

De vele werelden van de statistiek

Prof. dr. Herman Callaert, Centrum voor Statistiek, Universiteit Hasselt.

- **De wereld van het “ideale model” (= de populatie) om te beschrijven op welke manier uitkomsten van een experiment tot jou komen**

Deze wereld wordt beschreven met de taal van de wiskunde (een kansverdeling of een dichtheidsfunctie). Welk “ideaal model” het beste past bij een specifieke populatie is onderwerp van multidisciplinair onderzoek, onderbouwd met methoden uit de statistiek.

- **De wereld van het “concrete cijfermateriaal” (=de steekproefresultaten) dat je bekomt na het uitvoeren van een experiment**

Deze wereld kan vanuit verschillende standpunten worden onderzocht. De beschrijvende statistiek geeft een overzichtelijke weergave van het cijfermateriaal met de bedoeling informatie te winnen uit al die getallen. Of men kan ook exploratief op zoek gaan naar globale kenmerken met de bedoeling een idee te krijgen over de totale populatie waaruit de steekproefresultaten afkomstig zijn.

- **De wereld van het “ideale model” voor het gedrag van steekproeven en van grootheden die op steekproeven gebaseerd zijn**

Deze wereld vormt het fundament van de verklarende statistiek. Hierbij wordt een model gemaakt dat op een formele wiskundige manier de regulariteit van de toevalligheid beschrijft. Dit is de enige basis die de verklarende statistiek heeft om haar uitspraken wetenschappelijk te onderbouwen.

Notatie

“steekproefresultaten” **kleine letters** x_1, x_2, x_3, \dots

“grootheden gebaseerd op steekproefresultaten” **kleine letters** \bar{x}, S

“een populatie (is een ideaal model)” **hoofdletters** X

“een ideaal model voor steekproefresultaten” **hoofdletters** X_1, X_2, X_3, \dots

“een ideaal model voor grootheden gebaseerd op steekproefresultaten” **hoofdletters** \bar{X}, S

“eigenschappen van een populatie” **Griekse letters** μ, σ

“eigenschappen van een ideaal model voor grootheden gebaseerd op steekproefresultaten” (*)*(laatste blz)*

=====
 Hieronder doe ik een poging om het voorgaande wat te verduidelijken met een voorbeeld dat veel te beperkt is om als “echt voorbeeld” te dienen, maar waarbij ik probeer om de achtergrond van de gebruikte notatie te verklaren (dit is dus geen tekst die rechtstreeks geschikt is voor leerlingen).
 =====

Als start is er reeds het probleem om een “ideaal model” (een stochastische veranderlijke) te beschrijven. Als aan de universiteit een cursus statistiek in het eerste jaar gedoceerd zal worden zonder eerst formele kansentheorie te behandelen (wordt verschoven naar het tweede jaar), dan is het zeker niet aangewezen om in het SO een stochastische veranderlijke te definiëren op een formele manier. Er is dus geen nood aan gestructureerde uitkomstenruimten met bijhorende sigma-algebra en kansmaat, en ik denk zelfs dat er geen nood is aan een functie X van “omega” naar de reële getallen (je gebruikt daar een stukje van een formele definitie van een kansruimte, maar daarna moet je jezelf toch beperken). Bovendien heb ik ervaren dat er leerkrachten zijn waarvoor de begrippen “een uitkomstenruimte omega”, “een stochast”, “de functie X ” en “een kansverdeling” meer voor verwarring dan voor verheldering zorgen. Temeer als je je daarbij nog afvraagt: wat doe je daar dan mee in je verdere lessen?

Hoe kan het dan wel, als je met bovenstaande beperkingen moet leren leven?

Een mogelijk voorstel is dat je totaal van denkkader verandert, en dat je niets meer uit de formele kansruimten gebruikt.

Je begint met af te spreken dat je, voor de wiskundige behandeling van problemen, “uitkomsten van een experiment” altijd eerst zal vertalen in “cijfermateriaal” (= getallen). Dit kan meestal op een zeer natuurlijke wijze (lengte, gewicht, aantal “ogen” op een dobbelsteen), of soms doe je het bij conventie (nul en één voor munt en kruis). **Het is duidelijk dat bij de bespreking van het experiment (vóór, tijdens en na de analyse), de rol van de “context”, en dus ondermeer ook “eenheden”, “manier van opmeten”, enz., even belangrijk is al de rol van “wiskunde”, maar in deze tekst waarbij ik de notatie-afpraak probeer te verduidelijken, ga ik dit niet telkens herhalen.**

De afspraak dat “uitkomsten” (eventueel tijdelijk) met “getallen” worden geassocieerd kan op een natuurlijke manier “in woorden” worden geformuleerd. Hiervoor is geen functievoorschrift nodig.

