Het cumulatief gemiddelde of wat hebben ze nu

Wat is een cumul

Wie zich wel eens met statistisch onderzoek heeft beziggehouden zal gemerkt hebben dat er wel erg veel meetgegevens nodig zijn om tot betrouwbare uitspraken te komen. Velen van u zullen, net als ik, wel eens tussendoor wat hebben zitten rommelen met de getallen. Je merkt dan al snel dat de conclusies veel sneller getrokken kunnen worden dan de statistici willen toestaan. De indruk ontstaat dat de statistiek erg veel redundantie bevat, waardoor die grote aantallen metingen nodig zijn. In dit stuk willen we nagaan waar mogelijkheden zouden kunnen liggen om het aantal metingen terug te brengen.

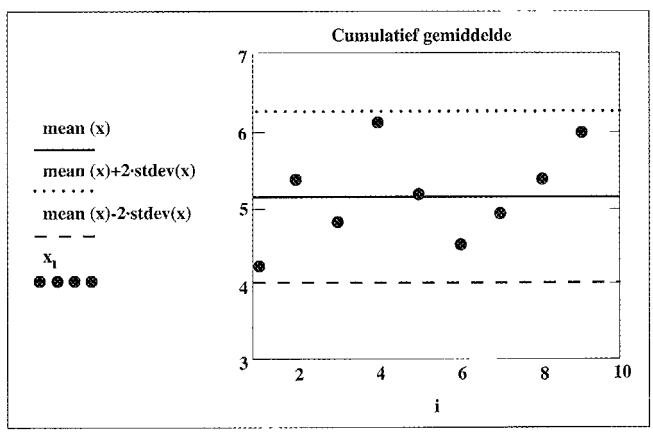
We starten met te laten zien wat een cumulatief gemiddelde (CG) is. In formulevorm ziet het er als volgt uit:

$$CG = \frac{1}{i}\sum_{1}^{i} x_i$$

Dus we tellen de eerste i getallen op en delen door i, dat is alles. We zullen aan de hand van het voorbeeld op de volgende pagina laten zien hoe dat uitpakt.

We berekenen van kolom 2 het gemiddelde en de spreiding en vinden.
gemiddelde = 5.117 en de
standaarddeviatie = 0.563

We zetten dit in een grafiek samen met de gevonden meetpunten (zie grafiek 1).
Wat opvalt is dat de 2-s grenzen buiten de meetpunten lopen, en dat is de eerste redundantie die wordt toegelaten.



Grafiek 1

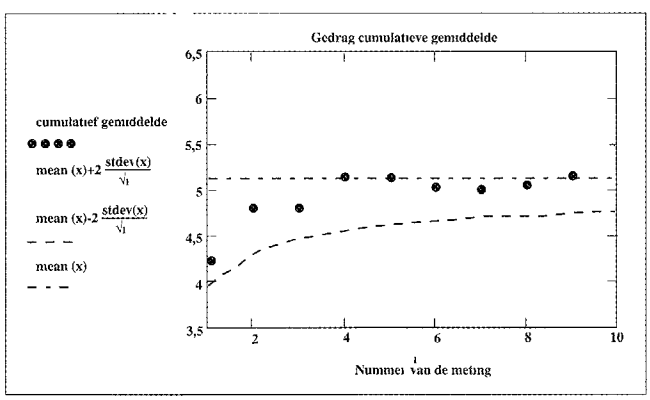
We gaan nu kijken naar het cumulatieve gemiddelde. Hierbij moeten we er rekening mee houden dat de steekproefdeviatie is afgenomen met $\frac{1}{\sqrt{i}}$ bij de i-de meting. Dat is ook logisch, na meer metingen is het gemiddelde beter in te perken. We zetten dat weer in een grafiek (grafiek 2)

weer verzonnen

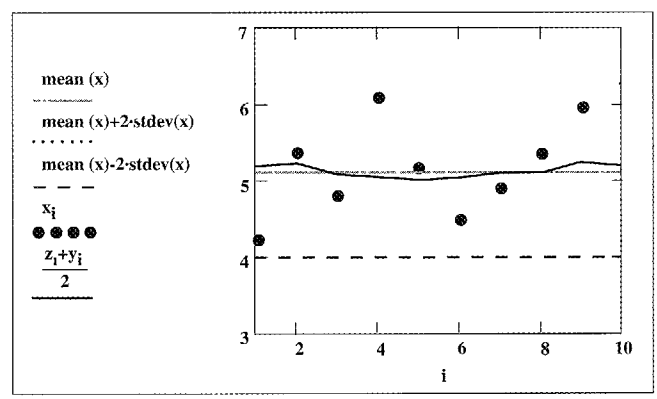
atief gemiddelde

Voorbeeld

nummer	meting M	CG	Meting oplopend A	Meting aflopend B	(A + B)/2
1	4.25	4.25	4.25	6.10	5.175
2	5.37	4.81	4.48	5.95	5.215
3	4.80	4.81	4.80	5.37	5.085
4	6.10	5.13	4.81	5.35	5.08
5	5.15	5.13	4.91	5.15	5.03
6	4.48	5.02	5.15	4.91	5.03
7	4.91	5.01	5.35	4.81	5.08
8	5.35	5.05	5.37	4.80	5.085
9	5.95	5.15	5.95	4.48	5.215
10	4.81	5.12	6.10	4.25	5.175



