alleen voor frequenties, maar ook voor klassemidden, cumulatieve en relatieve frequenties en nog veel meer getallen die horen bij een klasse.

Laten we eens kijken naar een rij getallen. Bijvoorbeeld de rij

30, 65, 34, 80, 82, 80

De elementen van die rij geven we aan met x en een index. Concreet:

$$x_1 = 30, x_2 = 65, x_3 = 34, x_4 = 80, x_5 = 82 \text{ en } x_6 = 80 .$$

Als we een formule willen opschrijven voor het berekenen van de som van al die getallen, dan kan dat natuurlijk door:

$$som = x_1 + x_2 + x_3 + x_4 + x_5 + x_6$$

Maar je begrijpt dat dit vervelend wordt als het om nog veel langere rijen getallen gaat. We schrijven daarom:

$$som = \sum_{i=1}^{6} x_i$$

lees dit als de som voor i = 1 tot aan i =6 van x i. Σ is de Griekse (hoofdletter) sigma.

Dit kunnen we gebruiken bij het netjes opschrijven van de formule voor het gemiddelde μ van deze rij getallen:

$$\mu = \frac{\sum_{i=1}^{6} x_i}{6}$$

Een formule voor de standaardafwijking (bij een populatie) wordt dan:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{6}(x_i - \mu)^2}{6}}$$

Dit ziet er ingewikkelder uit dan het in werkelijkheid is. Het geheim van het begrijpen en kunnen gebruiken van dergelijke formules is dat je er stap voor stap naar kijk.

In dit geval:

Je moet de wortel ergens uit nemen. Waaruit? Uit een breuk!

Hoe zit die teller in elkaar? Die teller is een opsomming van iets (het sigma teken).

Wat precies moet je optellen? Je moet kwadraten optellen!

Hoe kom je aan al die kwadraten? Daarvoor moet je telkens het gemiddelde μ aftrekken van de waarde x_i.

Vergelijk dit eens met de procedure zoals hiervoor beschreven om een standaardafwijking te berekenen, Dat is precies hetzelfde. Deze formule vervangt dus de procedure beschrijving in 6 stappen zoals hiervoor.

Tenslotte geven we de formules voor de standaardafwijking van een populatie een steekproef:

Gegeven een rij van N getallen (opgevat als populatie), dan is de standaardafwijking van die getallen:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

waarbij μ het gemiddelde van de populatie is.

Als we dezelfde getallen kunnen opvatten als een steekproef, dan definiëren we de standaardafwijking:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Let op de verschillen met de vorige formule:

We schrijven s in plaats van σ

We schrijven $\bar{x}$ in plaats van μ

We delen door N-1 in plaats van door N.

3.6.2. De standaardafwijking als fout.

Een praktische toepassing van het begrip standaardafwijking bij het studieonderdeel komt voor bij de kwaliteitscontrole.

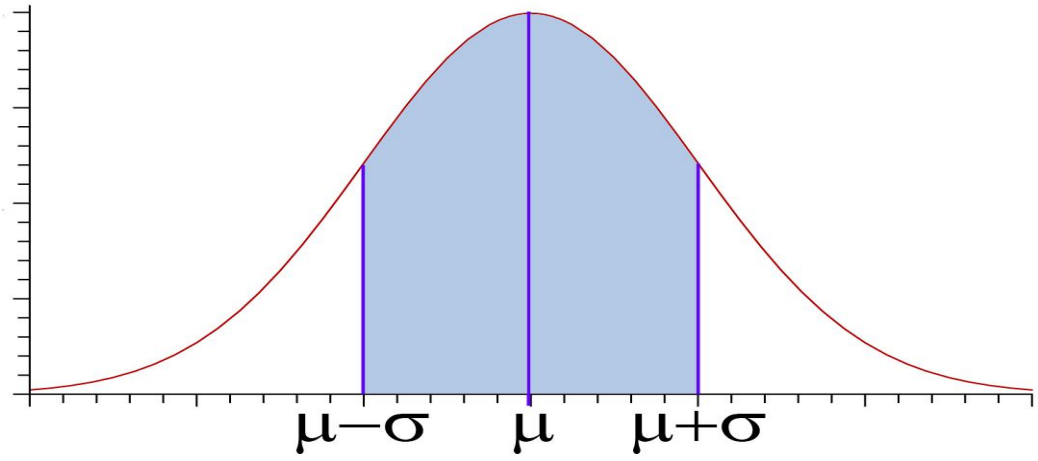
Bij het studieonderdeel kwaliteitscontrole worden metingen gedaan op verschillende doekconstrukties.

Stel we meten de treksterkte (symbool: X) van een bepaalde doekconstruktie. We vinden dan bijvoorbeeld X = 200 Newton. Als we de treksterkte daarna nog een keer meten, vinden we een iets andere waarde, bijvoorbeeld 205 Newton. Stel nu eens dat we de treksterkte 5 keer meten en de volgende 5 waarden vinden:

200 , 205 , 198 , 186 en tenslotte 212 Newton.

Wat moeten we nu als gevonden waarde voor de treksterkte X opgeven?

Om deze vraag te beantwoorden gaan we een klein statistisch uitstapje maken. In de praktijk blijkt vaak dat als je de treksterkte of überhaupt elke andere meetbare grootheid talloze malen meet, dat dan de meetresultaten ongeveer een normale verdeling (zie deze reader) gaan vormen. Elke normale verdeling wordt volledig gekarakteriseerd door de gemiddelde waarde μ en door de standaardafwijking σ. Deze begrippen zijn reeds besproken: de gemiddelde waarde is de meest centrale waarde en de standaardafwijking is een grootheid die de mate van spreiding rondom de gemiddelde waarde beschrijft. Zie als illustratie de onderstaande grafiek:



Horizontaal is de treksterkte uitgezet en verticaal is de frequentie uitgezet, d.w.z. het aantal keren dat een waarde van de treksterkte gevonden is voor elke bepaalde waarde van de treksterkte. Dus de grafiek laat zien dat de meest centrale waarde het meest frequent gevonden wordt en hoe meer de treksterktewaarde van de centrale waarde afwijkt, des te minder vaak zal hij als resultaat van een treksterktemeting gevonden worden.

