“Gemiddelde” en ander bekende woorde

Pieta van Deventer

Trefwoorde: niesentrale momente, populasie, rangordestatistieke, rekenkundige gemiddelde, sentrale momente, standaardafwyking, steekproef, verwagte waarde

Ons word van jongs af bekendgestel aan die begrip van “die gemiddelde”. Het iemand al vir jou gevra om te verduidelik wat met “die gemiddelde”, of meer korrek, “die rekenkundige gemiddelde” van ’n steekproef van getalle bedoel word? Indien wel, sou jou antwoord heel waarskynlik soos volg gelui het: Tel al die getalle bymekaar en deel deur die aantal getalle. Of anders gestel in formule- of resepvorm, die rekenkundige gemiddelde, voorgestel deur $\overline{X}$, word gevind mbv die formule $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$.

Dit is wel die regte antwoord op die vraag: Hoe bereken ek die rekenkundige gemiddelde van ’n steekproef van getalle? Ongelukkig is dit nie die antwoord op die oorspronklike vraag wat handel oor die betekenis van “die gemiddelde”, of wat daarmee bedoel word nie. Dit wat hier bo beskryf word, is slegs ’n formule of ’n berekeningsresep. Tegnies beskou is $\overline{X}$ daardie getal wat sodanig is dat die som van die afwykings vanaf hierdie getal 0 moet wees. Anders gestel, $\overline{X}$ word deur die volgende uitdrukking: $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)=0$ gedefinieer.

Dit is soos die steunpunt van ’n wipplank. In ingenieurstaal is dit die eerste moment om die nulpunt, ’n uitdrukking wat nou-nou verder toegelig sal word. Op hierdie wyse staan $\overline{X}$ sentraal of in ’n sekere sin in die middel teenoor al die getalle in die steekproef. ’n Vreemde eienskap van $\overline{X}$ is dat dit nie een van die getalle in die steekproef hoef te wees nie, ook nie ’n element van die populasie waaruit die steekproef getrek is nie. $\overline{X}$ is uit en uit ’n wiskundige begrip. Dink aan die resultate wat met ’n dobbelsteen verkry kan word – in hierdie geval die populasiewaardes self. Die (rekenkundige) gemiddelde van 1, 2, 3, 4, 5, 6 is 3,5, dws ’n onmoontlike waarneming. Dis egter wel so dat $\left(1-3.5\right)+\left(2-3.5\right)+\left(3-3.5\right)+\left(4-3.5\right)+\left(5-3.5\right)+\left(6-3.5\right)=0$.

As ’n sentrale getal gesoek word wat ’n element van die steekproef móét wees, gebruik ons die *mediaan* of die *modus*. Die mediaan is fisies die middelste van die geordende getalle, dws georden volgens grootte. Indien die aantal getalle onewe is, is dit die middelste getal en as die aantal getalle ewe is, word dit gewoonlik beskou as die gemiddelde van die middelste twee getalle. Die mediaan kan dus ook suiwer wiskundig met interpolasie bereken word en is dan óók nie noodwendig ’n lid van die steekproef/populasie nie. Daarteenoor is die modus dié getal wat meer as enige ander een voorkom. Indien daar meer as een getal is wat ewe veel meer as enige ander getal voorkom, het ons ’n bimodale populasie of selfs ’n multimodale populasie. Dit kan dan ook wees dat daar géén modus is nie.

Daar is nog verskeie ander sentrale maatstawwe, soos die *harmoniese gemiddeldes* en die *geknipte gemiddeldes*, om maar enkeles te noem, maar dis nie nou ter sake nie.

In hierdie skrywe word in die eerste plek ’n eenvoudige toepassing getoon van die verwerking van statistiese data. Dit word dan veralgemeen na die generiese notasie en bewoording in statistiek uit die oogpunt van verwagte waardes en momente. Samevattend word dan aangetoon dat hierdie prosedures almal slegs spesiale gevalle is van wat as *genererende funksies* bekend staan. Dit word dan uitgebrei tot ’n meer algemene toepassing van genererende funksies en daar word oa aangetoon hoe sekere afleidings baie vergemaklik kan word deur van hierdie verwagte waardes en in die besonder momente gebruik te maak.

