STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Populatiemodellen en normaal verdeelde populaties

2. Eigenschappen van kansmodellen

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Eigenschappen van kansmodellen

1. Het gemiddelde van kansmodellen

Een kansmodel zegt op welke manier getallen tot jou komen

- In het discrete geval vertelt een kansmodel je welke uitkomsten er allemaal mogelijk zijn en met welke kans zij optreden.

- In het continue geval wordt de verzameling van alle mogelijke uitkomsten aangeduid door een interval. Daar hoort een dichtheidsfunctie bij die je toelaat om voor alle mogelijke deelintervallen de kans te berekenen om daarin terecht te komen.

Over het gemiddelde van een kansmodel X kan je als volgt nadenken. Stel je voor dat je heel veel keren (een oneindig aantal keer) het kansmodel X gebruikt om een getal te genereren. Wat zou dan het gemiddelde van al die getallen zijn? Wat verwacht je?

Het gemiddelde dat je verwacht wordt ook "**verwachtingswaarde**" genoemd. Het Engelse woord voor "verwachting" is "expectation" en het is de **E** van **E**xpectation die als afkorting wordt gebruikt. Het gemiddelde of de verwachtingswaarde van het kansmodel X wordt genoteerd door $E(X)$.

Het gemiddelde van een kansmodel is **een vast getal**. Het is een kenmerk voor het centrum van dat model. Van zodra je het model kent zou je dus ook dat kenmerk moeten kunnen bepalen, zelfs zonder dat je eerst getallen door dat model hebt laten genereren. Kijk hoe dat gaat.

2. Het gemiddelde van discrete kansmodellen

2.1. Een eerlijke dobbelsteen

Denk aan het model X voor een eerlijke dobbelsteen. Je kan dat voorstellen door een vaas of door een staafdiagram of met een tabel. Die ziet er als volgt uit.

x	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Als je deze dobbelsteen heel veel keren zou gooien, bijvoorbeeld 6 miljoen keer, dan verwacht je dat je “ongeveer” 1 miljoen keer een 1, “ongeveer” 1 miljoen keer een 2, enz. zal hebben. Als je dan het gemiddelde zou maken van die 6 miljoen getallen dan zou je vinden:

$$\frac{1}{6\,miljoen}\left(ongeveer\,1\,miljoen\;keer\;1+ongeveer\,1\,miljoen\;keer\;2+ongeveer\,1\,miljoen\;keer\;3+....\right)$$

$$= \frac{ongeveer\,1\,miljoen}{6\,miljoen}\times 1 \quad + \quad \frac{ongeveer\,1\,miljoen}{6\,miljoen}\times 2 \quad + \quad \frac{ongeveer\,1\,miljoen}{6\,miljoen}\times 3 \quad + \quad .....$$

$$= \left(ongeveer\,\frac{1}{6}\right)\times 1 \quad + \quad \left(ongeveer\,\frac{1}{6}\right)\times 2 \quad + \quad \left(ongeveer\,\frac{1}{6}\right)\times 3 \quad + \quad ....$$

Als je nog veel meer keren zou gooien dan zal $\left(ongeveer\,\frac{1}{6}\right)$ dichter en dichter naar de kans $\frac{1}{6}$ naderen en er uiteindelijk mee samenvallen. Het gemiddelde van een eerlijke dobbelsteen (wat je als gemiddelde verwacht te vinden als je oneindig keer zou kunnen gooien) is dus gelijk aan

$$E(X)=\left(\frac{1}{6}\right)\times 1 \quad + \quad \left(\frac{1}{6}\right)\times 2 \quad + \quad \left(\frac{1}{6}\right)\times 3 \quad + \quad \left(\frac{1}{6}\right)\times 4 \quad + \quad \left(\frac{1}{6}\right)\times 5 \quad + \quad \left(\frac{1}{6}\right)\times 6$$

Als je dit uitrekent vind je dat $E(X)=3.5$.