Grafiek 2



Grafiek 3

Dat is natuurlijk mooi, we zien de punten heel snel naar het gemiddelde gaan en het valt op dat ze ver binnen de 2-s grenzen blijven. Door dit in een groot aantal praktijkgevallen toe te passen blijkt dat dat altijd zo is en pak je de 3-s grenzen dan wordt dat natuurlijk nog veel opvallender. Keren we terug naar het voorbeeld dan ziet u dat de metingen in de kolommen 4 en 5 in gesorteerde volgorde staan (oplopend en aflopend). In kolom 6 staat het gemiddelde van de twee voorgaande kolommen en we zien dat deze cijfers heel dicht bij het gemiddelde liggen. Hoe dicht laat grafiek 3 zien.

We zien de lijn die zich mooi om het gemiddelde kronkelt. De vraag die nu opkomt is of dit altijd opgaat en het antwoord is ja, als de cijfers tenminste uit een meting komen Natuurlijk ben ook ik in staat om een set cijfers te construeren die zich op dit punt misdraagt. Maar dit soort setjes blijkt in de praktijk niet voor te komen bij reële verschijnselen.

In de praktijk is een groot aantal metingen op een dergelijke manier bekeken (150) en in geen enkel geval ging het mis. Uit het gedrag van het cumulatieve gemiddelde zien we dat na 5 à 6 metingen er een soort gemiddelde wordt bereikt

waarmee men prima kan werken, we komen daar nog op terug.

De luieriken onder ons kunnen nu het verder lezen staken want in de rest wordt aannemelijk gemaakt waarom dat zo is. Zij hoeven alleen maar te onthouden dat ook bij een klein aantal metingen het gemiddelde bruikbaar is en waarschijnlijk werkt u zo ook al, dus er is niets nieuws onder de zon.

Ik hoop echter dat er een voldoende aantal lezers verder leest, want er komen leuke dingen aan de orde die u in de toekomst van dienst zouden kunnen zijn.

Kritisch geluid

1 In de statistiek praat men over twee soorten gegevens namelijk het populatiegemiddelde en -spreiding (μ, σ) en het steekproefgemiddelde en -spreiding ($\bar{x}$, s). Men gaat er dan van uit dat er een hele grote populatie van dingen is die een eigen gemiddelde en spreiding hebben. Uit een steekproef uit deze populatie probeert men dan een schatting te vinden van μ en σ. Statistisch gesproken is dit een prima procedure en met zekerheid bestaan er ook van deze populaties (denk b.v. aan de bevolking van Nederland). Maar hoe zit dat met nieuwe ontwikkelde dingen? Dan is het niet zo waarschijnlijk dat er een dergelijke populatie bestaat.

Nu moeten we met dit laatste voorzichtig zijn. De grote vraag is of nieuwe dingen ontdekt worden of dat ze worden uitgevonden. Er is een behoorlijke groep mathematici die van het eerste uitgaan (b.v. Erdös, toch niet de minste). Als dat waar is dan bestaan de objecten eigenlijk al, wij hebben ze slechts ontdekt. Als voorbeeld kunt u denken aan de Mandelbrot-verzameling, bestond die al voor de definitie of niet? Helaas kunnen we niet verder ingaan op deze boeiende discussie.

De statistiek haalt vaak, als voorbeeld, de dingen aan die in massa vervaardigd worden om over populaties te praten. Nu heb ik in het verleden op grote schaal kunnen testen wat er in de massafabricage optreedt en steeds is gebleken dat het populatieconcept niet vruchtbaar is. Als er überhaupt al gereageerd wordt op de gegevens uit het statistisch onderzoek dan blijkt steeds dat men het beste uit kan gaan van de steekproefgrootheden, b.v. om machines bij te stellen enz

2 De statistiek maakt slecht gebruik van de informatie die reeds over een te meten grootheid bestaat. Twee voorbeelden:

De leeftijdsopbouw van medewerkers in een organisatie. De leeftijden liggen dan niet tussen $-\infty$ en $+\infty$ maar tussen 18 jaar en 65 jaar. Een anekdote:

In een organisatie was een onderzoek geweest naar o.a. de opbouw van de leeftijden. De statisticus rapporteerde dat de gemiddelde leeftijd 40 jaar was en de spreiding s bedroeg 15 jaar. Voor alle duidelijkheid vermeldde hij er nog bij dat 95% van de waarnemingen nu lagen tussen de $\pm$ 2s grenzen. Rekenen we dit uit dan zien we dat de leeftijden dan liggen tussen de 10 en de 70 jaar. Dus op papier was er sprake van kinderarbeid en ouderenuitbuiting.

U zult opmerken dat dit een uitzondering is, maar daar vergist u zich in. Het is eerder regelmaat let u er maar eens op.

Het tweede voorbeeld heeft te maken met het meten van procenten Uiteraard is het duidelijk dat de gemeten grootheden dan tussen de 0 en 100% moeten liggen. Maar vaak is het mogelijk om dit veel meer in te perken, b.v. men weet dat het percentage hoger moet zijn dan 95%. En ook nu mogen de 2-s grenzen de 100% niet overschrijden. Van deze kennis maakt de statistiek over het algemeen geen gebruik. En als de 2-s grens nu de 100% overschrijdt dan kan men vaststellen dat die 2-s grens helemaal niets zegt over de uitgebreidheid van de grootheid.