Bereken $\bar{x}$ vir die waarnemings van ’n steekproef wat hierna volg. Let op dat die onderkas vir waargenome waardes en die bokas (vorige paragraaf), dws $\bar{X}$, vir generiese waardes gebruik word, dws vir dit wat nog (teoreties) waargeneem moet word. Veronderstel ons data is 6, 7, 8, 4, 4, 1, 3, 5, 6, 8, 8, 8, 8, 4, 3, 7, 1, 1, 3, 1, 6, 5, 8, 9. Die notasie vir hierdie getalle in die volgorde waarin dit waargeneem is, is

$$x_1 = 6, x_2 = 7, x_3 = 8, x_4 = 4, x_5 = 4, x_6 = 9, x_7 = 3, x_8 = 5, x_9 = 6, x_{10} = 8, x_{11} = 8, x_{12} = 8, x_{13} = 8,$$
$$x_{14} = 4, x_{15} = 3, x_{16} = 7, x_{17} = 1, x_{18} = 1, x_{19} = 1, x_{20} = 1, x_{21} = 6, x_{22} = 5, x_{23} = 8, x_{24} = 9.$$

waar x_i na die *i*-ste waarneming in tydvolgorde verwys. Wanneer die data volgens grootte gerangskik word, dui die voetskrif tussen hakies aan die hoeveelste getal dit is itv grootte:

$$x_{(1)} = 1, x_{(2)} = 1, x_{(3)} = 1, x_{(4)} = 1, x_{(5)} = 3, x_{(6)} = 3, x_{(7)} = 4, x_{(8)} = 4, x_{(9)} = 4, x_{(10)} = 5, x_{(11)} = 5, x_{(12)} = 6,$$
$$x_{(13)} = 6, x_{(14)} = 6, x_{(15)} = 7, x_{(16)} = 7, x_{(17)} = 8, x_{(18)} = 8, x_{(19)} = 8, x_{(20)} = 8, x_{(21)} = 8, x_{(22)} = 8, x_{(23)} = 9,$$
$$x_{(24)} = 9.$$

Hierdie laaste voorstelling met sy besondere notasie staan as *rangordestatistieke* (Eng *order statistics*, of *rank order statistics*) bekend. So kan ’n mens onmiddellik sien dat die mediaan $x_{Me} = \frac{x_{(12)} + x_{(13)}}{2} = \frac{6+6}{2} = 6.$ Ewe-eens is die modus $x_{Mod} = 8$. Daar is geen ander getal wat soveel keer of selfs meer kere as die waarde 8 verskyn nie. Die gemiddelde is $\bar{x} = \frac{124}{24} = 5.1667$. Die volgende resultaat volg direk, nl $\sum_{i=1}^{n} \left(x_i - 5.1667 \right) = 0.$

Waarom die beheptheid met die rekenkundige gemiddelde as dit dan so maklik is om die modus en die mediaan te bereken en dit terselfdertyd nie eens noodwendig ’n realistiese of werklike getal is nie? Een van die vernaamste redes is dat $\bar{X}$ wiskundig baie gemaklik tot verdere resultate en gevolgtrekkings lei. Die modus en die mediaan is lastige maatstawwe in die sin dat hulle nie geredelik tot verdere wiskundige afleidings lei nie. Verdere berekeninge en veralgemenings daarmee is besonder lastig.