2.2. De rode dobbelsteen

Het kansmodel van de rode dobbelsteen ziet er als volgt uit.

x	1	3	6
$P(X=x)$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Opdracht 1

Stel je nu voor dat je met deze dobbelsteen heel veel keren zou gooien en dan het gemiddelde berekenen van de gevonden getallen. Gebruik een redenering zoals hierboven en begin met te onderstellen dat je 6 miljoen keer zou gooien. Toon aan hoe je dan verder moet redeneren om aan het echte gemiddelde van de rode dobbelsteen te komen. Bereken tenslotte dat gemiddelde.

2.3. Algemene definitie

Als je goed naar de twee voorgaande voorbeelden kijkt dan kan je daar een algemene structuur in ontdekken.

De formule voor het gemiddelde $E(X)$ van een kansmodel X ziet er uit als een som. In die som komen alle mogelijke uitkomsten één keer voor. Maar zij staan daar niet zomaar. Zij worden vermenigvuldigd met een getal. In de wiskunde spreekt men dan soms over "gewichten" en over een "gewogen som". Het is alsof je elke uitkomst een gewichtje geeft waarmee het in die som moet voorkomen.

Het gemiddelde $E(X)$ van een discreet kansmodel X is een gewogen som van uitkomsten. Het gewicht dat je aan een bepaalde uitkomst geeft is de kans van die uitkomst. Als een uitkomst met een grotere kans voorkomt dan krijgt die uitkomst een groter gewicht.

Voor de rode dobbelsteen heb je:

$$\begin{aligned} E(X) &= 1 \cdot \left(\frac{3}{6}\right) + 3 \cdot \left(\frac{2}{6}\right) + 6 \cdot \left(\frac{1}{6}\right) \\ &= 1 \cdot P(X=1) + 3 \cdot P(X=3) + 6 \cdot P(X=6) \end{aligned}$$

Als je de drie mogelijke uitkomsten van die rode dobbelsteen met een algemene notatie voorstelt door x_1 , x_2 en x_3, dan kan je dat gemiddelde schrijven als:

$$E(X) = 1 \cdot P(X=1) + 3 \cdot P(X=3) + 6 \cdot P(X=6)$$

of

$$E(X) = x_1 \cdot P(X=x_1) + x_2 \cdot P(X=x_2) + x_3 \cdot P(X=x_3)$$

In woorden zeg je dat het gemiddelde van een discreet kansmodel gelijk is aan de som van "de uitkomsten maal hun kansen".

Voor een algemeen discreet kansmodel X met uitkomsten x (dus met uitkomsten van de vorm x_1 , x_2 , … , x_i ,… , x_n) wordt het gemiddelde gegeven door

$$E(X) = x_1 \cdot P\left(X = x_1\right) + x_2 \cdot P\left(X = x_2\right) + ... + x_n \cdot P\left(X = x_n\right)$$

of korter

$$E(X) = \sum_{i=1}^{n} x_i \, P\left(X = x_i\right)$$

Opdracht 2

- Wat is het verschil tussen het gemiddelde van een kansmodel X en het gemiddelde van 400 uitkomsten die door dat kansmodel X zijn gegenereerd? Om dat te onderzoeken gebruik je het model X van de rode dobbelsteen. Schrijf de frequenties van de 400 getallen die je vroeger hebt gevonden (in opdracht 16 van de werktekst "De wereld van de kansmodellen") over in onderstaande tabel. Bereken dan het gemiddelde van die 400 getallen. Herinner je van vroeger dat het gemiddelde van getallen voorgesteld wordt door $\overline{x}$.

uitkomst x_i	frequentie f_i
1	
3	
6	

- Wat heb jij gevonden voor $\overline{x}$ en wat was je uitkomst voor $E(X)$? Kijk nu wat je medeleerlingen gevonden hebben voor $\overline{x}$ en voor $E(X)$. Vat dit samen in onderstaande tabel (bijvoorbeeld voor 5 leerlingen).

	$E(X)$	$\overline{x}$
leerling 1		
leerling 2		
leerling 3		
leerling 4		
leerling 5		

- Welk belangrijk verschil tussen $\overline{x}$ en $E(X)$ heb je nu ontdekt? Is het gemiddelde $E(X)$ een vast getal of is $E(X)$ aan het toeval onderhevig? Is dat ook zo voor $\overline{x}$?