Een bekende eigenschap van de normale verdeling is dat bijna 70 % van de gevonden treksterktewaarden zich op een maximale afstand gelijk aan de standaardafwijking van de centrale waarde bevindt. Zie het gearceerde gebied in de bovenstaande grafiek, dat zo'n 70 % van de totale oppervlakte onder de normale verdeling beslaat. Omdat dit percentage van 70 % zo hoog is, is het redelijk om aan te nemen dat de correcte treksterktewaarde in dit gebied moet liggen. Als antwoord voor de treksterktewaarde wordt dan opgegeven:

treksterkte X = de gemiddelde waarde ± standaardafwijking.

Als concrete illustratie gaan we terug naar de 5 gevonden treksterktewaarden. De gemiddelde waarde is:

$$\bar{x} = \frac{200 + 205 + 198 + 186 + 212}{5} = 200{,}2$$

en de standaardafwijking s is gelijk aan

$$s = \sqrt{\frac{(200 - 200{,}2)^2 + (205 - 200{,}2)^2 + (198 - 200{,}2)^2 + (186 - 200{,}2)^2 + (212 - 200{,}2)^2}{5 - 1}} = 9{,}6N$$

Het uiteindelijke antwoord voor de treksterkte X wordt:

treksterkte X = 200.2 ± 9.6 Newton

Dit voorbeeld diende als illustratie van een toepassing van de standaardafwijking, waarbij deze grootheid dan geïnterpreteerd wordt als een fout.

Vermeld kan nog worden dat in de confectiepraktijk, om te komen tot een minimumkwaliteitseis, de gevonden waarde van 200.2 Newton verminderd wordt met een halve standaardafwijking , dus 200.2 — (1/2)x9.6 = 200.2 — 4.8 = 195.4 Newton, afgerond: 195 Newton. Er wordt als het ware een extra kwaliteitseis ingebouwd, want deze uiteindelijke waarde van 195 Newton moet vergeleken worden met de minimale kwaliteitseis en daar zeker gelijk aan zijn, maar liefst groter dan de minimale kwaliteitseis. De minimale kwaliteitseis zelf is op zijn beurt weer afhankelijk van de prijs-kwaliteitverhouding.

3.6.3. De berekening van de standaarddeviatie uit de frequentietabel.

Als laatste voorbeeld de frequentietabel van de militairen. We hebben dan te maken met een zeer grote populatie. We zien dan een probleempje opdoemen: welke lengte moeten we nu nemen voor dienstplichtigen in een bepaalde klasse?

Dit probleem kunnen we oplossen door iedere dienstplichtige een lengte toe te kennen gelijk aan het klassemidden van de klasse waarin de dienstplichtige zich bevindt. Dus alle dienstplichtigen in de eerste klasse krijgen de lengte 156.95 cm toegekend. Vervolgens wordt de afstand tussen het klassemidden van de eerste klasse en de gemiddelde lengte berekend

156.95 — 180.4 = 23.45 cm

De gevonden waarde moet gekwadrateerd worden:

$$(-23{,}45)^2 = 549{,}90$$

Daarna moet het gevonden getal vermenigvuldigd worden met het aantal dienstplichtigen in de eerste klasse:

105 x 549.90 = 57739.76

Dit moet voor iedere klasse gedaan worden .

We krijgen voor klasse 2 als resultaat: 391803.28 ;
voor klasse 3 wordt het 1002923.50 , enzovoorts.

Vervolgens moeten alle resultaten voor iedere klasse bij elkaar opgeteld worden (dat wordt dan : 5078301) en gedeeld worden door het totale aantal militairen

$$\frac{5078301}{104591} = 48.5539$$

En tenslotte moet de vierkantswortel getrokken worden , dus de standaarddeviatie is gelijk aan

$$\sigma = \sqrt{48.5539} = 6.968\,\text{cm}$$

We kunnen dit ook weer in een formule schrijven:

Als je niet de beschikking hebt over de afzonderlijke meetwaarden, maar enkel de gegevens van een frequentietabel hebt, dan wordt de standaardafwijking (voor een populatie) berekend met de volgende formule:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{A} f_1 (k_i - \mu)^2}{N}}$$

waarbij A het aantal verschillende klassen, k_i en f_i het klassemidden en de frequentie van klasse i zijn. N is de populatiegrootte en μ het gemiddelde (hier de gemiddelde lengte).

Die gemiddelde lengte μ wordt hier natuurlijk berekend zoals eerder aangegeven: we doen alsof alle soldaten in een klasse dezelfde lengte hebben: het klassemidden van die klasse en vermenigvuldigen. In formulevorm:

$$\mu = \frac{\sum_{i=1}^{A} f_i \cdot k_i}{N}$$

4. .De normale verdeling

De normale verdeling is de belangrijkste verdeling uit de statistiek. Deze verdeling werd ontdekt door de Duitse wis- en natuurkundige Carl Friedrich Gauss , die leefde van 1777 tot 1855), een wetenschapper die met kop en schouders uitstak boven al zijn wetenschappelijke tijdgenoten.

Als zuivere wiskundigen kunnen slechts Archimedes en Isaac Newton tot zijn gelijken gerekend worden. Zijn mathematische genie openbaarde zich reeds op driejarige leeftijd toen hij al uit het hoofd kon rekenen. Vreemd genoeg bleek deze buitengewoon intelligente man een gebrek aan zelfvertrouwen te hebben , dat hem remde bij het publiceren van zijn bevindingen. Pas na zijn dood zijn veel ongepubliceerde, hoogst belangrijke wiskundige en natuurkundige ideeën gevonden.

Eigenlijk heeft Gauss de normale verdeling herontdekt, samen met de Fransman Laplace, en is de oorspronkelijke ontdekker ervan de beroemde Franse wiskundige de Moivre (1667 — 1754). (wat wetenswaardigheden over het leven van de Moivre zijn te vinden op blz. 48).

De normale verdeling zal in de volgende paragrafen behandeld worden. Er komen zaken aan de orde als : wat is nu eigenlijk de praktische betekenis van de normale verdeling? En : wat moeten we weten van de normale verdeling om ermee te kunnen rekenen ? En waarom staat deze verdeling eigenlijk zo centraal in de statistiek? Ter beantwoording van de laatste vraag kunnen veel redenen aangevoerd worden.

Allereerst blijkt het zo te zijn dat vele grootheden een normale verdeling volgen, ofwel zoals men zegt: normaal verdeeld zijn. Voorbeelden hiervan zijn legio. We noemen slechts meetfouten, het intelligentiequotiënt en bv. de lengte (zie het voorbeeld van de lengteverdeling elders).