Rangskik nou die getalle en benoem die getalle anders. Dui die getal 1 aan deur x_1, 3 deur x_2 ens soos in die tabel aangetoon. Die simbool f_i is nou die frekwensie waarmee die getal x_i voorkom. Nou is die voorkoms van die datastel eenvoudiger en kan ’n mens $\bar{x}$ heel gemaklik bereken.

i	x_i	f_i	$f_i x_i$	$\frac{f_i}{n} x_i$	Of $\doteq p_i x_i$
1	1	4	4	$\frac{4}{24} \cdot 1$	$\doteq p\left(x_1\right) \cdot x_1$
2	3	2	6	$\frac{2}{24} \cdot 3$	$\doteq p\left(x_2\right) \cdot x_2$

3	4	3	12	$\frac{3}{24}\cdot 4$	$\doteq p(x_3)\cdot x_3$
4	5	2	10	$\frac{2}{24}\cdot 5$	$\doteq p(x_4)\cdot x_4$
5	6	3	18	$\frac{3}{24}\cdot 6$	$\doteq p(x_5)\cdot x_5$
6	7	2	14	$\frac{2}{24}\cdot 7$	$\doteq p(x_6)\cdot x_6$
7	8	6	48	$\frac{6}{24}\cdot 8$	$\doteq p(x_7)\cdot x_7$
8	9	2	18	$\frac{2}{24}\cdot 9$	$\doteq p(x_8)\cdot x_8$
		$n=\sum f_i = 24$	$\sum f_i x_i = 130$	$\sum \frac{f_i}{n} x_i = 5.167$	$\sum p(x_i) x_i \doteq 5.167$

’n Mens sou kon sê dat $\frac{f_i}{n} \doteq p(x_i)$, nl ’n benaderde waarskynlikheid, dws indien *n* groot genoeg is en die steekproef verteenwoordigend is van die populasie, sodat

$$\bar{x} = \frac{1}{n}\sum f_i x_i \doteq \sum p(x_i) x_i = \sum xp(x) = \mu .$$

Let op die gebruik van die simbool $\doteq$ om ’n benaderde waarde aan te dui. $\bar{x}$ dien dus as benadering vir die populasiegemiddelde, ook die verwagte waarde genoem, nl μ.

Steekproefwaarnemings kan die werklike waardes (populasieparameters) slegs benader, maar dis al wat tot ons beskikking is en dit is dus hierop dat ons berekening berus – ons weet mos van die halwe eier en die leë dop. Hier is op ’n intuïtiewe (puriste sal sê kru) manier die ooreenkoms tussen die steekproefgemiddelde en die populasiegemiddelde, beter bekend as die teoreties verwagte waarde, aangetoon. Hou in gedagte dat die steekproef slegs beperkte inligting oor die populasie bevat, sodat $\bar{X}$ die populasiegemiddelde, μ, slegs by benadering kan beraam. Om redes wat later aangetoon sal word, staan $\mu = \sum xp(x)$ as *die eerste moment om die oorsprong*, of ook as die *verwagte waarde van X*, bekend. Die simbool $p(x)$ word algemeen vir diskrete verdelings en $f(x)$ vir kontinue verdelings gebruik. In die teorie is dit dikwels die konvensie om $f(x)$ generies te gebruik ongeag of die onderliggende verdeling diskreet of kontinu van aard is.

Dis nou duidelik dat ek op hierdie wyse die verwagte waarde, dws $\mu = E(X)$, van enige waarskynlikheidsmassaverdeling of waarskynlikheidsdigtheid kan aflei, dis te sê as dit bestaan. Ja, daar is enkele vreemde verdelings wat die lewe baie moeilik kan maak, maar ons gaan nie met hulle heul nie. In die geval van digtheidsfunksies (kontinue veranderlike) word natuurlik oor al die waardes van X, ook die definisiegebied van X genoem, geïntegreer ipv gesommeer.

Beskou eers drie voorbeelde. Ons brei dan later verder op hierdie voorbeelde uit om sekere begrippe beter toe te lig.

