STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Steekproefmodellen en normaal verdeelde steekproefgrootheden

4. Populatie en steekproef

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Populatie en steekproef

1. Populatiemodellen: even opfrissen

1.1. Kansmodellen voor populaties

Als je iets wil weten over de geboortegewichten van de kinderen die in 2002 in Vlaanderen geboren zijn, dan zou je uit de volledige namenlijst van die kinderen een steekproef van grootte 400 kunnen trekken. Je hebt dan 400 namen. Die namen verwijzen naar kinderen. Bij die kinderen horen geboortegewichten. Uiteindelijk heb je 400 gewichten, en die bestudeer je dan. Op die manier kan je zeggen dat “de populatie van geboortegewichten” 400 getallen naar jou heeft gestuurd.

In de statistiek beschrijf je een populatie met behulp van een “geïdealiseerd model”. Zo’n model vertelt je welke getallen je kan vinden en met welke kans. Het is een kansmodel.

- Je kan spreken over een populatie X waar een **kansverdeling** bij hoort. Dan beschrijf je een populatie waarbij de mogelijke uitkomsten discreet zijn.

- Je kan ook spreken over een populatie X waar een **kansdichtheid** bij hoort. Dat doe je om populaties aan te duiden waarbij de uitkomsten een continuüm (een interval) bestrijken.

Het kansmodel X van de rode dobbelsteen zegt welke getallen met welke kans tot jou komen als je die dobbelsteen zal gooien. Die populatie X is volledig vastgelegd door tabel 1.

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

Kansmodel van de populatie X

Tabel 1

1.2. Populatieparameters

Kansmodellen hebben modeleigenschappen, zoals de verwachtingswaarde (het gemiddelde) en de standaardafwijking. De algemene notatie voor die modeleigenschappen ken je. Het gemiddelde van het kansmodel X noteer je met $E(X)$ en voor de standaardafwijking schrijf je $sd(X)$.

Een populatie X wordt beschreven door een kansmodel en je kan dus de gekende formules voor kansmodellen gebruiken om het gemiddelde en de standaardafwijking van die populatie te berekenen. De getallen die je zo vindt worden **populatieparameters** genoemd. Zij krijgen een heel bijzondere 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)

Opdracht 1

Hoeveel is μ en σ voor de populatie X uit tabel 1 (de rode dobbelsteen)?

Voor het gemiddelde en de standaardafwijking van modellen gebruik je de volgende notaties en formules:

		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	$sd(X)=\sqrt{\int_{-\infty}^{+\infty}\left(x-E(X)\right)^2\, f(x)\, dx}$	$\sigma=\sqrt{\int_{-\infty}^{+\infty}\left(x-\mu\right)^2\, f(x)\, dx}$

2. Kansmodellen voor steekproeven

2.1. Een steekproef van grootte n=2: experimenteel en benaderend

Er zijn heel veel manieren om uit een populatie een steekproef te trekken. In deze tekst beperken we ons tot de EAS, de enkelvoudige aselecte steekproef. Bij een vaasmodel betekent dat "lukraak trekken met terugleggen". Je trekt telkens uit dezelfde populatie en het resultaat bij een volgende trekking wordt niet beïnvloed door een vorige (onafhankelijkheid).

Als populatie X neem je hier tabel 1. Dat is de rode dobbelsteen.

Uit deze populatie ga je nu een steekproef trekken. Om de berekeningen eenvoudig te houden trek je een heel kleine steekproef, namelijk een steekproef van grootte $n = 2$.

De eerste keer gooien noem je de eerste "trekking". Noteer het resultaat als x_1. Als jij met jouw dobbelsteen een één gevonden hebt, dan is voor jou $x_1 = 1$.

Voor een tweede "trekking" uit X gooi je die dobbelsteen opnieuw. Het resultaat noem je x_2. Als jij de tweede keer een zes gegooid hebt, dan heb jij $x_2 = 6$.

Als je nu zegt: "ik heb een steekproef van grootte $n = 2$ getrokken en één van mijn uitkomsten was een één en de andere was een zes" dan weet ik niet wat jij voor x_1 gevonden hebt. Als je dus wil zeggen wat het resultaat van jouw steekproef was, dan moet je alle informatie geven. Daar hoort ook de volgorde bij.

Een mogelijke uitkomst van een steekproef van grootte $n = 2$ is dus, **in volgorde**, het getal dat je bij de eerste trekking vindt gevolgd door het getal bij de tweede trekking. In de wiskunde noteer je dat als (x_1, x_2). Je spreekt dan van een "koppel" of een "geordend tweetal". De volgorde van de getallen is belangrijk.