Van de laatstgenoemde grootheid zullen we aantonen dat de frequentieverdeling van de dienstplichtigen een normale verdeling volgt en er zullen een aantal rekenvoorbeelden gepresenteerd worden. Interessant is het trouwens om te vermelden dat een wiskundige afleiding bestaat die verklaart waarom deze normale verdeling overal zo vaak voortkomt. Deze afleiding is buitengewoon interessant en doet het inzicht in het wezen van de normale verdeling geweldig toenemen , maar zal in het kader van de beperkte doelstelling van deze reader niet gegeven worden.

Ook vele verschijnselen in de natuur zijn normaal verdeeld. Om slechts een enkel voorbeeld van zo'n normaal verdeelde grootheid te noemen : de afstand tussen de bladeren in planten is normaal verdeeld.

Maar het belangrijkste is dat de normale verdeling een beslissende rol speelt in de Centrale Limietstelling. Op dit behoorlijk moeilijke begrip (en ook woord !) zal totaal niet ingegaan worden, maar het is wel het vermelden waard dat een praktisch gevolg van de Centrale Limietstelling is dat de normale verdeling van essentieel belang is bij de interpretatie van steekproefresultaten , dus bij kwaliteitscontrole. En kwaliteitscontrole komt veelvuldig voor in de confectie-industrie. Ook hiervan zullen voorbeelden gegeven worden.

Carl Friedrich Gauss werd op 30april 1777 in Brunswijk (Duitsland) geboren in een arm en ongeletterd gezin. Zijn vader was tuinman. koopmanshulp en beheerder van een verzekeringsfonds. De jonge Carl was zeer 'voorlijk': hij leerde zichzelf rekenen en lezen; het verhaal gaat dat hij op driejarige leeftijd een fout ontdekte in een berekening van zijn vader en dat hij, toen hij acht jaar was, tijdens zijn eerste reken/es op school de som van 1.2 100 berekende door 1 bil 100.2 bij 99. enz. te nemen om zo tot het totaal van 50 x 101=5050 te komen. Op het gymnasium was Gauss 'even goed' in klassieke talen als in wiskunde. In de periode van zijn veertiende tot zijn zeventiende levensjaar had hij al bijna al zijn mathematische ontdekkingen gedaan. De ideeën kwamen sneller dan ze konden worden opgeschreven. Echter, door gebrek aan ervaring in het publiceren en een grote bescheidenheid, misschien zelfs een gebrek aan zelf vertrouwen. bleef dit alles verborgen. Een deel van zijn wiskundige vindingen werd (door anderen) enige tientallen jaren later herontdekt.

Omstreeks zijn negentiende jaar construeerde Gauss met passeren liniaal (zonder schaalverdeling!) een regelmatige zeventienhoek: de eerste ontdekking in de Eudidische meetkunde sinds tweeduizend jaar. Het deed hem besluiten niet in de klassieke, maar in de mathematische wetenschappen verder te gaan.

In 1795 begon Gauss een studie aan de Universiteit van Göttingen. waar hij toe - gang had tot de werken van Euler. Fermot. Lagrange en Legendre. Zes jaar later verscheen zijn eerste boek: 'Disquisitiones arithmeticae' (Over Getallentheorie). In zijn proefschrift bewees de twintigjarige Gauss de hoofdstelling van de algebra: Elk polynoom van derde graad met complexe coëfficiënten heeft ten minste één complexe wortel. Hij slaagde hier waar groten

als Newton, D 'Alembert en Lagrange geen succes hadden gehad.

Het principe der kleinste kwadraten is van Gauss afkomstig; het werd merkwaardig genoeg niet in eerste instantie ontwikkeld voor het oplossen van fysische problemen. maar voor het onderzoek naar regelmaat in de verdeling van de priem getallen. Het principe der kleinste kwadraten fundeerde hij op datgene wat nu grootste aannemelijkheid (maximum likelyhood) wordt genoemd.

Het is beslist ondoenlijk de verdiensten van Gauss hier naar waarde te vermelden. Zo werkte en publiceerde hij op het gebied van landmeetkunde (het veidwerk deed hij zelf). geodesie. differentiaalmeetkunde (de beroemde 'stelling van Gauss') is daar spreekwoordelijk). mathematische statistiek (hij analyseerde de efficiëntie van schatters; verder heeft de ontwikkeling en het gebruik van de normale verdeling geleid tot de wild verspreide naamgeving 'Gausverdeling'). optica (de berekening van een stralengang gaat nog steeds via de zo genoemde 'Gaussische optica') en magnetisme (in 1839 stelde hij een algemene theorie op over het aardmagnetisme en werkte op dit gebied ook samen met Weber). Verder publiceerde hij nog artikelen over mechanica, kristallografie en capillariteit.

Veel wiskundige ontdekkingen van Gauss zijn pas aan het licht gekomen na de vondst van zijn dagboek in 1799; sommige daarvan waren inmiddels al aan anderen toegeschreven.

