STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Steekproefmodellen en normaal verdeelde steekproefgrootheden

6. Proporties

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Proporties

1. Een nieuwe naam voor een gekende grootheid

In de krant lees je dat er bij de geboorten in Vlaanderen 48.7 % meisjes zijn, dat 54 % van de allochtone leerlingen thuis geen Nederlands spreekt, en dat 10 % van wie hersenvliesontsteking (meningitis) krijgt eraan sterft.

Heel veel informatie komt tot jou in de vorm van verhoudingen of proporties. We gaan daar nu uitgebreid aandacht aan besteden.

Werken met proporties is eigenlijk eenvoudig. Je kijkt naar een bepaalde eigenschap en de enige vraag die je dan stelt is: “Heeft iemand (of iets) die eigenschap, ja of neen?”. Die vraag herhaal je bij elk element van de groep die je bestudeert. Zo vind je de proportie van die groep die de eigenschap heeft. Als er bijvoorbeeld in een groep van 2000 baby’s 974 meisjes zijn dan is de proportie meisjes in die groep gelijk aan $\frac{974}{2000} = 0.487$ wat ook gelijk is aan 48.7 %.

In de statistiek gebruik je als klassieke benaming het woord “succes” als de eigenschap er wel is en het woord “mislukking” als die eigenschap er niet is. De woorden “succes” en “mislukking” mag je hierbij niet interpreteren als iets wat goed of slecht is. Bij een onderzoek naar de proportie meisjes bij de geboorte van kinderen spreek je over “succes” als de baby een meisje is en over “mislukking” als het een jongen is.

Om over te stappen van woorden op getallen vervang je “succes” door het cijfer 1 en “mislukking” door het cijfer 0. Dat is gemakkelijk te onthouden, want 1 betekent “de eigenschap WEL hebben” en 0 betekent “de eigenschap NIET hebben”.

> Dat je met de cijfers 0 en 1 werkt is niet zomaar een willekeurige keuze. *Zij zorgen ervoor dat je een proportie kan uitrekenen juist zoals je een gemiddelde berekent.* Dat zie je in volgend voorbeeld.

Aan een groep van 12 allochtone leerlingen vraag je of zij thuis een andere taal dan het Nederlands spreken. Bij 7 is het antwoord “ja” en 5 zeggen “neen”. Voor deze groep is de proportie leerlingen die thuis een andere taal spreekt gelijk aan $\frac{7}{12} \cong 0.58$ wat je ook kan schrijven als 58 %.

Om deze proportie te berekenen heb je vooraf het aantal “ja” antwoorden samengeteld. Dat hoefde niet. Je kan gewoon bij elke leerling noteren wat het antwoord is. Als de eerste leerling “ja” zegt schrijf je 1. Zegt de tweede ook “ja” dan schrijf je terug 1. Zegt de derde “neen” dan schrijf je 0 enz.. Het resultaat van die 12 leerlingen zou dus als volgt kunnen zijn: 1 1 0 1 0 0 0 1 0 1 1 1 .

Wat gebeurt er nu als je de som maakt van al die enen en nullen? Je krijgt dan gewoon het totaal aantal enen, want nullen erbij tellen verandert toch niets.

som = aantal enen = aantal dat de eigenschap wel bezit

Als je nu het "aantal dat de eigenschap wel bezit" deelt door het totale aantal dan krijg je de *proportie* die de eigenschap *wel* bezit. Als formule is dat juist hetzelfde als het gemiddelde van getallen.

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{\textit{som van alle getallen}}{\textit{totaal aantal getallen}} = \frac{\textit{aantal enen}}{\textit{totaal aantal}} \quad = \textit{proportie die de eigenschap wel bezit}$$

Juist zoals bij het gemiddelde kan je nu ook de proportie op verschillende manieren bekijken. Je kan kijken naar de proportie in een totale populatie of naar een proportie in jouw steekproef of zelfs naar een kansmodel waarbij je vooraf zegt welke proportie je in je steekproef kan vinden en met welke kans. Dit leer je in de volgende paragrafen.