3. Het gemiddelde van continue kansmodellen

3.1. Algemene definitie

Voor continue kansmodellen moet je wiskundige technieken voor discrete getallen overzetten naar wiskundige technieken voor een continuüm. In plaats van de uitkomsten x te vermenigvuldigen met hun bijhorende kans $P(X = x)$ moet je nu de waarden x vermenigvuldigen met de dichtheidsfunctie $f(x)$. En in plaats van te sommeren moet je nu integreren. Dat wordt dus

$$\boxed{E(X) = \int_{-\infty}^{+\infty} x\ f(x)\, dx}$$

De bepaalde integraal kan je in principe altijd laten lopen van $-\infty$ tot $+\infty$ want je kan elke dichtheidsfunctie uitgebreid definiëren over de volledige getallenas.
Als je een model hebt dat enkel “leeft” op het interval [0 ; 1] dan kan je evengoed schrijven dat het gemiddelde van dat model gelijk is aan $E(X) = \int_0^1 x\ f(x)\, dx$ want de functie $f(x)$ is dan toch gelijk aan nul voor alle x-waarden buiten het interval [0 ; 1].

3.2. De uniforme dichtheid op [0 ; 1]

Figuur 1

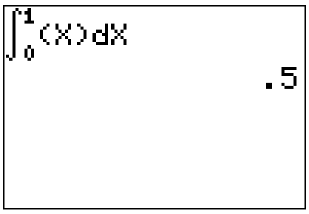
Het continue kansmodel voor lukraak trekken uit het interval [0 ; 1] heeft als dichtheidsfunctie f(x) = 1 voor $0 \le x \le 1$ (zie figuur 1).

Als je met dit model werkt, wat verwacht je dan gemiddeld uit te komen? Herinner je dat alle deelintervallen van gelijke lengte gelijkwaardig zijn en dat je daar met dezelfde kans in terechtkomt. Bij heel veel trekkingen verwacht je dus getallen “van alle soorten”, zowel dicht bij nul als in het midden als dicht bij één. Het gemiddelde van heel veel dergelijke getallen zal dan misschien wel in het midden van het interval [0 ; 1] liggen?

Het echte gemiddelde $E(X)$ van dit uniforme kansmodel X kan je eenvoudig berekenen met je GRM. Kijk daarbij goed naar de definitie van het gemiddelde van een continu kansmodel. In dit voorbeeld moet je een integraal uitrekenen die loopt van 0 tot 1 en de functie die je moet integreren is volgens de algemene definitie altijd gelijk aan $x \cdot f(x)$. Je moet daarbij de juiste dichtheid $f(x)$ invullen. Hier is $f(x) = 1$ zodat $x \cdot f(x)$ hier gewoon gelijk is aan x.

Het gemiddelde van de uniforme op [0 ; 1] is dus gelijk aan $E(X) = \int_0^1 x\ dx$. Dit gemiddelde bereken je nu met je GRM.

Opdracht 3

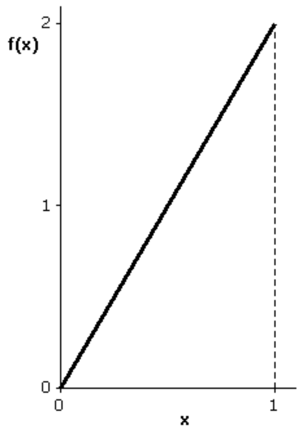


Druk MATH , loop naar 9:fnInt(en druk ENTER. Vul in zoals aangegeven en druk ENTER. De waarde van de integraal is gelijk aan 0.5. Dus is $E(X) = 0.5$ voor het kansmodel X waarbij je lukraak trekt uit [0 ; 1].

3.3. De dichtheid f(x)=2x op [0 ; 1]

Opdracht 4

- Werk met het kansmodel X waarvan de dichtheidsfunctie gegeven is door $f(x) = 2x, 0 \leq x \leq 1$. Kijk goed naar de grafiek van de dichtheid (figuur 2) en doe dan een gok naar wat het gemiddelde $E(X)$ hier zou kunnen zijn. Motiveer je antwoord.