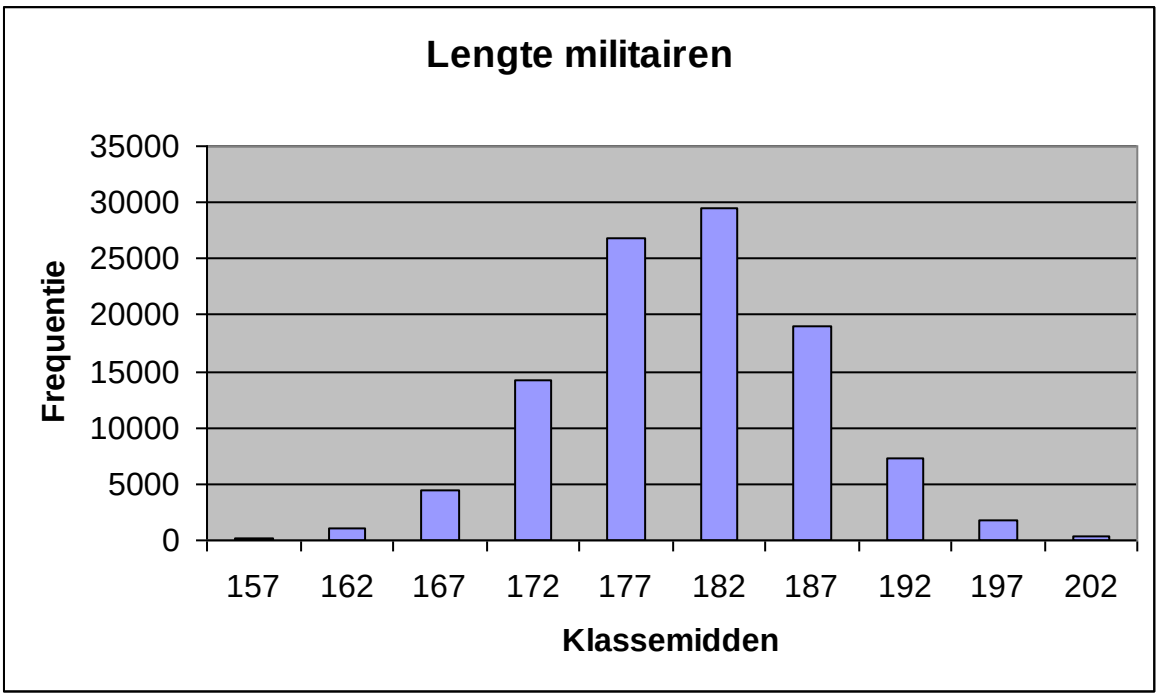
Gauss wordt als wis-, natuur- en sterrekundige wel op één lijn gesteld met groten als Newton en Archimedes. Hoewel zijn brief wisseling met tijdgenoten een totaal van zeker twintigduizend brieven moet hebben bevat, zocht Gauss geen communicatie met vakgenoten; hij schijnt zich nauwelijks bewust geweest te zijn van het werk van wiskundige tijdgenoten. Zijn publicaties waren bondig en moeilijk. zonder 'aanloop/es uit de praktijk', en weinig toegankelijk.

Gauss heeft in zijn persoonlijke leven maar één periode van geluk gekend. In 1805 trouwde hij met Johanna Osthof; ze kregen een zoon en een dochter. Na de geboorte van een derde kind (in 1809) overleed Johanna; de baby stierf kort daarop. Een klein jaar later hertrouwde Gauss; zijn tweede vrouw. Minna Wakieck. overleed in 1831 na een lange periode van ziekelijkheid en neerslachtigheid. Ook Gauss zelf had een zwakke gezondheid. Het belette hem echter niet nog op tweeénzestigjarige leeftijd vloeiend Russisch te leren lezen en spreken. Hij verliet Göttingen ze/den en leefde altijd sober. Hij over/eed daar op 23 februari 1855.

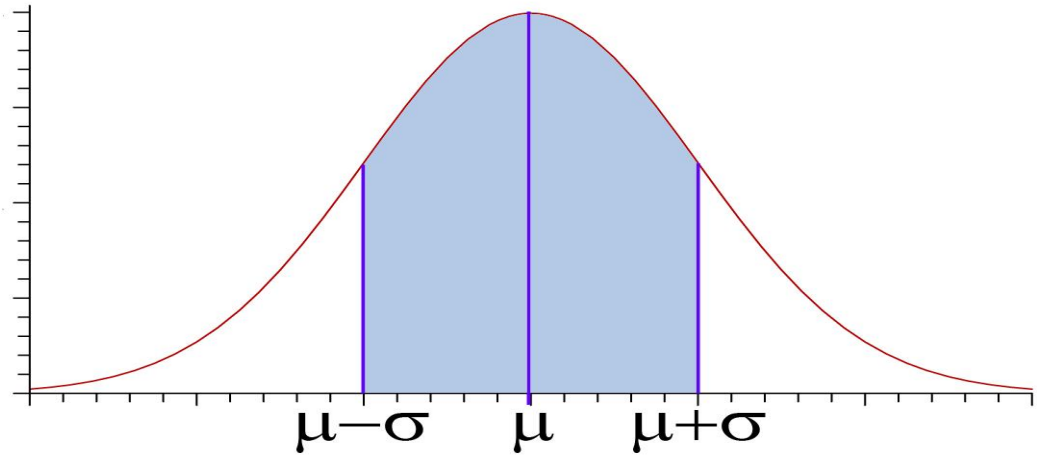
C. F. Gauss (1777—1855)

4.1. Eigenschappen van de normale verdeling

Hieronder staat een histogram van de gegevens uit de tabel met de lengtes van de militairen.



De vorm van deze grafiek lijkt op de grafiek die we eerder gebruikt hebben bij het begrip standaarddeviatie. Een dergelijke grafiek heet de grafiek van een normale verdeling



Wellicht ten overvloede opgemerkt: de grootheid, die op de horizontale as uitgezet wordt, wordt meestal x genoemd en de grootheid, die verticaal uitgezet wordt, wordt meestal y genoemd.

De normale verdeling heeft zeer vele eigenschappen, maar in het gestelde kader van deze reader zullen alleen de belangrijkste eigenschappen vermeld worden. Deze eigenschappen zijn ook terug te vinden in de bovenstaande grafiek.

(1) Het gemiddelde van de normale verdeling wordt meestal door het symbool μ aangegeven.

(2) De normale verdeling is volkomen symmetrisch t.o.v. de gemiddelde waarde, μ. Dit betekent dat de modus en de mediaan zich ook in het centrum bevinden. Voor de normale verdeling geldt derhalve dat:
Rekenkundig gemiddelde = mediaan = modus

(3) Het maximum van de normale verdeling bevindt zich bij de centrale waarde, bij het rekenkundig gemiddelde.

(4) Er zijn twee buigpunten te vinden in de grafiek , d.w.z. twee punten waar de grafiek van hol in bol overgaat. De afstand van het buigpunt tot de centrale waarde wordt de standaardafwijking σ genoemd.

(5) De twee staarten van de normale verdeling strekken zich uit tot in het oneindige en raken nooit de x-as. In de praktijk van alledag wordt slechts een zeer klein gedeelte van de normale verdeling gebruikt: grofweg het gebied dat loopt van:
(gemiddelde — 3 x standaardafwijking) tot: (gemiddelde + 3 x standaardafwijking).
Je hebt dan 99.7 % van de totale oppervlakte onder de normale verdeling te pakken. De overblijvende 0.3 % is verwaarloosbaar klein.

(6) De normale verdeling is volkomen bepaald door slechts twee parameters : het gemiddelde,U. en de standaardafwijking 0.

(7) Ongeveer 68 % van alle waarden in een normaal verdeelde verzameling bevindt zich op een afstand van maximaal 1 standaardafwijking vanaf de centrale waarde.