Beskou die verwagte waarde van die Bernoulliverdeling, dws waar $X \sim Bernoulli(p)$. Die Bernoulliverdeling bestaan uit slegs een eksperiment in ’n binomiaal scenario. Die waarskynlikheid van sukses in die enkele eksperiment is *p.* Die waarskynlikheidmassaverdeling is dus

$$\begin{aligned} f(x) &= p^x(1-p)^{1-x}, \quad x = 0,1 \\ &= 0 \qquad\qquad\qquad \text{andersins.} \end{aligned}$$

Die waarskynlikheid van ’n mislukking, dws dat $x = 0$, is $1-p$ en die waarskynlikheid van sukses, dws dat $x = 1$, is p. Kontroleer dit gerus.

Die verwagte waarde van die Bernoulliverdeling word soos volg afgelei:

$$\begin{aligned} \mu = E(X) &= \sum_{x=0}^{1} xf(x) \\ &= \sum_{x=0}^{1} xp^x(1-p)^{1-x} = 0\cdot p^0(1-p) + 1\cdot p^1(1-p)^{1-1} \\ &= p \end{aligned}$$

Die verwagte waarde van die binomiaalverdeling, d.w.s. waar $X \sim binomiaal(n, p)$

d.w.s. $f(x) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x = 0,1,2,3,\ldots,n$ word soos volg afgelei:

$$\begin{aligned} \therefore \mu = E(X) &= \sum_{x=0}^{n} xf(x) \\ &= \sum_{x=1}^{n} x\binom{n}{x} p^x(1-p)^{n-x} \quad (x = 0 \text{ lei tot 0 in eerste term}) \\ &= \sum_{x=1}^{n} x\frac{n\cdot(n-1)\cdot(n-2)\cdots 1}{x!(n-x)!} p^x(1-p)^{n-x} \\ &= \sum_{x=1}^{n} \cancel{x}\frac{n\cdot(n-1)\cdot(n-2)\cdots 1}{(x-1)!(n-x)!} p^x(1-p)^{n-x}, \text{stel nou } y = x-1; \\ &\qquad\qquad\qquad \therefore x = 1 \text{ tot } n \text{ lei tot } y = 0 \text{ tot } n-1 \\ &= np\sum_{y=0}^{n-1} \frac{(n-1)\cdot(n-2)\cdots 1}{y!((n-1)-y)!} p^y(1-p)^{n-1-y} \\ &= np\sum_{y=0}^{n-1} \frac{(n-1)\cdot(n-2)\cdots 1}{y!(n-(y+1))!} p^y(1-p)^{(n-1)-y} \\ &= np\sum_{y=0}^{n-1} \binom{n-1}{y} p^y(1-p)^{n-1-y} \\ &= np(p+1-p)^{n-1} \qquad (\text{volgens die binoomstelling}) \\ &= np \end{aligned}$$

Hierdie laaste afleiding was nie heeltemal so eenvoudig nie, maar in ’n latere artikel sal ons sien dat dit moontlik is om hierdie afleiding baie maklik direk te vind deur van die eienskappe van die sg *momentvoortbringende funksie* gebruik te maak.

As ’n voorbeeld in die geval van ’n kontinue veranderlike word die verwagte waarde van die eksponensiaalverdeling, waar $T \sim \exp(\lambda)$ d.w.s. $f(t) = \lambda e^{-\lambda t}, \quad t \geq 0$, afgelei.

$$\begin{aligned}
\therefore \mu = E(T) &= \int_{t=0}^{t=\infty} tf(t)\,dt \\
&= \int_{t=0}^{t=\infty} t\lambda e^{-\lambda t}\,dt \\
&= \lambda \int_{t=0}^{t=\infty} te^{-\lambda t}\,dt \\
&= \lambda \left[t \cdot \frac{e^{-\lambda t}}{-\lambda} - \int_0^{\infty} \frac{e^{-\lambda t}}{-\lambda}\,dt \right]_0^{\infty} \qquad \left(\int f \cdot g = f \int g - \int \left(f' \int g \right) = f \cdot G - \int f' \cdot G \right) \\
&= \lambda \left[t \cdot \frac{e^{-\lambda t}}{-\lambda} \right]_{t=0}^{t=\infty} - \lambda \left[\frac{e^{-\lambda t}}{\lambda^2} \right]_{t=0}^{t=\infty} \\
&= \lambda[0] - \lambda \left[0 - \frac{1}{\lambda^2} \right] \\
&= \frac{1}{\lambda}
\end{aligned}$$