Opdracht 2

Werk in een groepje van 5 leerlingen. Elke leerling voert zelf het experiment uit en daarna breng je de resultaten samen. Zoals eerder afgesproken experimenteer je hier met de rode dobbelsteen.

Jij hebt als uitkomst van je steekproef van grootte $n = 2$ bijvoorbeeld $(1, 6)$ gevonden. Wat zou er gebeuren als je nog eens een steekproef van grootte $n = 2$ uit diezelfde populatie zou trekken? En nog eens en nog eens? Probeer dat even uit. Zet het juiste vaasmodel in [L5] in je GRM en gebruik RANDVAAS waarbij je zegt dat je een steekproef van grootte $n = 2$ wil trekken. Het resultaat vind je dan in lijst [L1] waarbij het bovenste getal jouw x_1 is en het tweede getal jouw x_2 .Trek 10 keer zo'n steekproef van grootte $n = 2$ en noteer telkens welke uitkomst je vindt. Gebruik tabel 2.

	uitkomst 1e trekking	uitkomst 2de trekking	uitkomst steekproef
steekproef 1			
steekproef 2			
steekproef 3			
steekproef 4			
steekproef 5			
steekproef 6			
steekproef 7			
steekproef 8			
steekproef 9			
steekproef 10			

Tabel 2

Breng nu de resultaten van je groepje van 5 leerlingen samen zodat je ziet wat er gebeurt als je 50 keer een steekproef van grootte $n = 2$ trekt uit die populatie X. Gebruik de tabellen 3, 4, en 5.

uitkomst 1ste trekking	frequentie	relatieve frequentie
1		
3		
6		

Tabel 3

uitkomst 2de trekking	frequentie	relatieve frequentie
1		
3		
6		

Tabel 4

uitkomst steekproef	frequentie	relatieve frequentie
$(1,1)$		
$(1,3)$		
$(1,6)$		
$(3,1)$		
$(3,3)$		
$(3,6)$		
$(6,1)$		
$(6,3)$		
$(6,6)$		

Tabel 5

2.2. Een steekproef van grootte n=2: theoretisch en exact

Opdracht 3

Kijk nu eerst naar tabel 3. Daar staat een samenvatting van de resultaten bij de eerste trekking. Je merkt dat de enig mogelijke uitkomsten 1, 3 en 6 zijn. Dit zijn resultaten voor deze 50 herhalingen. Heb jij enig idee welke relatieve frequenties je in tabel 3 zou vinden bij oneindig veel herhalingen? Waarom?

Opdracht 4

Als je een gevonden resultaat bij de eerste trekking algemeen noteert door x_1 dan is het logisch dat je het bijhorende kansmodel noteert als X_1. Daarbij is X_1 de notatie voor het model dat zegt wat alle mogelijke uitkomsten en hun kansen zijn bij een eerste trekking. Kan jij nu X_1 voorstellen in tabelvorm?

Tabel 6

Opdracht 5

Je kan op dezelfde manier redeneren voor tabel 4. Daar gaat het over uitkomsten die je krijgt bij de tweede trekking. Wat je daar gevonden hebt stel je algemeen voor door x_2. Wat zijn hier alle mogelijke waarden die je kan hebben voor x_2 en wat zijn hun bijhorende kansen? Als je dat weet, dan ken je X_2, het kansmodel voor uitkomsten bij de tweede trekking. Stel nu ook X_2 voor in tabelvorm.

Tabel 7

Wat zouden bij oneindig veel herhalingen de relatieve frequenties van tabel 5 worden? Hoe geraak je daar aan de echte kansen?

Wat alle mogelijke uitkomsten zijn heb je reeds ontdekt. Herinner je dat de **volgorde** belangrijk is en dat **één** uitkomst van een steekproef van grootte $n = 2$ een **koppel** (x_1 , x_2) is.

Als je de eerste keer een 1 en de tweede keer een 6 gooit, dan is je uitkomst (x_1 , x_2) = (1 , 6). Wat is de kans dat zoiets gebeurt? Wat is de kans dat het model X_1 je de uitkomst 1 oplevert en dat tegelijkertijd het model X_2 je daarna een 6 oplevert? Zoiets schrijf je als $P(X_1 = 1, X_2 = 6)$. De komma betekent "en" zodat $P(X_1 = 1, X_2 = 6)$ hetzelfde is als $P(X_1 = 1 \text{ en } X_2 = 6)$ of voluit: de kans dat je de eerste keer een 1 **en** de tweede keer een 6 gooit.

Welk getal je de tweede keer gooit hangt niet af van het getal dat je de eerste keer vond zodat je de rekenregels voor "onafhankelijkheid" mag gebruiken: $P(X_1 = 1, X_2 = 6) = P(X_1 = 1) \cdot P(X_2 = 6)$.