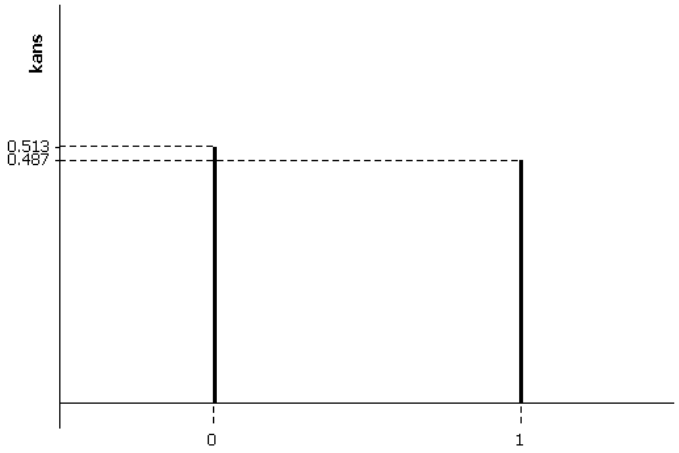
2. De populatieproportie

In 2003 zijn er in Vlaanderen meer dan 60 000 kinderen geboren en daarvan waren er 48.7 % meisjes. Deze uitspraak gaat over de totale populatie van alle kinderen die in 2003 in Vlaanderen geboren werden. Als eigenschap neem je "meisje zijn" zodat je met een "neen" of een "ja" kan antwoorden op de vraag: "Is de baby een meisje?" Op die manier heb je hier te maken met een 0–1 populatie. Zo'n 0–1 populatie wordt ook een Bernoulli populatie genoemd en je kan ze op verschillende manieren voorstellen

2.1. Bernoulli of 0–1 populatie

Als je werkt met een vaasmodel dan kan je een vaas vullen met heel veel kaartjes waarop een 1 staat voor een meisje en een 0 voor een jongen. Je moet er dan voor zorgen dat je de juiste verhouding gebruikt, bijvoorbeeld 3000 kaartjes waarbij er op 1461 een 1 staat (voor een meisje) en op de andere 1539 kaartjes een 0 (voor een jongen).

Je kan ook een staafdiagram gebruiken om de populatie van deze baby's voor te stellen. Dat ziet er zo uit.



Kansmodel voor de 0–1 populatie X met succeskans 0.487
(0=jongen, 1=meisje)

Je kan de populatie ook met een tabel voorstellen. De kans op succes (en dat is in dit voorbeeld een meisje, voorgesteld door het cijfer 1) is hier gelijk aan 0.487.

x	0	1
$P(X=x)$	0.513	0.487

Kansmodel voor de 0–1 populatie X met succeskans 0.487
(0=jongen, 1=meisje)

Tabel 1

Een andere 0–1 populatie ziet er als volgt uit. Bemerk dat de kans op "succes" hier gelijk is aan 3 %.

x	0	1
$P(X=x)$	0.97	0.03

Kansmodel voor de 0–1 populatie X met succeskans 0.03
(0=mislukking, 1=succes)

Tabel 2

Over het algemeen stel je bij een 0–1 **populatie** de kans op succes voor door π (de Griekse letter pi). De kans op mislukking is dan gelijk aan $1-\pi$. Een algemeen kansmodel voor een 0 – 1 populatie ziet er als volgt uit.

x	0	1
$P(X=x)$	$1-\pi$	π

Kansmodel voor een 0–1 populatie X met succeskans π
(0 = mislukking, 1 = succes)

Tabel 3

Zoals vroeger μ en σ noem je ook π een **populatieparameter**. Het is een vast getal dat een bepaalde eigenschap van de populatie beschrijft. De letter π is eigenlijk een Griekse letter "p" en dat is de eerste letter van het woord proportie. De letter π duidt aan dat **de proportie** meisjes in de **totale populatie** gelijk is aan 0.487. Dat zijn 487 meisjes per duizend baby's. Dit betekent ook dat, als je lukraak een baby uit die databank van meer dan 60 000 baby's trekt, de **kans op succes** (kans op een meisje) gelijk is aan 0.487.