Figuur 2

- Bereken $E(X)$ met je GRM.
  - Schrijf eerst de algemene formule voor het gemiddelde van een continu kansmodel:

    $E(X) =$ ....

  - Vul in de algemene formule de juiste dichtheidsfunctie in waarmee je hier moet werken (let ook op de grenzen van de integraal):

    $E(X) =$ ....

  - Om het gemiddelde te kennen moet je die integraal berekenen. Doe dat met je GRM zoals je dat in opdracht 3 gedaan hebt. Vul dan de onderstaande uitspraak in:

  Het kansmodel X met dichtheidsfunctie $f(x) = 2x, 0 \leq x \leq 1$ heeft een gemiddelde dat gelijk is aan $E(X) =$ ......

4. De standaardafwijking van kansmodellen

Bij een statistische studie zijn de getallen die je observeert aan het toeval onderhevig. Maar toeval is helemaal niet gelijk aan willekeur. Het kansmodel bepaalt op welke manier je die toevallige getallen te zien krijgt.

De manier waarop een kansmodel X de getallen rond het gemiddelde $E(X)$ laat vallen kan nogal verschillend zijn. Sommige kansmodellen zorgen ervoor dat de meeste getallen dicht bij het centrum $E(X)$ terechtkomen. Bij andere kansmodellen komen de getallen met een veel grotere spreiding rond het gemiddelde terecht. De **standaardafwijking** is een maat voor de spreiding rond $E(X)$. De standaardafwijking van een kansmodel X is een vast getal, genoteerd door $sd(X)$. Die afkorting komt van het Engelse **s**tandard **d**eviation.

De standaardafwijking van een kansmodel is, juist zoals de verwachtingswaarde, een *modeleigenschap*. Het is **een vast getal**. Als je het model kent dan kan je de standaardafwijking berekenen, je hebt daarvoor geen observatiegetallen nodig.

4.1. De standaardafwijking van discrete kansmodellen

Het kansmodel van de rode dobbelsteen staat hieronder.

x	1	3	6
$P(X=x)$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Het gemiddelde van dit model heb je al berekend, dat is $E(X) = 2.5$. Een maat voor de spreiding waarmee de uitkomsten 1, 3 en 6 rond het gemiddelde 2.5 vallen bereken je als volgt.

Begin met de eerste bouwsteen. Dat is het kwadraat van de afwijking van elke mogelijke uitkomst tot aan het gemiddelde. Hier is dat $(1-2.5)^2$, $(3-2.5)^2$ en $(6-2.5)^2$. Daarna maak je de som van die kwadratische afwijkingen. Niet zomaar een som, maar een gewogen som. Uitkomsten die met grotere kans voorkomen krijgen bij hun kwadratische afwijking een groter gewicht. Dat gewicht is, zoals vroeger, de bijhorende kans. Het model zegt dat de uitkomst 1 met kans $\frac{3}{6}$ voorkomt.

Dus moet je $(1-2.5)^2$ met $\frac{3}{6}$ vermenigvuldigen. Alles samen krijg je dan

$$(1-2.5)^2 \cdot \frac{3}{6} \quad + \quad (3-2.5)^2 \cdot \frac{2}{6} \quad + \quad (6-2.5)^2 \cdot \frac{1}{6}$$

Die gewogen som noem je de variantie van het model X, genoteerd als $\text{var}(X)$. Voor de rode dobbelsteen is $\text{var}(X) = (1-2.5)^2 \cdot \frac{3}{6} + (3-2.5)^2 \cdot \frac{2}{6} + (6-2.5)^2 \cdot \frac{1}{6} = 3.25$.
De standaardafwijking is de positieve vierkantswortel uit de variantie. Dat is hier $sd(X) = \sqrt{3.25} \approx 1.80$.

Je hebt zopas stap voor stap de formule opgebouwd voor de standaardafwijking van de rode dobbelsteen. Het is niet moeilijk om nu over te stappen op de algemene notatie bij discrete kansmodellen, waar de uitkomsten door x en de kansen door $P(X = x)$ worden voorgesteld. Let er hierbij op dat je het gemiddelde 2.5 vervangt door de algemene notatie voor het gemiddelde $E(X)$.