Die verwagte wagtyd voor ’n gebeurtenis plaasvind is dus direk omgekeerd eweredig aan die koers van voorkoms van die gebeurtenis. Bepeins dit gerus: 1 gebeurtenis per 60 minute is ekwivalent aan gemiddeld $\frac{1}{60}$-ste gebeurtenis per minuut! Die eerste moment om die oorsprong, dws om nul, is slegs ’n spesiale geval van ’n baie kragtige gereedskapstuk in die hande van die wiskundige en daarmee saam die wiskundige statistikus. Fisies kan dit gesien word as die wig steunpunt waarom’n wipplank balanseer. Dit is die rede waarom die steunpunt ook as die draaipunt bekend staan.

In die algemeen staan $E(X) = \mu = \sum xf(x)$ as die verwagte waarde van X onder $f(x)$ bekend. Die term *genererende funksie* word ook dikwels in hierdie verband gebruik. Die definisie van *genererende funksie* strek egter heelwat verder as wat ons hier gaan bespreek, maar tans is dit vir ons genoeg. In die algemeen word die verwagte waarde van ’n algemene funksie $h(X)$ onder $f(x)$ as

$$E(h(X)) = \sum_x h(x) f(x)$$

geskryf waar $h(X)$ enige nuttige funksie kan wees, soos hier onder aangedui. Weereens sal integrasie in die geval van ’n kontinue veranderlike gebruik word. Ons gaan ’n aantal daarvan voorstel en bestudeer. Hoe onwaarskynlik dit ook al mag lyk, is daar dikwels eenvoudige, direkte verbande tussen die verwagte waardes wat so geskep is. Op hierdie manier kan dan aangetoon word dat daar verskeie maniere is waarop populasiegemiddeldes, variansies, en ’n wavrag ander maatstawwe bereken kan word – sonder om jou met ingewikkelde afleidings te vermoei. Die een volg mbv ’n eenvoudige vervanging (substitusie) dikwels direk uit die ander. So sal ons by ’n latere geleentheid waarneem hoe dit selfs in gevorderde wiskunde soos Laplace-transformasies, waarskynlikheidsvoortbringende funksies, toustaanteorie, momentvoortbringende funksies, magsfunksies, Fouriertransformasie en ’n horde ander plekke ’n onontbeerlike rol speel. Hierdie elegansie wat in die genererende funksies skuil, is juis wat dit so aantreklik en nuttig maak.

Definieer nou $h(X)$ om die beurt as een van die volgende funksies:

$$h(X) = (X-a)^m,\ m = 1,2,3,4 \text{ en } a = 0 \text{ of } \mu \text{ of die mediaan of}$$
$$h(X) = e^{sX} \text{ of } s^X \text{ of } e^{isX}$$

Die waarde van a word bepaal deur die scenario ter sprake. Daar is natuurlik ’n hele klomp ander vorme van $h(X)$ wat in die praktyk voorkom.

Hier bo is reeds die geval vir $f(X) = (X-a)^m$ met $a = 0, m = 1$ met drie toepassings aangetoon. Beskou nou hierdie funksie eers vir $a = 0$ en $m = 1,2,3$ en 4 en dan vir $a = \mu$ en $m = 1,2,3$ en $4.$ Op hierdie manier kan ’n mens die patrone en die besondere notasie ter sprake makliker waarneem. Let op die notasie. Die res volg eers later. Wanneer $a = 0$, praat ons van *momente om die oorsprong*, of die *niesentrale momente*, omdat dit nie om die sentrale waarde, μ, bereken word nie. Wanneer $a = \mu$ praat ons van die *momente om die gemiddelde* of die *sentrale momente*, omdat dit om die gemiddelde as bron van sentraliteit geskep word. Die mediaan word nie so dikwels gebruik nie. Die notasie wat hier gebruik word, geld vir diskrete verdelings en daarom is die operator die sommasie. Presies dieselfde formulerings geld in die geval van kontinue verdelings met dié verskil dat van integrasie ipv sommasie gebruik gemaak word. Dus word dit nie vir die kontinue geval herhaal nie.