Dan komt het nieuwe denkkader.

Aangezien het toeval in elk statistisch experiment aanwezig is, komen de getallen op een toevallige manier tot mij. Elke particuliere uitkomst is een toevallig getal, getrokken uit een populatie, en wat het zal zijn weet ik niet vooraf. Maar er is iets wat ik wel weet. Op basis van heel veel trekkingen zal ik een patroon zien verschijnen (relatieve frequentie als “goede” benadering van het begrip kans). Dat patroon is stabiel (of nadert naar stabiliteit als je dat probeert te verduidelijken met simulatiestudies). Er is dus regulariteit “in the long run”. **Deze “regulariteit van het toeval” formaliseren in een wiskundige taal en die dan verder bestuderen in het kader van een “context”, dat doet de statistiek.**

Dus, vanaf nu bestudeer ik het **toevalsmechanisme** dat er voor verantwoordelijk is dat “deze” getallen op “deze toevallige manier” tot mij komen. Als ik dit mechanisme ken, dan **ken ik alles** (dit is een uitspraak in de denkwereld van de statistiek: alles kennen betekent “het volledige toevalsmechanisme” kennen, het betekent niet: weten welk getal ik vandaag zal vinden als steekproefresultaat).

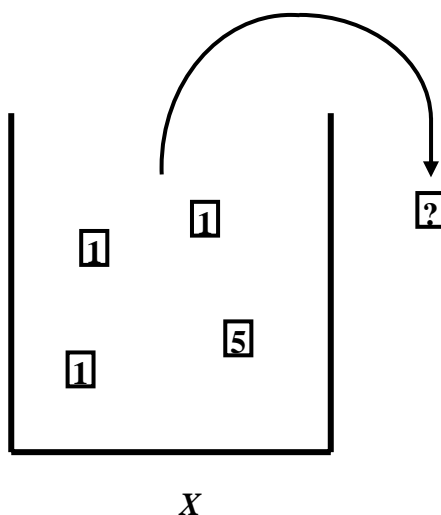
Wat we nu nodig hebben is een formele beschrijving van dit toevalsmechanisme, en een geëigende notatie.

Vanaf nu werk ik met een zeer beperkt voorbeeld voor een situatie met discrete uitkomsten. Het continue geval werkt met identiek hetzelfde denkkader (toevalsmechanisme) maar daar gaat het over intervallen en bijhorende oppervlakte onder curven.

Het kan nuttig zijn om op verschillende manieren een toevalsmechanisme voor te stellen. In het begin is een grafische voorstelling te verkiezen (of kan zij minstens dienen als een extra steun om een “tabelvorm” of een “formulevorm” beter te begrijpen).

Het toevalsmechanisme (in zijn totaliteit) stellen we voor door een hoofdletter X .

Grafische voorstelling: **het vaasmodel**



Bovenstaande figuur stelt het toevalsmechanisme voor waarmee getallen tot mij komen wanneer ik lukraak kaartjes trek uit deze vaas. De notatie X staat dus niet voor het vraagteken in de figuur, als je daarmee bedoelt: wat staat er op dat kaartje? X staat voor de volledige figuur, die aangeeft “welke getallen te verwachten zijn, en met welke kans”, wanneer je uit deze vaas “**een kaartje zou gaan trekken**” (formuleer uitspraken over toevalsmechanismen in de *voorwaardelijke wijs* – dat helpt om antwoorden te krijgen die het hele proces beschrijven, en niet de toevallige uitkomst op jouw getrokken kaartje). Bemerkt ook dat een vaasmodel niet zomaar de uitkomstenverzameling geeft (dat is $\{1,5\}$), maar een gelijktijdige voorstelling is van “wat zijn alle mogelijke uitkomsten” en “met welke kans komen deze uitkomsten naar mij toe”

De uitspraak $P(X = 5)$ kan je lezen als: de kans dat het toevalsmechanisme (zoals éénduidig gedefinieerd door X) het getal 5 oplevert. Of je kan ook spreken over de kans om een 5 te hebben als je lukraak een kaart zou gaan trekken uit die welbepaalde vaas. En uiteindelijk kan je ook zeggen dat dit de kans is om een 5 te vinden als je lukraak zou trekken uit de populatie die vastgelegd is door X .

Zodra een leerling goed begrijpt dat er een verschil is tussen een toevallig observatiegetal en het achterliggende mechanisme, is de belangrijke stap gezet om met verklarende statistiek te gaan werken. Meer is niet nodig (maar onderschat de moeilijkheidsgraad niet om telkens terug in “toevalsmechanismen” of in “achterliggende modellen” te denken en te redeneren). Veel voorbeelden zijn hier zeer nuttig.