Voor de variantie heb je dan:

$$\text{var}(X) = (x_1 - E(X))^2 \cdot P(X = x_1) + \ldots \quad + (x_i - E(X))^2 \cdot P(X = x_i) + \ldots \quad + (x_n - E(X))^2 \cdot P(X = x_n)$$

of korter

$$\text{var}(X) = \sum_{i=1}^{n} \left[(x_i - E(X))^2 \cdot P(X = x_i) \right]$$

De standaardafwijking van een discreet kansmodel X is gelijk aan

$$sd(X) = \sqrt{\text{var}(X)} = \sqrt{\sum_{i=1}^{n} (x_i - E(X))^2 \cdot P(X = x_i)}$$

Opdracht 5

Het kansmodel X heeft maar 2 mogelijke uitkomsten: 0 (= mislukking) of 1 (= succes). De bijhorende kansverdeling zie je hieronder.

x	0	1
P(*X*=*x*)	0.4	0.6

- Zoek het gemiddelde $E(X) = \sum_{i=1}^{n} x_i \, P(X = x_i) = x_1 \, P(X = x_1) + x_2 \, P(X = x_2) = \ldots\ldots\ldots\ldots..$

- Pas de formule voor de standaardafwijking aan voor dit voorbeeld (zoals hierboven ook gedaan is voor het gemiddelde). Vul dan de waarden in en gebruik je GRM om de uitkomst te berekenen. Dit levert: $sd(X) = \ldots\ldots\ldots\ldots$

4.2. De standaardafwijking van continue kansmodellen

Bij continue modellen moet je terug overstappen van discrete wiskundetechnieken op continue wiskundetechnieken, juist zoals je dat bij de verwachtingswaarde hebt gedaan.

De eerste bouwsteen is terug de kwadratische afwijking van de waarden x tot aan het gemiddelde $E(X)$. Dat is dus $\left(x-E(X)\right)^2$. Deze uitdrukking moet je nu vermenigvuldigen met de dichtheidsfunctie $f(x)$ en in plaats van te sommeren moet je nu integreren.

Voor continue kansmodellen geldt dat $\text{var}(X) \;=\; \int\limits_{-\infty}^{+\infty} \left(x-E(X)\right)^2 \, f(x)\, dx$.

De standaardafwijking van een continu kansmodel X is gelijk aan
$sd(X)=\sqrt{\text{var}(X)} \;=\; \sqrt{\int\limits_{-\infty}^{+\infty} \left(x-E(X)\right)^2 \, f(x)\, dx}$

Voor het kansmodel X met dichtheid $f(x)=2x\,,\,0\le x\le 1$ heb je berekend dat $E(X)=\dfrac{2}{3}$. Als je dit invult in de algemene formule dan krijg je voor de variantie van X dat $\text{var}(X) \;=\; \int\limits_{-\infty}^{+\infty} \left(x-E(X)\right)^2 \, f(x)\, dx \;=\; \int\limits_{0}^{1} \left(x-\dfrac{2}{3}\right)^2 (2x)\, dx$. Deze variantie bereken je nu met je GRM. Daarna trek je de positieve vierkantswortel om de standaardafwijking te vinden.

Opdracht 6

Je weet nu al hoe je integralen met de GRM berekent. Zoek $\text{var}(X)=\int\limits_{0}^{1} \left(x-\dfrac{2}{3}\right)^2 (2x)\, dx$.

Vul in:

het kansmodel X met dichtheid $f(x)=2x\,,\,0\le x\le 1$ heeft

- een variantie die gelijk is aan $\text{var}(X)=$
- een standaardafwijking die gelijk is aan $sd(X)=$

5. Populatiemodellen

5.1. Populatieparameters

Je hebt geleerd dat voor elk kansmodel X

- het gemiddelde genoteerd wordt als $E(X)$
- de standaardafwijking genoteerd wordt als $sd(X)$ [en de variantie als $\text{var}(X)$].