$\mu_1' = \sum_x xf(x) = \mu,$ wat agv konvensie vereenvoudig word na μ alleen, (nie μ_1 nie)

- die eerste niesentrale moment

$\mu_2' = \sum_x x^2 f(x)$ - die tweede niesentrale moment

$\mu_3' = \sum_x x^3 f(x)$ - die derde niesentrale moment

$\mu_4' = \sum_x x^4 f(x)$ - die vierde niesentrale moment

En die eerste vier sentrale momente

$$\mu_1 = \sum_x (x-\mu) f(x) = \sum_x xf(x) - \sum_x \mu f(x) = \mu - \mu \sum_x f(x) = \mu - \mu \cdot 1 = 0$$

$-$ die eerste sentrale moment

$\mu_2 = \sum_x (x-\mu)^2 f(x)$ $-$ die tweede sentrale moment

$$= \sum_x x^2 f(x) - 2\mu \sum_x xf(x) + \mu^2 \sum_x f(x) = \sum_x x^2 f(x) - 2\mu^2 + \mu^2 \quad \left(\sum_x f(x) = 1 \right)$$

$= \mu_2' - \mu^2 = \mathrm{var}(X) = \sigma^2$ $\quad$ –’n maatstaf van (ver)spreiding.

$$\mu_3 = \sum_x (x-\mu)^3 f(x) - \text{ die derde sentrale moment}$$
$$= \sum_x x^3 f(x) - 3\mu\sum_x x^2 f(x) + 3\mu^2 \sum_x x f(x) - \mu^3 \sum_x f(x)$$
$$= \sum_x x^3 f(x) - 3\mu\mu_2' + 3\mu^2\mu - \mu^3$$
$$= \mu_3' - 3\mu\mu_2' + 2\mu^3 \qquad \text{–'n basis vir 'n maatstaf van skeefheid}$$
$$\mu_4 = \sum_x (x-\mu)^4 f(x) - \text{ die vierde sentrale moment}$$
$$= \ldots$$
$$= \mu_4' - 4\mu\mu_3' + 6\mu^2\mu_2' - 3\mu^4 \qquad \text{–'n basis vir 'n maatstaf van spitsheid}$$

Dit beteken dus dat die sentrale momente nooit direk bereken hoef te word nie, maar bereken kan word deur gebruik te maak van die verbande met die niesentrale momente, wat nie net makliker is om te bereken nie, maar in die praktyk ook tot minder afrondingsfoute lei. Op hierdie wyse kan die variansie, σ^2, dus direk van die eerste twee niesentrale momente afgelei word.

Bostaande momente geld vir enige diskrete waarskynlikheidsverdeling. Weer eens geld dieselfde verbande in die geval van 'n kontinue verdeling met die enigste verskil dat integrasie ipv sommasie gebruik word.

In die geval van werklike steekproefwaardes is die notasie feitlik identies, maar die $\mu-$simbool word deur *m* (vir *moment*) vervang, en binne-in die formules word die onderskeie waardes vir $f(x)$ deur die relatiewe frekwensieverhoudings vervang, sodat die notasie vir die steekproefmomente nou so lyk:

$$m_1' = m, m_2', m_3', m_4', m_1 = 0, m_2, m_3, m_4 \text{ en } m_2 = s^2 \text{ as die beramer vir } \sigma^2.$$

Beskou nou weer die numeriese voorbeeld aan die begin om die mees basiese momente te bereken.