Overigens is het ook nog zo dat men vaak van een populatie het gemiddelde en spreiding opgeeft, maar men verzuimt om erbij te zetten welke verdeling de grootheid heeft. Men blijft er kennelijk van uitgaan dat alle verdelingen normaal zijn

Tot slot van dit hoofdstuk nog een onverkwikkelijke episode uit februari van het vorige jaar In de pers wordt melding gemaakt van de medewerker van het RIVM dr. ir. de Kwaadsteniet die zijn schorsing voor rechtbank in Utrecht aanvecht. In de pers zag ik hoe de Kwaadsteniet uitlegde dat men bij metingen vaak 'vergat' om de betrouwbaarheid op te geven, en tevens begreep ik dat een hoop gegevens niet eens gemeten werden, maar het resultaten waren van simulatieprogramma's. Nu is dat grosso modo ook mijn ervaring met statistisch onderzoek, dus wachtte ik met meer dan gewone belangstelling de reactie af van de leiding van het RIVM. Nou, die was om van te huilen. In een onfrisse scheldkanonnade, zonder een inhoudelijk argument, werd de arme medewerker de grond in geboord. Wat was het toch eenvoudig geweest om aan de hand van controleerbare gegevens te laten zien dat de medewerker fout was geweest. Dat deed men echter niet (waarom niet?) zodat er nu maar

een conclusie mogelijk is. Het is niet duidelijk of het RIVM betrouwbare gegevens produceert, maar wel is duidelijk dat het RIVM oncontroleerbaar is. En dat jammer voor al die goedwillende medewerkers van het instituut. En we zien dat je gedonder krijgt als je wel de betrouwbaarheid vermeldt, maar het geeft ook gedonder als je het niet doet.

Een nuttige transformatie

Waarschijnlijk is het duidelijk geworden dat er in de praktijk van een stochastische grootheid de statistische verdeling zelden bekend is. Dit is zeker zo als men metingen doet in de onderzoekssfeer, maar meestal is dat ook zo bij de massafabricage. We weten dan helemaal niets van de statistische grootheid en de gemeten getalletjes kunnen qua grootte overal liggen. We laten nu een transformatie uit de lucht vallen die vaak nut oplevert.

Stel $\underline{x}$ is een stochastische variabele met kansdichtheid $f(\underline{x})$, dan is de bedoelde transformatie:

$$y = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

We zien dat als $x_i = x_{min}$ dan wordt $y = 0$ en als $x_i = x_{max}$ dan wordt $y = 1$, dus ligt y tussen 0 en 1.

Wil men hiermee statistiek bedrijven dan moeten we in de gaten houden dat y geen kansverdeling is, dat is wel $2 \times y$. Bij integratie tussen 0 en 1 levert 2y het getal 1 op zoals het hoort. We zullen er niet verder op ingaan.

Passen we deze transformatie toe op een uniforme (rechthoekige, homogene) verdeling dan wordt deze getransformeerd in de standaard rechthoekige verdeling waarbij $\underline{y}$ tussen 0 en 1 ligt en de kansdichtheid $f(\underline{y}) = 1$ is.

Nu kan men in principe elke willekeurige verdeling transformeren in de standaard rechthoekige verdeling met de transformatie

$\underline{y} = F(\underline{x})$, waarbij $F(\underline{x})$ de cumulatieve verdeling is van $f(\underline{x})$

Aan deze wijsheid heb je natuurlijk geen pest als je de verdeling niet kent. Ook in het algemene geval van een transformatie kan men uitspraken doen over de verdeling van $f(\underline{y})$ als men de verdeling van $f(\underline{x})$ kent, maar helaas zoals gezegd, kennen we die niet. We zullen dus een list moeten verzinnen, we komen er nog op terug.

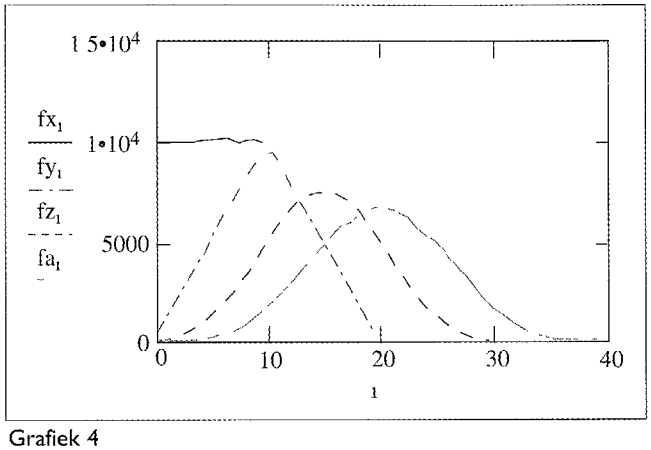
Kijken we nog eens goed naar de voorgestelde transformatie, en slaan we er een statistiekboekje op na, dan zien we dat dit precies de cumulatieve verdeling is van de rechthoekige verdeling.

Optellen van statistische grootheden

Een behoorlijk deel van de statistiek gaat over het optellen van statistische grootheden Het kost echter heel wat moeite om uit te vinden wat men daar dan wel onder verstaat. Het blijkt dat men het volgende bedoelt.