In het algemeen stelt π **de proportie successen in de populatie** voor, wat ook de **populatieproportie** wordt genoemd. π is ook de **succeskans** die je hebt als je uit die populatie trekt. Die succeskans kan je ook schrijven als $P(X=1)$ want succes wordt door het cijfer 1 voorgesteld. Bemerk dat in deze context π niets te maken heeft met het getal 3.14159…

2.2. Kenmerken van een 0–1 populatie

Voor elk kansmodel kan je het gemiddelde en de standaardafwijking berekenen. Een 0–1 populatie heeft maar twee mogelijke uitkomsten, namelijk 0 en 1. Het is een discreet kansmodel. Je moet dus de formules voor discrete kansmodellen gebruiken.

Het gemiddelde van een discreet kansmodel ken je al. Maak de gewogen som van de uitkomsten.

Voor de baby's van 2003 volgt uit tabel 1 dat

$$E(X) = 0 \cdot (0.513) + 1 \cdot (0.487) = 0.487$$

x	0	1
$P(X=x)$	0.513	0.487

en voor de algemene 0–1 populatie X van tabel 3 vind je

$$E(X) = 0 \cdot (1-\pi) + 1 \cdot \pi = \pi$$

x	0	1
$P(X=x)$	$1-\pi$	π

Het gemiddelde van een 0–1 populatie is gelijk aan haar succeskans π. Het is dan ook logisch dat je voor 0–1 populaties de Griekse letter π gebruikt om het populatiegemiddelde aan te duiden en niet de Griekse letter μ.

Zodra je het gemiddelde kent kan je ook de spreiding rond dit gemiddelde kenmerken. Dat doe je met de standaardafwijking. Je gebruikt daarbij de algemene formule voor discrete kansmodellen. Zo krijg je voor de 0–1 populatie van tabel 3 dat

$$\text{var}(X) = (0-\pi)^2 \cdot (1-\pi) + (1-\pi)^2 \cdot \pi = \pi(1-\pi)$$

zodat

$$sd(X) = \sqrt{\pi(1-\pi)}\,.$$

Voor de 0–1 populatie van de baby's van het jaar 2003 (tabel 1) is de standaardafwijking gelijk aan $sd(X) = \sqrt{\pi(1-\pi)} = \sqrt{0.487(1-0.487)} \cong 0.50$.

Een 0–1 populatie X met succeskans π heeft:

- als kansverdeling:

x	0	1
$P(X=x)$	$1-\pi$	π

- als gemiddelde: $E(X)=\pi$
- als standaardafwijking: $sd(X)=\sqrt{\pi(1-\pi)}$

Opdracht 1

Hoe groot is de standaardafwijking van de 0–1 populatie in tabel 2?

x	0	1
$P(X=x)$	0.97	0.03

3. De steekproefproportie

Als je een steekproef trekt uit een 0–1 populatie dan heb je een steekproefresultaat (x_1 , x_2 , …..x_i ,….., x_n) waarbij elke gevonden x ofwel een 0 ofwel een 1 is. Als je dan van al die getallen het gemiddelde maakt dan vind je:

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{som}{totaal\ aantal} = \frac{aantal\ enen}{totaal\ aantal}$$

$$= proportie\ elementen\ in\ de\ steekproef\ \ die\ de\ eigenschap\ wel\ bezit$$
$$= proportie\ successen\ in\ de\ steekproef$$

De proportie successen die jij in je steekproef vindt, stel je voor door een kleine letter p met een hoedje erop $\hat{p}$. Je spreekt dat uit als “p–hoed” en de formule is $\hat{p}=\frac{1}{n}\sum_{i=1}^{n} x_i$.

Bij een steekproef van grootte $n = 5$ zou je bijvoorbeeld (1, 0, 1, 0, 0) kunnen gevonden hebben. Dan is voor jou $\hat{p} = \frac{2}{5} = 0.4$. Maar je had natuurlijk ook iets helemaal anders kunnen vinden zoals (1, 0, 1, 1, 1) . Dan zou jij $\hat{p} = \frac{4}{5} = 0.8$ gevonden hebben.

Nu komt de klassieke vraag. Kan jij vooraf zeggen wat je als steekproefproportie **zou** vinden als je uit een 0–1 populatie **zou** trekken? Met het woord "steekproefproportie" wordt "de proportie successen in de steekproef" bedoeld.