x_i	f_i	$f_i x_i$	$f_i x_i^2$
1	4	4	4
3	2	6	18
4	3	12	48
5	2	10	50
6	3	18	108
7	2	14	98
8	6	48	384
9	2	18	162
	$n = \sum f_i = 24$	$\sum f_i x_i = 130$	$\sum f_i x_i^2 = 872$

$$m = \overline{x} = \frac{1}{n}\sum f_i x_i = \sum \frac{f_i}{n} x_i = 5.167$$

$$m_2' = \frac{1}{n}\sum f_i x_i = \sum \frac{f_i}{n} x_i^2 = 36.\dot{3}$$

$m_2 = m_2' - m^2 = m_2' - \overline{x}^2 \doteq 9.635 = s^2$ as steekproefberaming vir σ^2.

Indien die steekproef verteenwoordigend is van die populasie waaruit die data verkry is, is 5,167 'n redelik goeie beraming vir die onbekende populasiegemiddelde, μ. (Let op dat die woord *beraming* vir die numeriese resultaat van 'n *beramer* gebruik word. Engels *estimate vs estimator*.) Ewe-eens is 9,635 'n redelik goeie beraming vir die onbekende populasievariansie, σ^2.[1] Wat meer is, indien die populasie benaderd normaal (klokvormig) verdeel is, kan die volgende bewerings met 'n redelike mate van vertroue gemaak word, nl dat rofweg 90%, 95%, 98% en 99% van die populasiewaardes tussen die volgende grense lê:

$$\text{90\%-vertrouensinterval: } \overline{x} \pm 1.645s = 5.167 \pm 1.645\sqrt{9.635} = 0.070;10.282$$

$$\text{95\%-vertrouensinterval: } \overline{x} \pm 1.96s = 5.167 \pm 1.96\sqrt{9.635} = -0.908;11.260$$

$$\text{98\%-vertrouensinterval: } \overline{x} \pm 2.335s = 5.167 \pm 2.33\sqrt{9.635} = -2.056;12.408$$

$$\text{99\%-vertrouensinterval: } \overline{x} \pm 2.576s = 5.167 \pm 2.576\sqrt{9.635} = -2.820;13.172$$

Dis belangrik om daarop te let dat 'n vertrouensinterval nie 'n waarskynlikheidsinterval is nie. Die populasiegemiddelde is in die interval, of dit is nie in die interval nie. Daar is geen waarskynlikheid by betrokke nie, want dis 'n feit dat óf die een óf die ander een geld. Ons kan slegs 'n vertroue uitspreek oor of die interval die populasiegemiddelde bevat of nie. Dis al. Daarom word van 'n vertrouensinterval gepraat. In die geval van waarnemings uit 'n populasie wat uit positiewe waardes alleenlik bestaan, sal 'n negatiewe berekende linkergrens hier bo natuurlik deur 0 vervang word.

Die algemene uitdrukkings vir die verdere momente kan ook nou afgelei word om direk bereken te kan word. Ingenieurs sal die gebruik van sentrale sowel as niesentrale momente dadelik herken as deel van hulle daaglikse gereedskap in die berekeninge van kragte om een of ander steunpunt.

Gebruik weer die voorbeelde hier bo en lei die tweede niesentrale momente asook die variansies af.

Net om bostaande vir 'n diskrete sowel as 'n kontinue verdeling aan te toon, word die afleidings vir die Bernoulliverdeling sowel as die eksponensiale verdelings gegee:

Die Bernoulliverdeling, waarvoor die verwagte waarde reeds hier bo afgelei is.