(8) Ongeveer 95.5 % van alle waarden in een normaal verdeelde verzameling bevindt zich op een afstand van maximaal 2 standaardafwijkingen vanaf de centrale waarde.

(9) Ongeveer 99.7 % van alle waarden in een normaal verdeelde verzameling bevindt zich op een afstand van maximaal 3 standaardafwijkingen vanaf de centrale waarde.

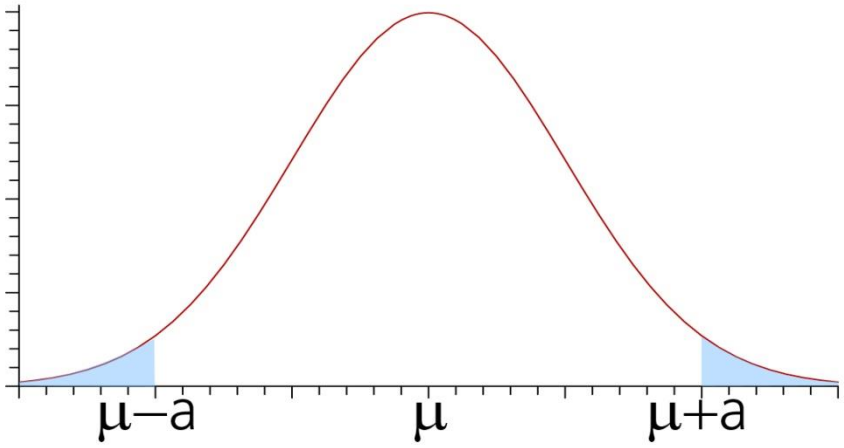
Er zijn duidelijk verschillen tussen de geschetste theoretische grafiek van een normale verdeling en de grafieken die je krijgt uit praktische meetwaarden. Zo lopen de beide staarten van de lengtegrafiek zeker niet tot in het oneindige door en is het ook lang niet zeker dat de grafiek wel helemaal symmetrisch is. De meetwaarden benaderen alleen de theoretische waarden van een normale verdeling.

4.2. De oppervlakte onder de grafiek van een normale verdeling.

De oppervlakte onder een grafiek van een normale verdeling zegt iets over **hoeveel** metingen in een bepaald gebied liggen.

Zo geeft het gearceerde deel van de grafiek hiervoor aan hoeveel meetwaarden en tussen μ-σ en μ+σ liggen. En omdat deze grafiek symmetrisch is, liggen er evenveel meetwaarden tussen μ-σ en μ als tussen μ en μ+σ.

Ook zijn er om dezelfde reden evenveel meetwaarden groter dan een zeker getal μ+a als dat er meetwaarden zijn kleiner dan μ-a.



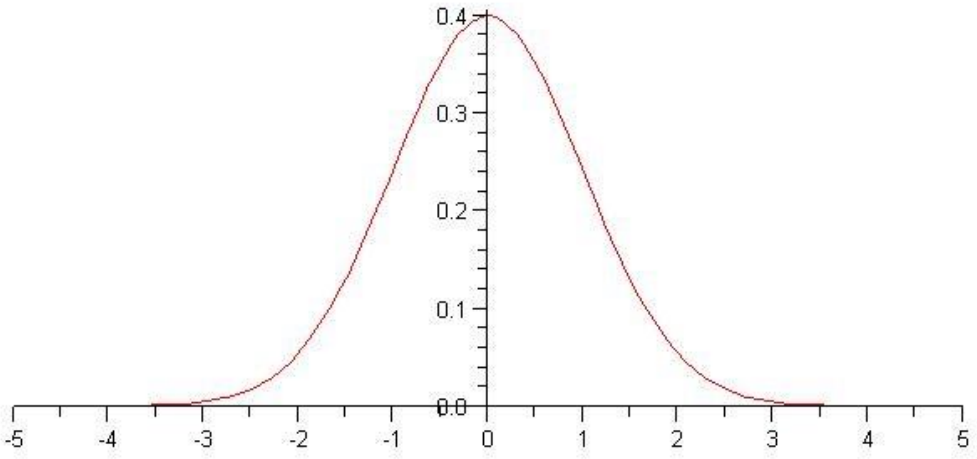
4.3. De standaardnormale verdeling

Eigenlijk is er niet een enkele normale verdeling, maar is er sprake van een oneindige verzameling van normale verdelingen. Dat kan je je gemakkelijk voorstellen als je in gedachten houd dat het gemiddelde zich overal op de horizontale x-as kan bevinden en als je er bovendien aan denkt dat de standaardafwijking ook iedere waarde groter dan 0 kan aannemen , dus de normale verdelingen kunnen variëren in hun plek op de x-as en kunnen variëren in hun breedte.

Een van de krommen uit deze oneindige verzameling van normale verdelingen heeft bijzondere eigenschappen en wordt de standaardnormale verdeling genoemd. Het bijzondere van deze standaardnormale verdeling is, dat het gemiddelde nul is en de standaardafwijking is gelijk aan 1 . Daardoor wordt de oppervlakte onder de normale verdeling gelijk aan 1. Dus

De standaardnormale verdeling is een normale verdeling met een gemiddelde,.μ = 0 en een standaardafwijking σ = 1 en waarvan de totale oppervlakte gelijk is aan 1.

Een grafiek van de standaardnormale verdeling is hieronder gegeven.



De rechter overschrijdingskans in de standaardnormale verdeling: P(X > x)

	0	1	2	3	4	5	6	7	8	9
0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002

4.4. Hoe gebruik je de standaardnormale verdeling?

Als we van een normale verdeling het gemiddelde en de standaardafwijking kennen, dan kunnen we die normale verdeling omrekenen naar de standaard normale verdeling. Dat heeft het grote voordeel dat we bij berekeningen met die verschillende verdelingen gebruik kunnen maken van dezelfde tabel met gegevens van de standaard normale verdeling. We kunnen met slechts één tabel volstaan. Die tabel staat elders in deze syllabus.

4.4.1. Het aflezen van de tabel

We kijken eerst alleen naar de eerste gegevenskolom: de kolom onder het getal 0.