Je weet dat je op deze vraag alleen maar kan antwoorden met een kansmodel. Je hebt hier het kansmodel van de steekproefproportie nodig, en dat stel je voor met een hoofdletter, dus met $\hat{P}$.

Als $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ dan is $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$. De formule $\frac{1}{n}\sum_{i=1}^{n} X_i$ herken je. Het is juist dezelfde formule die je gebruikt hebt voor het kansmodel van het steekproefgemiddelde. **Alle eigenschappen die je daar geleerd hebt blijven ook hier gelden. Je kan ze gewoon herhalen, met de aangepaste benaming**.

3.1. Het gemiddelde van de steekproefproportie

Werk met een 0–1 populatie X waarbij de kans op succes gelijk is aan π, of anders gezegd, waarbij de proportie successen in de populatie gelijk is aan π. Trek uit zo'n populatie een steekproef van grootte n en bereken de proportie successen in jouw steekproef. Als je dat heel veel keren zou herhalen dan zou het gemiddelde van al die gevonden steekproefproporties ("in the long run") samenvallen met de populatieproportie.

> Voor een steekproef $(X_1, X_2, \ldots X_n)$ uit een 0–1 populatie X met succeskans π geldt:
>
> het gemiddelde van de steekproefproportie is gelijk aan de populatieproportie
>
> $$E\left(\hat{P}\right) = \pi$$

Opdracht 2

Welke formule voor het steekproefgemiddelde heb je hier vertaald?

3.2. De standaardfout van de steekproefproportie

Herinner je dat voor een 0–1 populatie X de standaardafwijking gelijk is aan $sd(X) = \sqrt{\pi(1-\pi)}$. Als je dit toepast op de algemene eigenschap van het steekproefgemiddelde dan krijg je het volgende.

> Voor een steekproef $(X_1, X_2, \ldots X_n)$ uit een 0–1 populatie X met succeskans π geldt:
>
> de standaardfout van de steekproefproportie is gelijk aan
> de standaardafwijking van de populatie gedeeld door de wortel uit de steekproefgrootte
>
> $$se\left(\hat{P}\right) = \frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}$$

Opdracht 3

Welke formule voor het steekproefgemiddelde heb je hier vertaald?

3.3. De vorm van het kansmodel

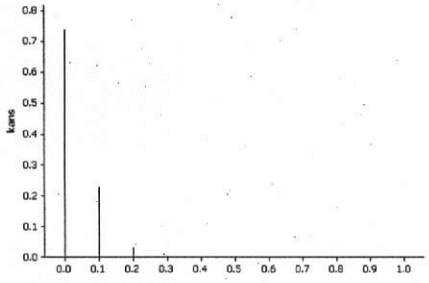
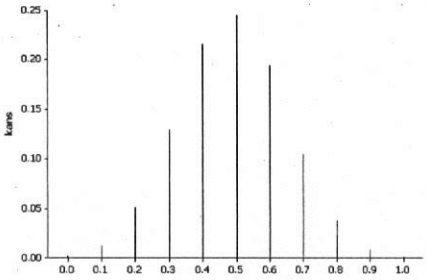
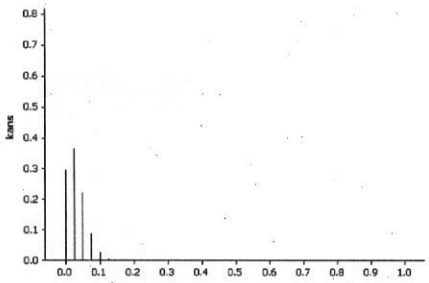
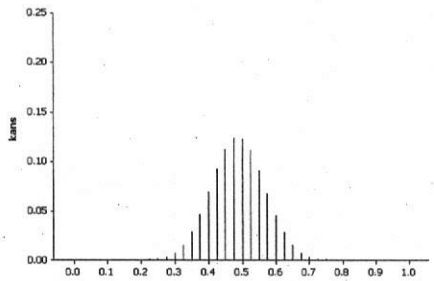
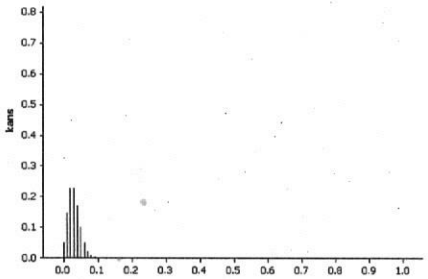
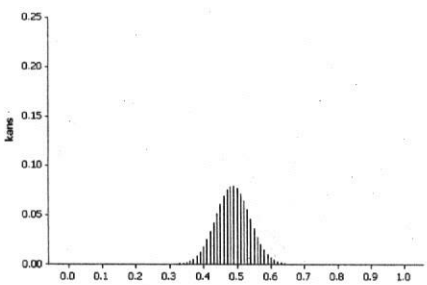
Juist zoals bij het kansmodel van het steekproefgemiddelde $\overline{X}$ is ook hier de globale vorm van het kansmodel van de steekproefproportie $\hat{P}$