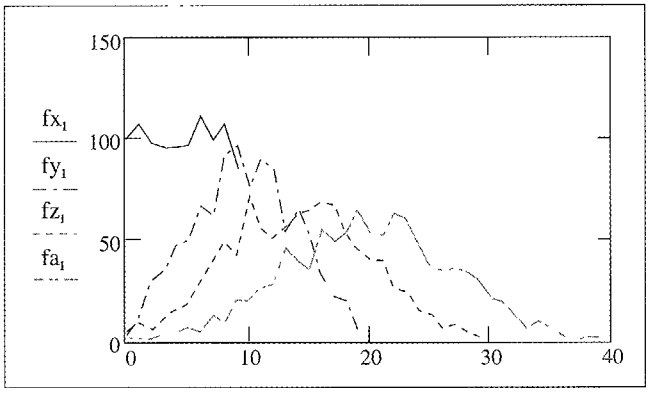
Stel we hebben een stochastiek $x_1, x_2, x_3, ...x_n$ en een tweede stochastiek $y_1, y_2, y_3, ...y_n$ dan bedoelt men met het optellen dat men kijkt naar: $x_1 + y_1, x_2 + y_2, x_3 + y_3$ ... enz. Mijns inziens zou het woord samenvoegen beter op zijn plaats zijn geweest, maar we zullen niet zeuren.

Stel nu verder dat we een aantal stochastieken hebben die komen uit een uniforme verdeling waarbij de getallen liggen tussen 0 en 10. In onderstaande figuur kunt u zien wat optellen dan oplevert.



Grafiek 4

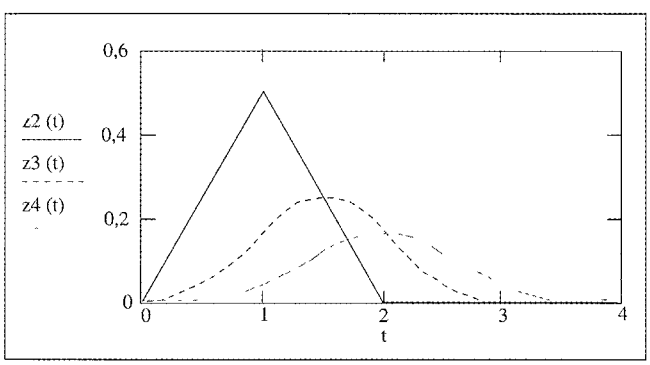
In grafiek 4 zien we links de uitgangsverdeling, dus de rechthoekige verdeling, dan de eerste rechts ervan, is een driehoekige verdeling en dat is de som van twee rechthoekige. Daar weer rechts van zien we twee mooie klokvormen en dat zijn respectievelijk de som van drie en de som van vier rechthoekige verdelingen. Om deze mooie vormen te krijgen heb ik 100.000 trekkingen moeten doen (uiteraard met de computer). Nemen we er honderd of duizend van



Grafiek 5

dan zijn de vormen nauwelijks te onderkennen en hier zien we weer hoe traag de statistiek eigenlijk reageert, zie grafiek 5 (1000 trekkingen).

Wat we nu met een simulatie hebben gedaan kunnen we ook theoretisch natrekken. We kunnen in de statistiekboekjes vinden dat voor de som van twee stochastieken (f en g) geldt dat de verdeling de bekende convolutie f*g is. Ook kan worden bewezen dat voor Laplace-getransformeerden geldt.
L(f*g) = L(f).L(g).
Meestal laat men het bij het vermelden van deze heugelijke feiten en als je probeert om dat na te rekenen snap je ook waarom. Je rekent je namelijk te blubber. Echter onze slaaf de computer kan ons nu weer van nut zijn. De berekeningen zijn uitgevoerd met Mathcad, we zullen er niet op ingaan en alleen het resultaat als grafiek 6 weergeven



Grafiek 6

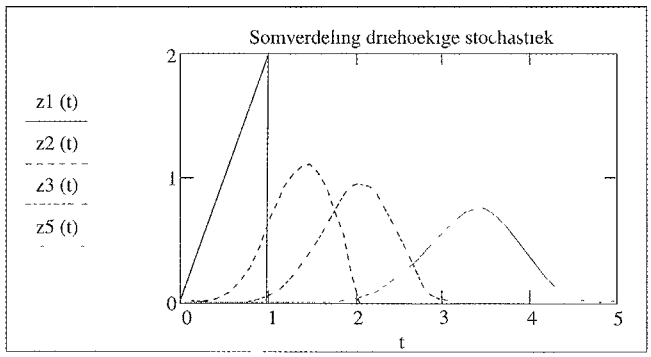
We zien gelukkig dezelfde figuren terug als bij de simulatie, de rechthoek heb ik maar weggelaten.

Ter illustratie geven we de kromme voor z3(t) in formulevorm weer In deze formule is b v. Φ(t–2) de Heavisidestapfunctie (voor t<2 is de functie 0 en voor t>2 is hij 1 en blijft hij 1)

$$z3(t) = \frac{1}{2} \cdot t2 - \frac{3}{2} \cdot \Phi(t-1) \cdot (t-1)^2 + \frac{3}{2} \cdot \Phi(t-2) \cdot (t-2)^2 - \frac{1}{2} \cdot \Phi(t-3) \cdot (t-3)^2$$

Zonder commentaar laten we nog zien hoe de som van stochastieken eruitziet als ze afkomstig zijn van een driehoekige asymmetrische verdeling (grafiek 7).