Uit de kansmodellen voor de afzonderlijke trekkingen (tabel 6 en tabel 7) haal je dat $P(X_1 = 1) = \frac{3}{6}$ en $P(X_2 = 6) = \frac{1}{6}$ zodat $P(X_1 = 1, X_2 = 6) = \frac{3}{6} \cdot \frac{1}{6} = \frac{3}{36}$.

Opdracht 6

Nu je dit systeem kent kan je alle mogelijke uitkomsten (x_1 , x_2) met hun bijhorende kansen uitrekenen. Vul tabel 8 aan en controleer dat de som van de kansen gelijk is aan 1.

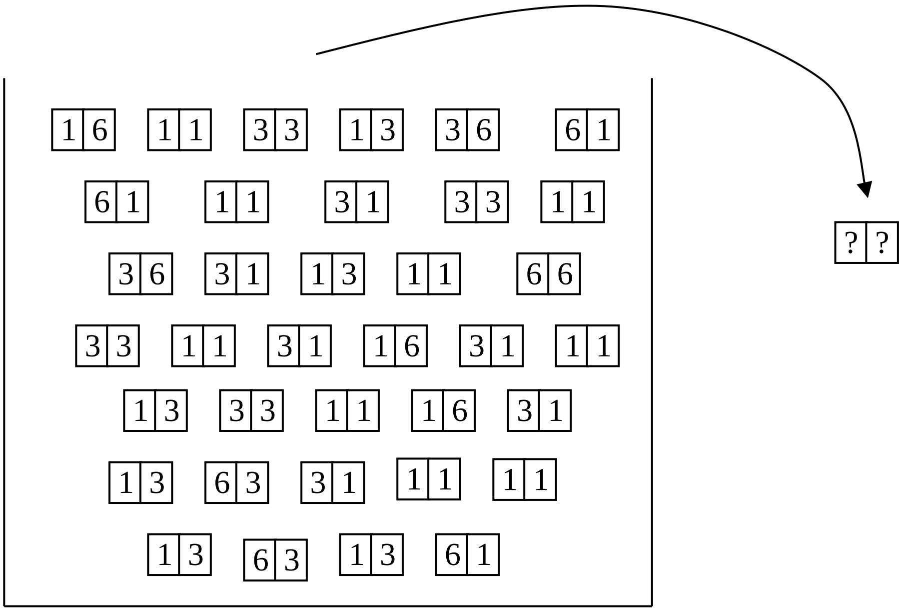
eerste trekking	$P(X_1 = x_1)$	tweede trekking	$P(X_2 = x_2)$	uitkomst (x_1 , x_2)	kans van deze uitkomst $P(X_1 = x_1 , X_2 = x_2)$
1	$P(X_1 = 1) = \frac{3}{6}$	1	$P(X_2 = 1) = \frac{3}{6}$		$P(X_1 = 1, X_2 = 1) =$
1	$P(X_1 = 1) = \frac{3}{6}$	3	$P(X_2 = 3) = \frac{2}{6}$		$P(X_1 = 1, X_2 = 3) =$
1	$P(X_1 = 1) = \frac{3}{6}$	6	$P(X_2 = 6) = \frac{1}{6}$		$P(X_1 = 1, X_2 = 6) =$
3	$P(X_1 = 3) = \frac{2}{6}$	1	$P(X_2 = 1) = \frac{3}{6}$		$P(X_1 = 3, X_2 = 1) =$
3	$P(X_1 = 3) = \frac{2}{6}$	3	$P(X_2 = 3) = \frac{2}{6}$		$P(X_1 = 3, X_2 = 3) =$
3	$P(X_1 = 3) = \frac{2}{6}$	6	$P(X_2 = 6) = \frac{1}{6}$		$P(X_1 = 3, X_2 = 6) =$
6	$P(X_1 = 6) = \frac{1}{6}$	1	$P(X_2 = 1) = \frac{3}{6}$		$P(X_1 = 6, X_2 = 1) =$
6	$P(X_1 = 6) = \frac{1}{6}$	3	$P(X_2 = 3) = \frac{2}{6}$		$P(X_1 = 6, X_2 = 3) =$
6	$P(X_1 = 6) = \frac{1}{6}$	6	$P(X_2 = 6) = \frac{1}{6}$		$P(X_1 = 6, X_2 = 6) =$

Tabel 8

Bemerk dat je hier 9 verschillende mogelijke uitkomstenkoppels hebt. Die hebben allemaal hun bijhorende kans.

Oplossing: zie laatste bladzijde.