- **WEL** afhankelijk van de populatie waaruit je trekt
- **WEL** afhankelijk van de grootte van de steekproef.

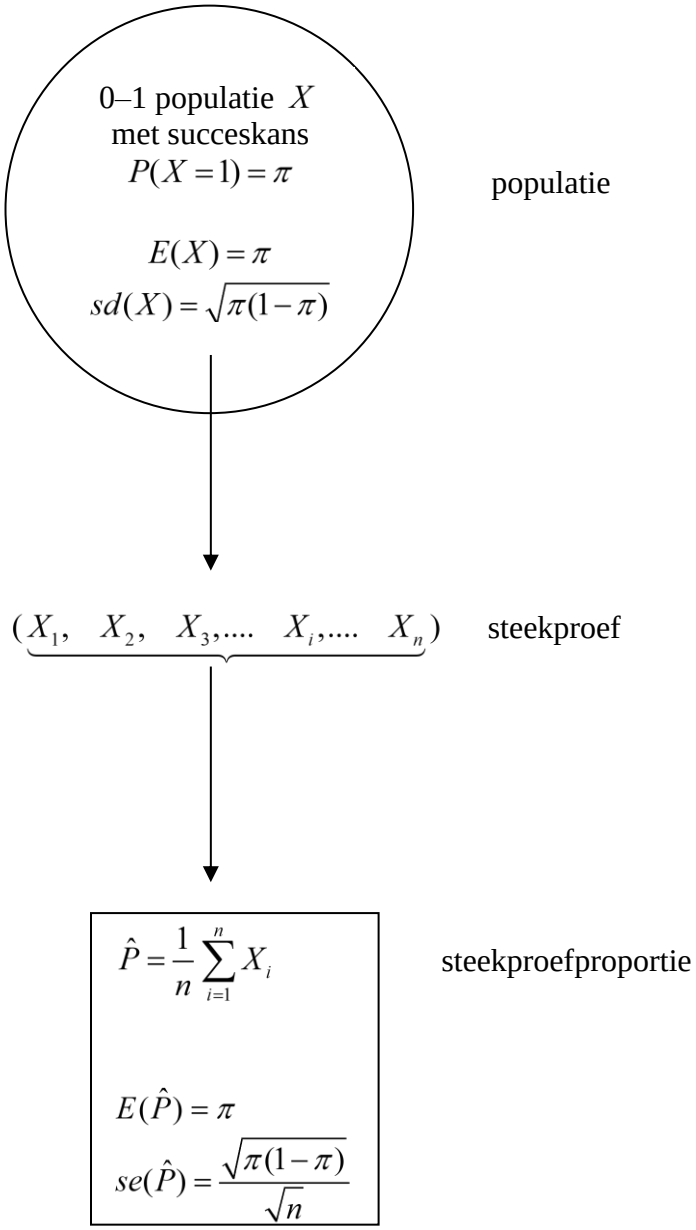
In de volgende figuur zie je in de linkerkolom het kansmodel van de steekproefproportie $\hat{P}$ wanneer je een steekproef trekt uit een 0–1 populatie waarbij de kans op succes gelijk is aan 0.03. Bij een steekproef van grootte $n = 10$ en zelfs bij $n = 40$ lijkt de globale vorm niet op een curve die symmetrisch is rond één top. Voor $n = 100$ is dit (benaderend) wel het geval.