Wat hebben we nu eigenlijk uit dit hoofdstuk geleerd? Ja, eigenlijk niet zo heel veel. Het is echter zo dat de meeste technici wiskundige formules best leuk vinden, maar nog leuker vinden ze het als ze in grafiekvorm kunnen zien hoe het werkt en dat was de achtergrond van dit hoofdstuk.
Technici willen nu eenmaal overal het vingertje achter krijgen Je kan denken met het hoofd maar het gaat ook heel goed met de handen en technici denken met beide, vandaar.



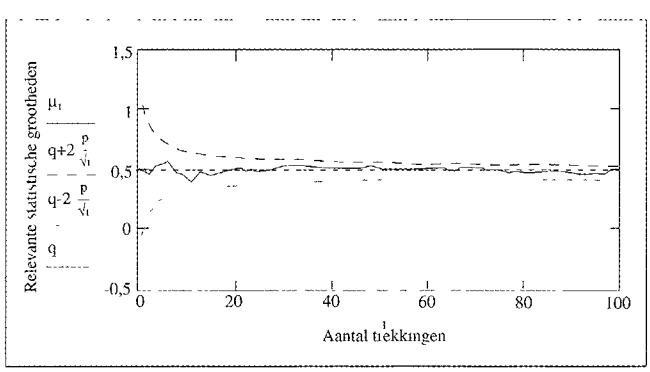
Grafiek 7

Over het gedrag van het cumulatieve gemiddelde

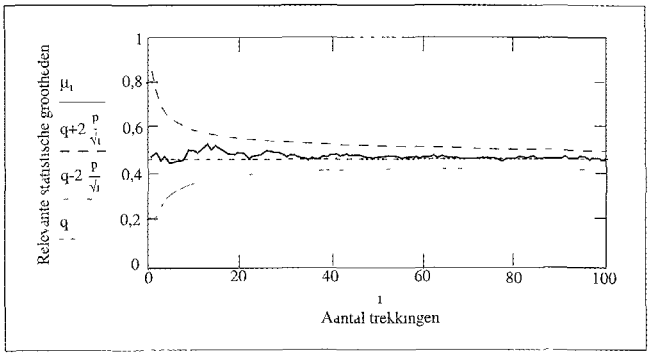
In het eerste hoofdstuk hebben we aan de hand van een klein voorbeeld laten zien wat een cumulatief gemiddelde is en hoe het zich gedraagt Met de hiervoor ontwikkelde inzichten zouden we nu theoretisch kunnen nagaan hoe het een en ander werkt. Dit leidt echter tot niets. Al spoedig worden de formules onoverzichtelijk groot en ook de computer houdt ermee op.

> Bovendien valt niet te verwachten dat we iets anders zullen vinden dan datgene wat de klassieke statistiek ons biedt. Statistiek is een goed gefundeerde wetenschap, laat daar geen misverstand over ontstaan.

Wat we echter wel kunnen doen is het gedrag praktisch simuleren. In dat geval doen we een groot aantal aselecte trekkingen uit een gekozen statistische verdeling en laten de computer de relevante grootheden uitrekenen en in een grafiek zetten. Dat is wel een rotklus, maar het leert ons erg veel. Het blijkt dat het gedrag van de output erg wordt beïnvloed door de eerste trekking die wordt gedaan. Als die binnen de trekking groot of klein is duurt het wat lang om naar het gemiddelde te kruipen. Hier kunnen we iets aan doen door als eerste waarde altijd het steekproefgemiddelde te nemen.
Hierdoor wordt het gemiddelde niet beïnvloed, en naar de spreiding kijken we even niet.
Alle bekende verdelingen zijn op een dergelijke manier bewerkt en beoordeeld. Voor de goede orde laat ik een tweetal grafieken zien.