[1] Hier kom ook die begrippe *sydige* en *onsydige beramers* en die invloed van die grootte van die steekproef ter sprake. Dis egter 'n gesprek vir 'n ander dag. Soms word daar werklik baie moeite gedoen om akkurater antwoorde te kry, maar met min wins. So 'n geval vind 'n mens byvoorbeeld in die sg kontinuïteitskorreksie wanneer die binomiaalverdeling met die normaalverdeling benader word. Die oes op die moeite is onbenullig as in ag geneem word dat 'n mens in elk geval met benaderings en dus per definisie met onakkuraathede te doen het. Soms is dit egter die moeite werd, maar dis nie nou se gesprek nie.

$$
\begin{aligned}
E\left(X^2\right) &= \sum_{x=0}^{1} x^2 p^x \left(1-p\right)^{1-x} \\
&= p \\
\sigma^2 &= E\left(X^2\right) - \left[E\left(X\right)\right]^2 = \mu_2' - \mu^2 \\
&= p - p^2 \\
&= p\left(1-p\right)
\end{aligned}
$$

Wanneer dieselfde prosedure met die binomiaalverdeling gevolg word, word dinge effens lastig. Hier is die gebruik van die eienskappe van die momentvoortbringende funksie baie nuttig, soos in ’n volgende artikel uiteengesit sal word.

Die eksponensiaalverdeling, waar $T \sim \exp\left(\lambda\right)$ dws $f\left(t\right) = \lambda e^{\lambda t}, \quad t \geq 0$ se variansie word nou afgelei:

$$
\begin{aligned}
E\left(T^2\right) &= \int_0^\infty t^2 \cdot \lambda e^{-\lambda t} dt \\
&= t^2\left(-e^{-\lambda t}\right)\Big|_0^\infty + 2\int_0^\infty t e^{-\lambda t} dt \quad \left(\int f \cdot g = f\int g - \int\left(f'\int g\right) = f \cdot G - \int f' \cdot G\right) \\
&= \left(-0+0\right) + 2\frac{1}{\lambda^2} \quad \text{- verwys die afleiding van } \mu \\
&= 2\frac{1}{\lambda^2}
\end{aligned}
$$

Dus volg dit dat die variansie

$$
\begin{aligned}
\sigma^2 &= E\left(T^2\right) - \left[E\left(T\right)\right]^2 \\
&= 2\frac{1}{\lambda^2} - \frac{1}{\lambda^2} \\
&= \frac{1}{\lambda^2}
\end{aligned}
$$

Hierdie afdeling was slegs ’n inleiding tot verwagte waardes en momente. Ons het ook gesien dat die afleidings, soos in die geval van die binomiaalverdeling, soms lastig kan word. By ’n volgende geleentheid sal ons sien hoe sekere vorme van $h\left(X\right)$ soms van hierdie afleidings kinderspeletjies kan maak. Hou dus dop vir die *momentvoortbringende funksies* (*mvf*) by ’n volgende aflewering. Die naam is presies wat dit sê – dit genereer die momente in ’n kits, selfs van meer komplekse verdelings. Teen hierdie tyd het jy hopelik genoeg om aan te herkou, maar moenie die volgende episode oor die mvf misloop nie. Dit maak die lewe sommer baie makliker.

’n Laaste opmerking is miskien nodig. Hoekom word die maatstaf van verspreiding, nl die variansie, as ’n som van kwadrate bereken en dan word die vierkantswortel daarvan, nl die standaardafwyking, gebruik? Hoekom eers kwadreer? Dit lyk mos nie logies nie. Wel, indien slegs die gemiddelde afwykings soos volg bereken word, nl $\sum_x \left(x-\mu\right) f\left(x\right)$

$\sum_x (x-\mu) f(x) = \sum_x x f(x) - \mu \sum_x f(x) = \mu - \mu \cdot 1 = 0.$ Dus help dit ons op geen manier nie. ’n Meer korrekte berekening van die gemiddelde afwyking is natuurlik $\sum_x |x-\mu| f(x)$, die sg *absolute verwagte verskil*, wat ook as die *absolute gemiddelde spreiding* bekend staan. Dit is egter ’n baie lastige uitdrukking om verdere analises mee te maak. Die standaardafwyking $\sigma = \sqrt{\sum_x (x-\mu) f(x)}$ is ’n intuïtiewe uitbreiding van die Wet van Pythagoras in *n* dimensies en is baie meer hanteerbaar vir verdere ontledings, al lyk dit op die oog af meer kompleks.