In de rechterkolom zie je het kansmodel van $\hat{P}$ wanneer je trekt uit een 0–1 populatie met succeskans 0.487. Bij $n = 100$ zie je een zo goed als perfecte globale klokvorm, maar ook reeds bij $n = 40$ en zelfs bij $n = 10$ zie je figuren die behoorlijk symmetrisch rond één top zijn.

Het kansmodel van de steekproefproportie $\hat{P}$
bij een steekproef van grootte n uit een 0–1 populatie X

	Kansmodel van $\hat{P}$ bij een steekproef uit de 0–1 populatie met $\pi = 0.03$	Kansmodel van $\hat{P}$ bij een steekproef uit de 0–1 populatie met $\pi = 0.487$
$n = 10$		
$n = 40$		
$n = 100$		

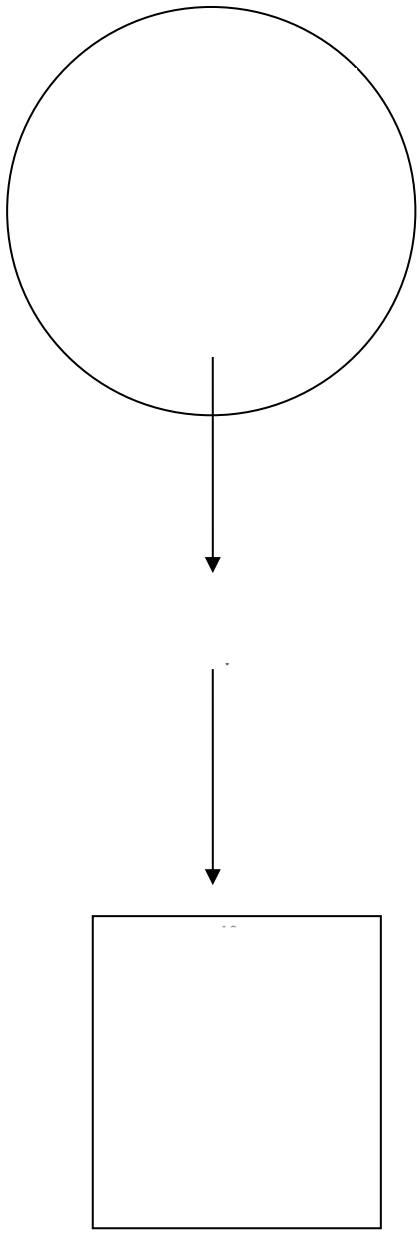
3.4. Overzicht



Opdracht 4

Bovenstaande figuur toont hoe je, uit de kennis van de 0–1 populatie, eigenschappen kan afleiden voor de steekproefproportie. Teken nu zelf ook zo'n figuur waarbij je voor de grootheden die je kent hun waarde invult wanneer je te maken hebt met een steekproef van grootte $n = 10$ uit de 0–1 populatie van tabel 1.

x	0	1
$P(X=x)$	0.513	0.487



4. Een kansmodel benaderen met een ander kansmodel

Soms is het handig om een kansmodel te kunnen vervangen door een ander kansmodel. Dat doe je natuurlijk alleen maar als dat andere kansmodel aan twee voorwaarden voldoet. Het moet het oorspronkelijke kansmodel goed benaderen en het moet tegelijkertijd eenvoudig te gebruiken zijn.

Een kansmodel dat je intussen goed kent is de normale. Onder bepaalde voorwaarden kan je de normale gebruiken als benaderend kansmodel voor de steekproefproportie $\hat{P}$. Je spreekt dan over "de normale benadering".

4.1. De normale benadering

Dit weet je. Met de normale dichtheidsfunctie zoek je de kans om in een bepaald **gebied** (in een bepaald deelinterval) terecht te komen en NIET om op één welbepaalde x-waarde terecht te komen. Je berekent een kans als een **oppervlakte** onder de curve en NIET als een hoogte (functiewaarde).