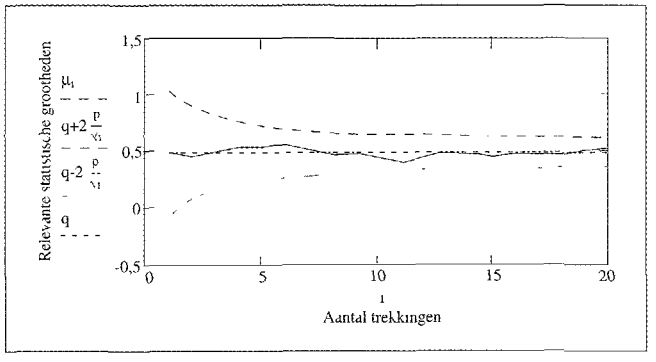


Trekkingen uit een uniforme verdeling

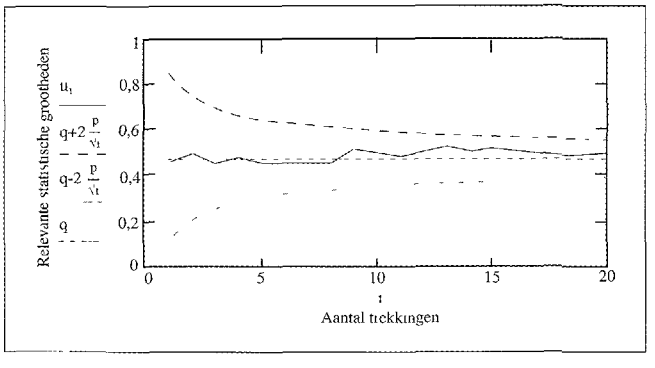


Trekkingen uit een normale verdeling

We zien de μ steeds op het gemiddelde q starten (hebben we zelf zo ingesteld) en zich mooi binnen de 2s grenzen bewegen. Heel soms loopt het gemiddelde wel eens tegen de 2s grenzen aan, maar echt overschrijden doet hij niet
Nu laten we hier het aantal trekkingen doorlopen tot 100, maar eigenlijk zijn we geïnteresseerd in het gedrag voor een klein aantal trekkingen. Daarom nogmaals twee grafieken maar nu voor maximaal 20 trekkingen



Trekkingen uit een uniforme verdeling



Trekkingen uit een normale verdeling

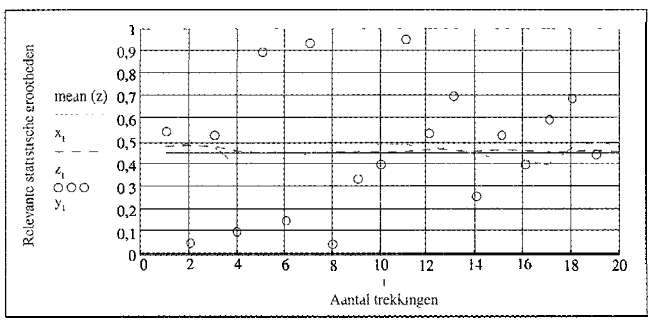
We hebben in de grafieken een paar fraaie uitkomsten laten zien. Gaat men langer met de getallen experimenteren dan ziet men de gemiddeldelijn (μ) wel naar de 2s grenzen lopen, maar zoals reeds voren aangegeven, het herstelt zich steeds weer. Wat we ons goed moeten realiseren is dat we in werkelijkheid het gemiddelde q en de spreiding p niet kennen. We hebben dan alleen de μ-lijn, en wat we beoordelen is het gedrag van deze lijn als functie van het aantal trekkingen i dat we gedaan hebben. Als dat tendeert naar stabiliseren, dan gaan we met het steekproefgemiddelde verder als goede schatting.

Bij het experimenteren met de verschillende verdelingen komt een ding prominent naar voren. Als we kijken naar de histogrammen van een klein aantal trekkingen (<40) dan valt op dat die in de verste verte niet lijken op de verdeling waar ze uit zijn ontstaan. Door ze te vergelijken met random getallen tussen 0 en 1 (natuurlijk eerst op dezelfde schaal gebracht) dan blijkt dat er geen principieel onderscheid zichtbaar is. Echter random getallen tussen 0 en 1 genereren is precies hetzelfde als aselect trekkingen doen uit een uniforme verdeling tussen 0 en 1.
Met de voorgestelde transformatie kunnen we nu alle verdelingen tussen 0 en 1 leggen, en als de aantallen klein zijn kunnen we net doen alsof deze getallen afkomstig zijn uit een uniforme verdeling. Waarom dat zo is zal wel samenhangen met de opmerking gemaakt in hoofdstuk 2. Ontdekken we de verdeling (die dus eigenlijk al lang bestaat), of hebben we hem uitgevonden. Aangezien we in de techniek niet geloven in bovennatuurlijke verschijnselen gaan we van het laatste uit. En in dat geval kunnen de eerste getallen die gegenereerd worden niet 'weten' waar ze vandaan komen, er is nog geen verdeling. Als er meer metingen komen dan kan zich een verdeling gaan ontwikkelen die we, via het maken van histogrammen, kunnen opsporen.
Tot slot nog het volgende: Men gaat er in de statistiek heel vaak van uit dat men te maken heeft met normale verdelingen. Door het bovenstaande consequent toe te passen blijkt dat dat niet zó vaak het geval is. Uniforme verdelingen, weibull-verdelingen en beta-verdelingen komen heel vaak voor. Door dit te negeren worden er, b.v bij steekproefkeuringen, economische verliezen geleden en dat is jammer.

Het sorteren van getallen

In hoofdstuk 1 hebben we kennisgemaakt met het sorteren van de meetgegevens. Nu was het in de oudheid al bekend dat met het sorteren van getallen soms tot mooie resultaten te komen is. Zo wist b.v. Archimedes via deze methode sommen van reeksen te berekenen. Helaas past men het in de praktijk niet vaak toe.
Aangezien we in het voorgaande gezien hebben dat de uniforme verdeling best bruikbaar is zullen we eerst met een

simulatie nagaan wat het sorteren (oplopend en aflopend) te bieden heeft.
In de volgende grafiek zijn 20 aselecte trekkingen uit een standaard uniforme verdeling (0,1) uitgezet. Dit zijn de o's. De lijn z_i geeft het gemiddelde van de twee gesorteerde reeksen, en om het feest complect te maken hebben we van z_i nog het cumulatieve gemiddelde x_i berekend.
Mean(z) is het gemiddelde van de oorspronkelijke trekkingen en we zien dat dit steekproefgemiddelde niet samenvalt met het populatiegemiddelde (0.5)



Gedrag van gesorteerde reeksen

We zien dat zowel de lijn x als z heel dicht om het gemiddelde 'mean(z)' kronkelen zoals ook te verwachten was.