Je kan het kansmodel voor deze steekproef van grootte $n = 2$ ook voorstellen door een vaas met kaartjes. Op elk kaartje staat een getallenkoppel, met twee getallen in de juiste volgorde. Als je bij de eerste trekking een 3 kan vinden en bij de tweede een 6 dan moet er in het vaasmodel zeker een kaartje van de vorm $\boxed{3 \mid 6}$ zitten.



Vaasmodel voor (X_1, X_2)

In deze vaas zitten 36 kaartjes. Twee van deze kaartjes zien eruit als $\boxed{3 \mid 6}$ wat betekent dat het koppel (3,6) met kans 2/36 voorkomt. Je hebt dus
een kans van 2/36 dat je bij een steekproef van grootte $n = 2$ eerst een 3 en dan een 6 zal trekken.

Opdracht 8

Het kansmodel voor (X_1, X_2) wordt soms in een tabel voorgesteld die er iets anders uitziet dan tabel 8. Begrijp je hoe tabel 9 gemaakt is? Wat staat er in de randen en wat staat er in het "midden"? Kan je hieruit het volledige kansmodel van (X_1, X_2) aflezen?

		Uitkomsten van X_2		
		1	3	6
Uitkomsten van X_1	1	$\frac{9}{36}$	$\frac{6}{36}$	$\frac{3}{36}$
	3	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$
	6	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Kansmodel voor (X_1, X_2)

Tabel 9

Nota. Ook uit een continue populatie kan je steekproeven trekken en ook daarvoor kan je kansmodellen opstellen. Als de populatie X gedefinieerd is door een dichtheidsfunctie $f(x)$ en je trekt hieruit een steekproef van grootte $n=2$ dan is het kansmodel voor (X_1, X_2) een dichtheidsfunctie in twee veranderlijken $f_{X_1,X_2}(x_1, x_2)$ die het gezamenlijk gedrag van het koppel (X_1, X_2) vastlegt. Kansmodellen voor steekproeven uit een continue populatie worden beschreven door functies in meerdere veranderlijken.

2.3. Kansmodel voor een steekproef van willekeurige grootte

Als voorbeeld kan je blijven denken aan de populatie X die zich gedraagt zoals de rode dobbelsteen.

Uit die populatie X ga je nu een steekproef van grootte n trekken. Je gooit een eerste keer met die rode dobbelsteen. Daarna een tweede, een derde, ….en tenslotte een n^{de} keer.

Je hebt nu 2 eigenschappen:

- Wat je bij elke afzonderlijke trekking zal vinden wordt beschreven door het kansmodel van de populatie waaruit je trekt. Dat betekent dat elke X_i zich gedraagt zoals de populatie X. Het is telkens hetzelfde model, met dezelfde uitkomsten en dezelfde bijhorende kansen.

- Wat je zal vinden bij de vierde trekking (of bij om het even welke trekking) hangt niet af van wat je gevonden hebt bij de andere trekkingen. Dus heb je "onafhankelijkheid".

Als je dat nu allemaal samen bekijkt, dan kan je over een steekproef spreken *in de voorwaardelijke wijs*. Wat zou ik allemaal kunnen uitkomen en met welke kansen als ik uit deze populatie X een steekproef van grootte n zou gaan trekken? Het model dat je hierop een antwoord geeft noteer je met hoofdletters.

$$\left(X_1 , X_2 ,.... X_i ,.... X_n\right)$$

kansmodel voor een steekproef
elke X_i gedraagt zich zoals de populatie X
en de X_i's zijn onafhankelijk van elkaar

Als je nu echt je steekproef trekt dan vind je bij de eerste trekking een getal. Dat getal is één van de mogelijke uitkomsten van het model X_1 en dat noteer je met een kleine letter x_1. De uitkomst na de tweede trekking noteer je met x_2, enz. De notatie met kleine letters betekent: dit zijn mijn toevallige resultaten die ik na het trekken van mijn steekproef gevonden heb.

$$\left(x_1 , x_2 ,.....x_i ,..... x_n\right)$$

het resultaat dat ik in mijn steekproef gevonden heb

3. Schematisch overzicht

In het algemeen	**Met een voorbeeld**
Een **populatie** X "ken" je volledig als je het kansmodel kent	Kansmodel van de populatie X (zie tabel 1)
↓	↓
Het kansmodel van een **steekproef** $(X_1, X_2, \ldots X_n)$ kan je afleiden uit het kansmodel van de populatie X	Kansmodel voor (X_1, X_2) (zie tabel 2)

Tabel 1:

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

Kansmodel van de populatie X

Tabel 2:

		Uitkomsten van X_2		
		1	3	6
Uitkomsten van X_1	1	$\frac{9}{36}$	$\frac{6}{36}$	$\frac{3}{36}$
	3	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$
	6	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Kansmodel voor (X_1, X_2)