Dit weet je ook. De mogelijke uitkomsten van de steekproefproportie vormen geen aaneengesloten continuüm maar nemen discrete waarden aan. Bij een steekproef van grootte $n=10$ bijvoorbeeld kan je 6 successen hebben en dan heb jij 0.6 gevonden als je steekproefproportie. Je kan ook 7 successen hebben en dan is jouw steekproefproportie gelijk aan 0.7. Maar je kan niet zoiets als 6.12 successen hebben bij een steekproef van grootte $n=10$ en dus kan je 0.612 niet als steekproefproportie uitkomen (en ook geen enkel ander getal tussen 0.6 en 0.7 bij $n=10$). Bij elke discrete uitkomst hoort een kans zodat je hier het kansmodel kan tekenen als een staafdiagram. Daar geeft de **hoogte** van de staafjes de kans aan.

Hoe kan je nu die twee tegenstrijdige systemen met elkaar verzoenen?

De oplossing is: werk met intervallen. Hoe je dat doet met de normale is duidelijk, want die is continu. Voor de steekproefproportie gaat dat ook, als je maar goed begrijpt wat er bedoeld wordt.

Bij een steekproef van grootte $n=10$ kan je de vraag stellen naar de kans dat de steekproefproportie in het interval [0.4 ; 0.7] terechtkomt. Dat betekent niet dat de mogelijke uitkomsten van de steekproefproportie nu plots continu geworden zijn. Neen, zij zijn discreet. Maar ook dan kan je kijken naar alle mogelijke discrete waarden die in het interval [0.4 ; 0.7] liggen. Voor al die waarden bereken je de kansen en die tel je samen. In formulevorm ziet dat er als volgt uit:

$$P(0.4 \le \hat{P} \le 0.7) = P(\hat{P} = 0.4) + P(\hat{P} = 0.5) + P(\hat{P} = 0.6) + P(\hat{P} = 0.7)$$

De normale benadering zegt dat je $P(0.4 \le \hat{P} \le 0.7)$ goed kan benaderen zonder eerst al die afzonderlijke kansen te berekenen. Hoe dat werkt leer je in de volgende paragraaf.

4.2. Criterium voor de normale benadering

Je mag de normale benadering gebruiken om de kans te berekenen dat de steekproefproportie in **intervallen** valt van zodra het verwachte **aantal** successen en het verwachte **aantal** mislukkingen beide minstens gelijk zijn aan 15.

In formulevorm betekent dit dat bij een bepaalde waarde van π je ervoor moet zorgen dat de steekproefgrootte n zo groot is dat er gelijktijdig voldaan is aan

$$\begin{cases} n\,\pi \geq 15 \\ n\,(1-\pi) \geq 15 \end{cases}$$

Voor $\pi = 0.487$ bijvoorbeeld betekent dit:

$$\begin{cases} n\,\pi \geq 15 \;\rightarrow\; n\,(0.487) \geq 15 \;\rightarrow\; n \geq 30.8 \;\rightarrow\; n \geq 31 \\ n\,(1-\pi) \geq 15 \;\rightarrow\; n\,(1-0.487) \geq 15 \;\rightarrow\; n \geq 29.2 \;\rightarrow\; n \geq 30 \end{cases}$$

zodat je moet werken met een steekproefgrootte die minstens gelijk is aan 31.

Nota.
In de meeste teksten over statistiek lees je dat je reeds met de normale benadering mag werken zodra $n\pi$ en $n(1-\pi)$ "minstens 10" zijn. Soms wordt "minstens 5" als criterium genomen. Recent wetenschappelijk onderzoek heeft aangetoond dat in bepaalde situaties deze grenzen te klein zijn en dat het veiliger is om te werken met "minstens 15".

Als de steekproef groot genoeg is dan mag je dus met de steekproefproportie $\hat{P}$ werken zoals met een normaal verdeelde grootheid.

De volgende kansuitspraak ken je: "Elk normaal model X valt met 95 % kans niet verder van zijn gemiddelde dan 1.96 standaardafwijkingen".

$$P\left(gemiddelde - 1.96\ standaardafwijking \leq X \leq gemiddelde + 1.96\ standaardafwijking\right) = 0.95$$

De normale benadering zegt dat je de steekproefproportie $\hat{P}$ ook mag beschouwen als een normaal model. Voor dit model is het gemiddelde $E(\hat{P}) = \pi$ en de standaardafwijking $se(\hat{P}) = \dfrac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}$.