Omdat het om de standaard normale verdeling gaat is het gemiddelde hier 0 en is de standaardafwijking 1. Het gemiddelde = 0 (en de mediaan en de modus ook), betekent dat de helft van de metingen een waarde onder 0 heeft en de helft een waarde boven 0. In de tabel (we kijken eerst alleen naar de eerste data kolom) zie je dat door door te kijken in het eerste vakje: daar staat 0,5000. Dit betekent dat de rechter overschrijdingskans van 0 gelijk is aan 0,5. Oftewel dat de kans dat de betreffende variabele groter dan 0, gelijk is aan 0,5..

Op dezelfde manier kun je aflezen dat de kans dat de betreffende variabele een waarde heeft groter dan 2,0 gelijk is aan 0,228. En dat de kans dat de variabele groter is dan 3,4 gelijk is aan 0,0003, Waardes groter dan 3,4 komen in de praktijk in deze situatie dus nauwelijks voor.

De betekenis van de volgende kolommen is dat het getal erboven een extra decimaal aangeeft die je achter het eerste opzoekgetal kunt plakken.
Laten we eens kijken bij het getal 1,3 in de eerste kolom en dan opschuiven naar kolom met 6 erboven. Daar staat 0.0869. Hiermee wordt bedoeld dat de kans dat de betreffende variabele groter is dan 1,36 gelijk is aan 0,0869. De 6 is een extra decimaal die van 1,3 nu het getal 1,36 maakt..

4.4.2. Een notatie

Omdat “de rechteroverschrijdingskans van de variabele X” een hele mond vol is, gebruiken we een verkorte schrijfwijze:

P(X>x) is de kans dat de waarde van een variabele X groter is dan een bepaalde waarde x

Een paar voorbeelden uit de tabel van de standaardnormale verdeling:

P(X>0,85)=0.1977
P(X>3,09)=0,0010

En

P(X>0,08)=0.4681

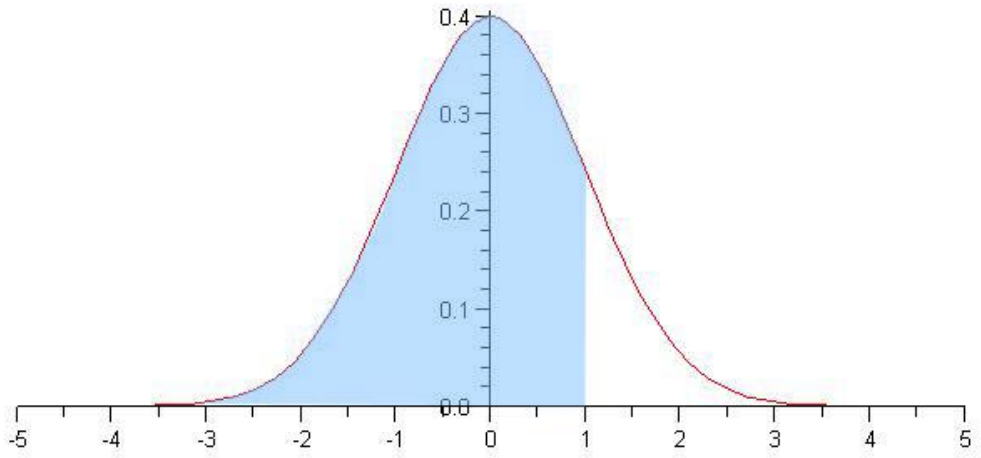
Je kunt de rechteroverschrijdingskansen van standaardnormaal verdeelde variabelen dus direct uit te tabel aflezen!

Maar er kan (veel) meer. We geven een paar voorbeelden:

Bereken: P(X<1)

Oplossing:

P(X<1) kun je in de grafiek als volgt weergeven:



Omdat de totale oppervlakte onder de grafiek gelijk is aan 1, geldt:

$$P(X<1)=1-P(X>1) = 1-0,587 = 0,413.$$

En het probleem is opgelost.

Een ander voorbeeld:

Stel we willen weten wat de kans is dat een variabele zich bevindt tussen 2 waarden, bijvoorbeeld tussen 1 en 2.

Oftewel bereken:

$$P(1<X<2)$$

Om dit te berekenen verdelen we de grafiek in drie delen: het deel kleiner dan 1, het deel tussen 1 en 2 en het deel groter dan 2. Die drie delen samen hebben oppervlakte gelijk aan 1. P(X>2) lezen we direct uit de tabel af (=0,228) en het deel P(X<1) hebben we zojuist berekend.

Dus:

$$P(1<X<2) =1-0,413-0,228 =0,359$$

4.5. Transformatie naar de standaardnormale verdeling

In de vorige paragraaf hebben we geoefend met het aflezen en berekenen van kansen bij de standaardnormale verdeling. Nu zijn de meeste normale verdelingen natuurlijk helemaal geen standaard normale verdeling. Het gemiddelde zal bijna nooit 0 zijn en de standaard afwijking bijna nooit 1.

Gelukkig kunnen we via een transformatie (omzetting) elke willekeurige normale verdeling omzetten naar de standaardnormale verdeling. We doen dit door:

(1) Trek van de gevraagde waarde het gegeven gemiddelde af

(2) Deel de uitkomst door de standaardafwijking

(3) De zo getransformeerde variabele is standaardnormaal verdeeld en kan in de tabel worden opgezocht.

Een voorbeeldje:

Stel we hebben een of andere variabele X die Normaal verdeeld is met gemiddelde = 125 en standaardafwijking = 20. We geven dat aan met

N(125,20)

Nu willen we weten wat de kans is dat de variabele groter is dan 150

Dus we vragen naar:

P(X>150)

waar X verdeeld is volgens

N(125,20)

Dan geldt:

$$P(X > 150) = P(Z > \frac{150-125}{20})$$

Waarbij de variabele Z nu standaard normaal verdeeld is. Dus:

$$P(X > 150) = P(Z > 0{,}25) = 0{,}0062$$

Er is één detail waar je rekening mee moet houden. In het bovenstaande voorbeeld is het getal 0,25 positief. Maar dat zou in een andere situatie ook negatief hebben kunnen zijn. In dat geval moet je van het negatieve getal een positief getal maken. Vandaar dat je in boeken hierover de volgende formule ziet:

$$\left|\frac{150-125}{20}\right|$$

Het getal staat tussen absolute waarden strepen. Hiermee wordt niets anders bedoeld dan dat het getal als het negatief is moet worden “omgeklapt” naar het positieve getal.