Als het kansmodel X een populatiemodel is, dan gebruik je een speciale notatie.

Het gemiddelde $E(X)$ van een populatie X stel je voor als μ (Griekse letter mu)
De standaardafwijking $sd(X)$ van een populatie X stel je voor als σ (Griekse letter sigma)

In het kader van statistische studies maak je een duidelijk onderscheid tussen een populatie en een steekproef. Om goed te zien dat het over een populatie gaat gebruik je die speciale notatie.

Als je ergens μ ziet staan dan weet je dat het gaat over het gemiddelde van de populatie die in dat onderzoek wordt bestudeerd. Als je σ tegenkomt dan gaat het over de standaardafwijking van die populatie. Het gaat dus altijd over de bestudeerde populatie, of die nu de naam X of Y of Z heeft. Het gemiddelde van een populatie noem je altijd μ en de standaardafwijking altijd σ.

Het populatiegemiddelde μ en de populatiestandaardafwijking σ heten **populatieparameters**.

5.2. Notatieafspraak

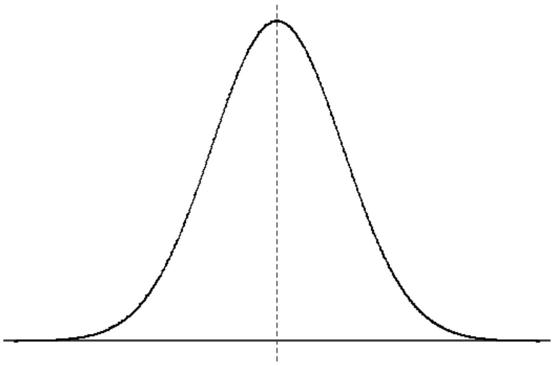
		algemeen kansmodel X	populatiemodel X
gemiddelde	discreet	$E(X)=\sum_{i=1}^{n} x_i\, P\left(X=x_i\right)$	$\mu=\sum_{i=1}^{n} x_i\, P\left(X=x_i\right)$
	continu	$E(X)=\int_{-\infty}^{+\infty} x\, f(x)\, dx$	$\mu=\int_{-\infty}^{+\infty} x\, f(x)\, dx$
standaardafwijking	discreet	$sd(X)=\sqrt{\sum_{i=1}^{n}\left(x_i-E(X)\right)^2\cdot P(X=x_i)}$	$\sigma=\sqrt{\sum_{i=1}^{n}\left(x_i-\mu\right)^2\cdot P(X=x_i)}$
	continu	$sd(X)=\sqrt{\int_{-\infty}^{+\infty}\left(x-E(X)\right)^2\, f(x)\, dx}$	$\sigma=\sqrt{\int_{-\infty}^{+\infty}\left(x-\mu\right)^2\, f(x)\, dx}$

6. Centrum, spreiding en vorm

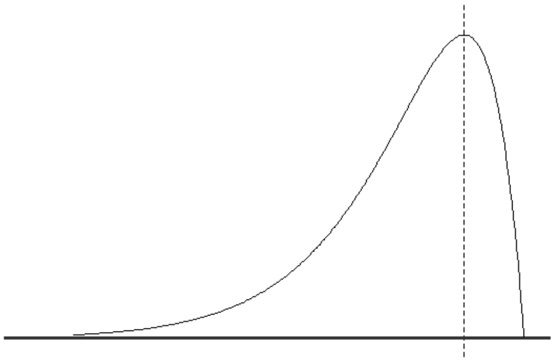
Sommige eigenschappen van kansmodellen kan je weergeven met een getal. Dat is zo voor het gemiddelde $E(X)$ en voor de standaardafwijking $sd(X)$.

Maar je moet ook altijd naar de globale vorm kijken. Om te beginnen kijk je naar symmetrie en scheefheid bij curven met één top. Daarna kan je eventueel andere eigenaardige patronen ontdekken zoals 2 verschillende toppen enz.

Een symmetrische dichtheidsfunctie.



Een dichtheidsfunctie die “scheef naar links” is.



Een dichtheidsfunctie die “scheef naar rechts” is.