In de mathematische statistiek komen we hoofdstukken tegen die zich bezighouden met de statistiek van gesorteerde getallen. Men noemt dat de statistiek van de geordende waarnemingen (order statistics). Dat is prachtig materiaal, maar het is niet zo eenvoudig toegankelijk en dat zal dan ook wel de reden zijn dat we het niet in de toegepaste statistiek, laat staan, in de techniek tegenkomen. Persoonlijk heb ik het nooit zien toepassen en dat zal ook niet lukken als de resultaten niet verder uitgewerkt worden. We zullen in het volgende alleen kijken naar de geordende statistiek van de standaard uniforme verdeling.
Stel $\underline{x}_1.....\underline{x}_n$ zijn de aselecte getallen getrokken uit een continue verdeling f(x) en f(x) is de standaard uniforme verdeling (0,1). Stel verder dat $\underline{y}_1.....\underline{y}_n$ dezelfde getalletjes zijn maar nu oplopend geordend.
Voor de kansdichtheid vinden we dan:

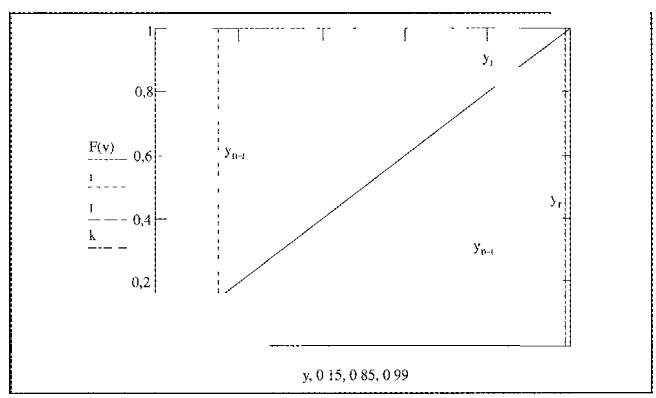
$$\mathbf{f(y_i)} = \frac{\mathbf{n!}}{\mathbf{(i-1)!(n-i)!}} \bullet \mathbf{y_1^{i-1}} \bullet \mathbf{(1-y_1)^{n-i}} \quad \mathbf{(1)}$$

Nu zeg je tegen dit soort verdelingen alleen maar ó, nooit ha, wat moet je ermee. Voor de liefhebbers deel ik zonder uitleg mee dat dit een beta-verdeling is met de parameters (n–i) en (i–1).

Door naar de simultane verdeling te kijken van y_1 en y_n en wat mathematisch gegoochel kunnen we de verdeling voor de range **r** vinden $\mathbf{r=y_n-y_1}$.
Voor de standaard uniforme verdeling (0,1) vinden we dan:

$$\mathbf{f(r)=n\bullet(n-1)\bullet r^{n-2}(1-r)} \quad \mathbf{(2)}$$

We komen hier zo op terug, we gaan eerst nog eens wat nader kijken naar formule (1). Het optellen van de twee gesorteerde verdelingen kunnen we vervangen door het optellen van (y_i+y_{n-i+1}). Voor i=1 staat er y_1+y_n, voor i=2 y_2+y_{n-1} enz En zoals we al in het numerieke voorbeeld zagen zijn dit schattingen voor '2 maal het gemiddelde' dus de range. We kijken nog eens wat beter naar het optellen van de grootheden via de cumulatieve verdeling $F(y_i)$. Voor een uniforme standaardverdeling is dat gewoon een rechte lijn onder 45°.



We zien inderdaad dat $y_1+y_{n-1}\sim y_r$, dat geldt overigens ook voor F(y) omdat de lijn onder 45° loopt.
Substitueren we dat in formule 1 met i=n, dan vinden we:

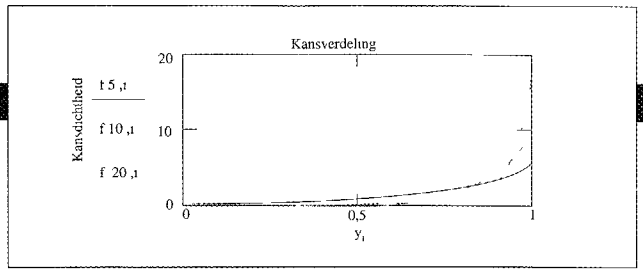
$$\mathbf{f(y_n)} = \frac{\mathbf{n!}}{\mathbf{(n-1)!(n-n)!}} \bullet \mathbf{y_n^{n-1}} \bullet \mathbf{(1-y_n)n-n}$$

en dit vereenvoudigt tot:

$$\mathbf{f(y_n)= n\bullet y_n^{n-1}}$$

Dit is voor elke n een kansverdeling omdat integreren tussen 0 en 1 voor y de waarde 1 oplevert zoals dat hoort.
We zullen voor een aantal waarden van **n** de kansdichtheid in grafiek 8 laten zien.