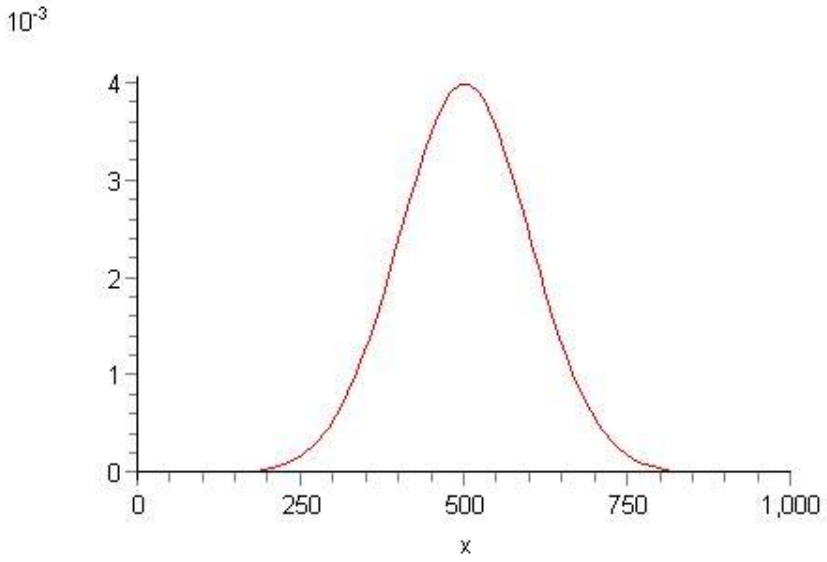
Een grootheid X, die normaal verdeeld is, kan gestandaardiseerd worden door X te transformeren in een andere grootheid Z volgens de onderstaande transformatieformule:

$$Z = \left| \frac{X - \mu}{\sigma} \right|$$

Deze formule met zijn absoluutstrepen (= de twee verticale strepen) zal nu uitvoerig verklaard en toegelicht worden. Dit wordt gedaan aan de hand van het onderstaande praktijkvoorbeeld:

Het bovenstaande praktijkvoorbeeld is samengevat in de onderstaande grafiek:

Er is een trainingsprogramma ontworpen om de bekwaamheid van toezichthouders bij lopende band-werkzaamheden te doen toenemen. Omdat het programma pure zelfwerkzaamheid vereist, zullen verschillende toezichthouders een onderling verschillend aantal uren nodig hebben om het programma te volgen en met succes af te sluiten. Bestudering van de studieresultaten van vele toezichthouders gaf aan dat de gemiddelde tijd , die zij nodig hadden , 500 uren bedroeg en dat deze benodigde tijd normaal verdeeld was met een standaardafwilking van 100 uren.



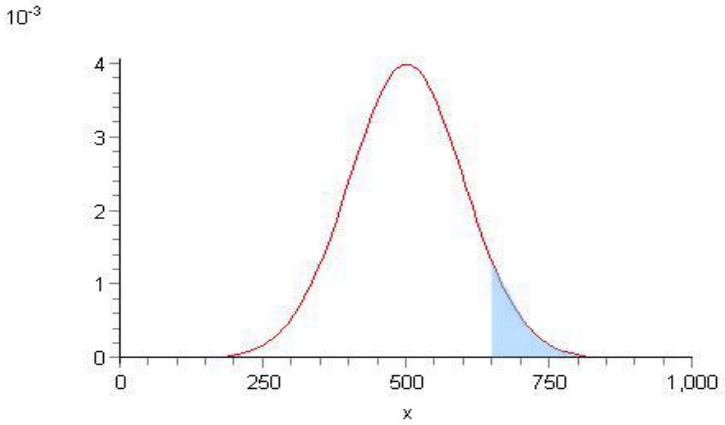
Er worden nu vijf voorbeelden gegeven om te laten zien hoe z berekend moet worden in 5 verschillende situaties. Verder wordt uitgelegd hoe z, bij het gebruiken van de standaard normale verdeling, leidt tot oppervlaktewaarden.

Voorbeeld 1:

Wat is de kans dat een toezichthouder er meer dan 650 uren over zal doen om zich de cursus met succes eigen te maken?

Oplossing

Deze kans is gelijk aan het gearceerde gebied in de onderstaande grafiek:



De berekening van de oppervlakte van dit gearceerde gebied gaat als volgt:

$$z = \left|\frac{x-\mu}{\sigma}\right| = \left|\frac{650-500}{100}\right| = |+1{,}5| = 1{,}5$$

We zoeken z=1.50 op in de tabel voor de standaardnormale verdeling. Je moet dan kijken op het kruispunt van de rij achter 1.5 en de kolom onder

Je vindt het getal 0668. De kansen in de tabel zijn echter met 10000 vermenigvuldigd. De werkelijke kans is dus: 668/10 000 = 0.068 ofwel , in procenten, 668%.

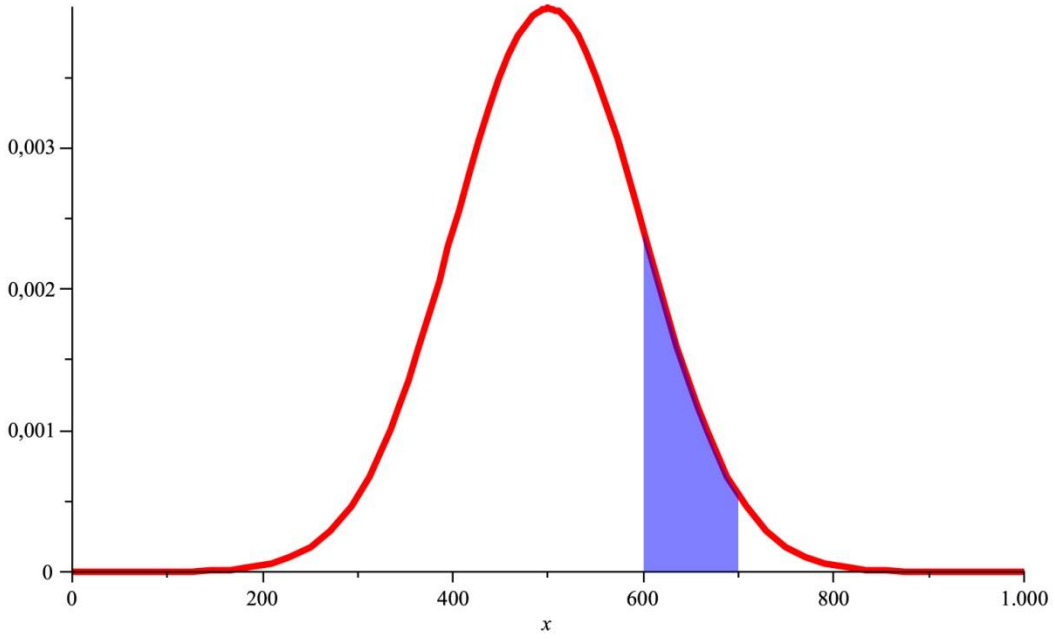
Conclusie : de kans dat een toezichthouder er meer dan 650 uren over zal doen om zich de cursus met succes eigen te maken , is 6.68 %.