Voor de steekproefproportie $\hat{P}$ krijg je dan:

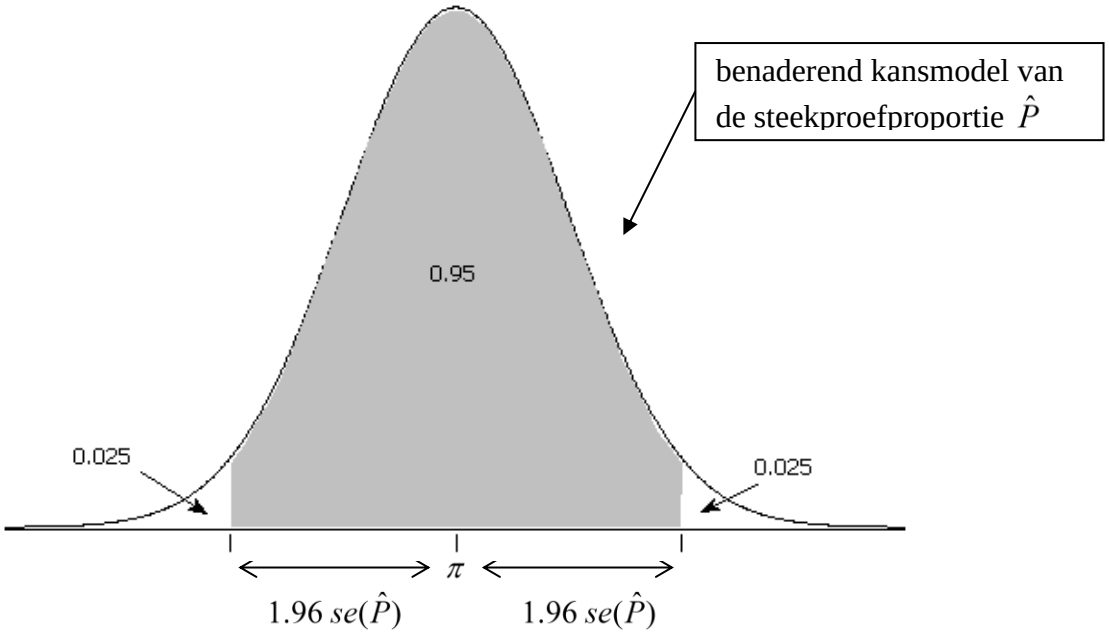
$$P\left(gemiddelde - 1.96\ standaardfout \le \hat{P} \le gemiddelde + 1.96\ standaardfout\right) = 0.95$$

of

$$P\left(E(\hat{P}) - 1.96\ se(\hat{P}) \ \le\ \hat{P}\ \le E(\hat{P}) + 1.96\ se(\hat{P})\right) = 0.95$$

of voluit

$$P\left(\pi - 1.96\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}} \le\ \hat{P}\ \le \pi + 1.96\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}\right) = 0.95$$



De steekproefproportie $\hat{P}$ levert resultaten die rond haar gemiddelde $E(\hat{P})$ terechtkomen. Je weet ook dat het gemiddelde van de steekproefproportie gelijk is aan de populatieproportie π. Dus kan je evengoed zeggen dat de steekproefproportie $\hat{P}$ rond de populatieproportie π terechtkomt. Met kans 95 % komt $\hat{P}$ terecht in het interval $\left[\pi - 1.96\ se(\hat{P})\ ;\ \pi + 1.96\ se(\hat{P})\right]$. Je kan dit interval ook verkort noteren als $\pi \pm 1.96\ se(\hat{P})$.

Opdracht 5

Als $\pi = 0.487$ en $n = 100$ dan komt de steekproefproportie $\hat{P}$ met 95 % kans in het interval [0.389 ; 0.585] terecht. Dat betekent dat jij met 95 % kans minstens 39 en hoogstens 58 meisjes in je steekproef zal vinden. Reken dat zelf maar eens na en motiveer je manier van werken.