We zien hier het typische gedrag van een beta-functie.
Indien we de kansdichtheid van een stochastiek te pakken hebben kunnen we via het eerste en tweede moment het gemiddelde en de spreiding bepalen.

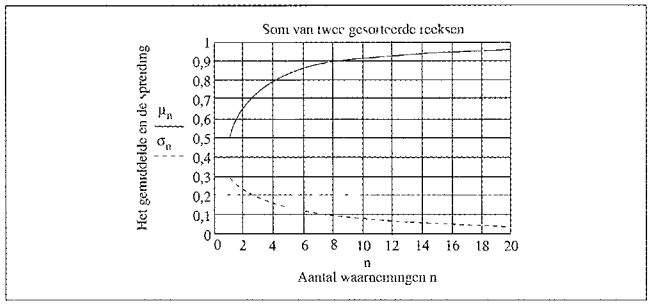


Grafiek 8

Hieronder ziet u de gevonden formules

$$\mu_n := \frac{n}{(1+n)} \qquad \sigma_n := \sqrt{\frac{n}{[(2+n)\bullet(1+n)^2]}}$$

We laten deze grootheden ook nog even in een grafiek zien:



Uit de grafiek lezen we af dat we al na 8 trekkingen op 90 % van het verwachte maximum van 1 zitten
Maar meer opmerkelijk is het dat in het gemiddelde en de spreiding de waarden van de individuele metingen niet meer voorkomen. Kennelijk bepaalt n het hele gedrag.

Als we met de transformatie gewerkt hebben moeten we de hier gevonden waarden natuurlijk terugtransformeren met $x_i = y(x_{max} - x_{min}) + x_{min}$ naar de oorspronkelijke waarden.

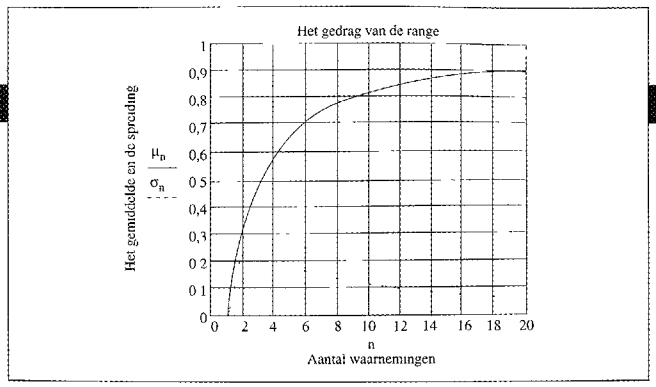
Dus bij de metingen hoeven we alleen nog maar te kijken naar het aantal metingen **n** en de grootste gemeten waarde. Het gemiddelde is dan gewoon deze '**grootste gemeten waarde gedeeld door 2**'. Zonder verder commentaar geef ik nog een handige formule.

> Als men **n** metingen heeft gedaan is de kans dat men een waarde vindt die groter is dan de reeds gevonden grootste waarde: $\frac{1}{n+1}$

We kunnen dezelfde exercitie natuurlijk ook uithalen met formule 2 van pagina 12 waarin de kansdichtheid staat van de range. We vinden dan:

$$\mu_n := \frac{(n-1)}{(n+1)} \qquad \sigma_n = \sqrt{\frac{(2\bullet n-2)}{[(2+n)\bullet(n+1)^2]}}$$

We zien weer hetzelfde, namelijk het gemiddelde en de spreiding worden alleen beheerst door het aantal metingen **n** dat is uitgevoerd (zie voren) We laten ook dit in grafiekvorm zien.



Het gedrag van de range

Vergelijken we dit met de vorige grafiek dan zien we dat het werken met de range minder efficient is, bij 8 waarnemingen zitten we op 80 %, maar dit is nog altijd heel mooi

Als er nu meer metingen beschikbaar komen is het mogelijk om, via histogrammen, na te gaan of er zich een statistische verdeling ontwikkelt die niet uniform is. Als deze verdeling symmetrisch is heeft dat geen consequenties voor het gemiddelde. Is de verdeling niet symmetrisch dan zien we dat en moet er natuurlijk worden bijgestuurd

Ir. F Doorschot, Eindhoven

Conclusies

1 Het cumulatief gemiddelde levert zeer snel (binnen 5 à 6 metingen) een bruikbaar gemiddelde.
2 Het toepassen van gesorteerde reeksen heeft hetzelfde effect.
3 Het combineren van deze twee methoden is eenvoudig en levert een optimaal resultaat.
4 Met de gegevens uit de statistiek zijn deze conclusies zeer goed te onderbouwen.
5 In de statistiek van de geordende waarnemingen zitten zeer veel mogelijkheden die helaas niet goed gebruikt worden.
6 Het populatie-concept is vaak niet houdbaar.
7 De statistiek maakt nauwelijks gebruik van informatie die al bekend is via andere wegen.
8 De normale verdeling komt minder vaak voor dan vaak wordt verondersteld.
9 De voorgestelde transformatie is bruikbaar.
10 De eerste metingen bij onderzoek gedragen zich als aselecte getallen.
11 Deze studie is overal toe te passen, maar vooral bruikbaar bij research/onderzoek, marketing en bij onderzoekingen die erg duur zijn.