Voorbeeld 2:

Wat is de kans dat een toezichthouder er tussen de 600 en 700 uren over zal doen om zich de cursus met succes eigen te maken?

Oplossing:

Deze kans is gelijk aan het gearceerde gebied in de onderstaande grafiek:



We weten dat de variabele $Z = \left| \frac{X - \mu}{\sigma} \right|$ N(0,1) verdeeld is.

Hier moeten we weten wat:

$$P(600 < X < 700)$$

Is. We berekenen die kans door:

$$P(600 < X < 700) = P(X > 600) - P(X > 700) = P(Z > \frac{600-500}{100}) - P(Z > \frac{700-500}{100})$$

Uit de tabel lezen we de bijbehorende waarden af:

$$P(Z > 2) = 0,0228 \text{ en } P(Z > 1) = 0,1587$$

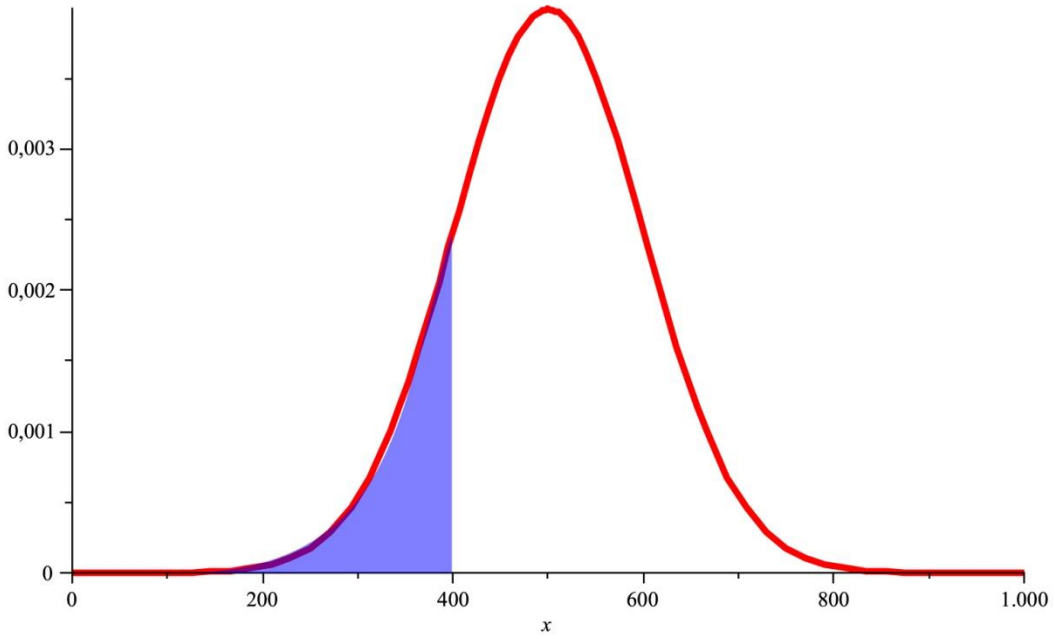
Dus de gevraagde kans is: 0,1587 – 0,0228 = 0,1359

Voorbeeld 3:

Wat is de kans dat een toezichthouder er minder dan 400 uren over zal .doen om zich de cursus met succes eigen te maken (een snelle cursist dus)?

Oplossing:

Deze kans is gelijk aan het gearceerde gebied in de onderstaande grafiek:



De berekening van de oppervlakte van dit gearceerde gebied gaat als volgt:

De oppervlakte links van 400 is, omdat de grafiek symmetrisch tov x=500, precies gelijk aan de oppervlakte rechts van 600. Dat hadden we in het vorige voorbeeld al berekend.

De tabel geeft 1587 , dus 15.87 % ligt links van x = 400 uren.

(b) Conclusie : de kans dat een toezichthouder er minder dan 400 uren over zal doen om zich de cursus met succes eigen te maken bedraagt 15.87 %.

Voorbeeld 4:

Wat is de kans dat een toezichthouder er tussen de 300 en 400 uren over zal doen om zich de cursus met succes eigen te maken?

Oplossing:

Het wordt saai. Beredeneer zelf dat deze kans gelijk is aan de kans van voorbeeld 2.

Conclusie : de kans dat een toezichthouder er tussen de 300 en 400 uren over zal doen om zich de cursus met succes eigen te maken bedraagt 15.87 % — 2.28 % = 13.59 %.

Laatste voorbeeld (ook saai ;-) :

Wat is de kans dat een toezichthouder er tussen de 400 en 700 uren over zal doen om zich de cursus met succes eigen te maken?

Oplossing:

We hadden berekend dat rechts van 700 een gebied van 2.28 % lag en ook hadden we berekend dat links van 400 zich een gebied van 15.87 % bevond. Daar de totale oppervlakte onder de standaardnormale verdeling gelijk is aan 100 %, ligt tussen 400 en 700 een gebied van:

100% - 2.28% - 15.87% = 81.85%

Conclusie : de kans dat een toezichthouder er tussen de 400 en 700 uren over zal doen om zich de cursus met succes eigen te maken bedraagt 81.85 %.

De voorgaande vijf voorbeelden dienen ervoor het gebruik van de tabel voor de standaardnormale verdeling volkomen te verklaren , onafhankelijk van het feit of je nu opereert in het gebied links van de centrale waarde van de normale verdeling of het gebied rechts van de centrale waarde van de normale verdeling.

De normale verdeling kan overigens in tal van situaties toegepast worden. Maar er moet in het kader van deze reader een beperking aangebracht worden en daarom zullen nu enige situaties beschreven worden waarin het gebruik van de normale verdeling zonneklaar is of waarin de normale verdeling als benadering gebruikt zal worden.

5. Index