Nota. Als we aan concrete populaties denken dan stellen we ons meestal “grote” populaties voor waaruit later een “kleine” steekproef wordt getrokken. Daarom kan het in het begin interessant zijn om ook zo te starten. Maak dan een vaas met 300 000 kaartjes waar een één op staat en 100 000 kaartjes met een 5 erop. En vertel dan dat die heel goed door elkaar geschud worden en dat er dan lukraak wordt uit getrokken. Wat je dan bemerkt is dat de regulariteit van het toevalsmechanisme geen gebruik maakt van “aantallen” maar van “structuur”. Alles wat je moet weten is “welke getallen” en “met welke kans”.

Voorstelling in **tabelvorm** (kansverdeling).

Hetzelfde toevalsmechanisme kan je ook voorstellen (discrete eindige populatie) in tabelvorm. In zo’n tabel geef je terug twee dingen tegelijk aan: wat zijn de mogelijke uitkomsten en wat zijn de bijhorende kansen. En het geheel van deze informatie noteer je met de hoofdletter X .

Een specifieke uitkomst van een toevalsmechanisme X noteer je door de corresponderende kleine letter x .

Een tabel is voor bepaalde leerlingen waarschijnlijk reeds moeilijker om zich een concreet idee te vormen over een “achterliggend ideaal model”.

het “ideale model” X	de uitkomsten van het model X	x	1	5
	de bijhorende kansen voor dit model X	$P(X=x)$	$\frac{3}{4}$	$\frac{1}{4}$

Op het gepaste ogenblik komt ook de voorstelling **in formulevorm** (zoals Binomiale of Poisson), maar voer die slechts in als je ze ook echt gebruikt. Het enige wat je hier nodig hebt is de definitie van het “toevalsmechanisme” (uitkomsten en hun kansen), en dan kan je zonder verdere berekeningen (maar met eventuele hulp van ICT) je statistische toetsen uitvoeren en ten volle begrijpen wat zij betekenen.

Steekproef

=====

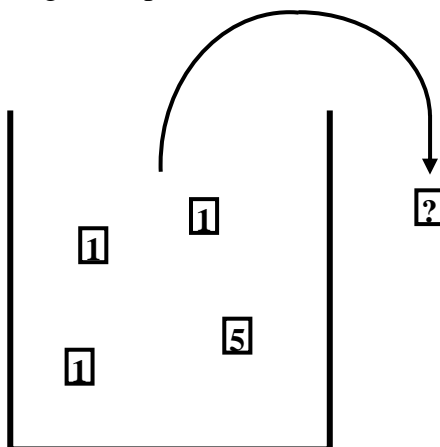
Uit de bovenstaande populatie X trek ik een steekproef van grootte 2. *Bemerk terug dat hier het volledige toevalsmechanisme aangeduid wordt met “populatie X ”. Hiermee wordt bedoeld dat ik niet zomaar trek uit de verzameling $\{1,5\}$ maar dat ook de kansen van de uitkomsten meespelen. Misschien is het goed om regelmatig de zin “ik trek uit een populatie X ” te vervangen door het synoniem “ik kijk naar de manier waarop het toevalsmechanisme X de getallen naar mij stuurt als ik dat toevalsmechanisme voor mij getallen laat genereren”*

Ik trek lukraak uit de vaas en zie dat er op mijn kaartje het cijfer 5 staat. Dat noteer ik door x_1 . Dan leg ik het kaartje terug in de vaas, schud eens goed, en trek terug een kaartje. Nu heb ik het cijfer 1. Dat resultaat noteer ik door x_2 . En algemeen noteer ik elk resultaat (of het nu eenzelfde getal oplevert of niet) gewoon in de volgorde waarin die getallen tot mij komen, dus x_1, x_2, \dots, x_n voor een steekproef van grootte n .

En nu beginnen we helemaal opnieuw. *Ik denk dat dit de juiste weg is om het aan te leren aan leerlingen: van concrete uitkomst naar achterliggend model, en niet omgekeerd.* Dus eerst altijd aan leerlingen vragen: wat zou je concreet doen? Dan laten aanvoelen dat, als zij dat morgen opnieuw doen, er iets anders zal uitkomen. En dan de vraag stellen naar: welk onderliggend mechanisme is er dat er voor zorgt dat jij dergelijke dingen uitkomt?

Uit de bovenstaande populatie X wil ik een steekproef van grootte 2 trekken. Wat zal er op mijn kaartje staan als ik een eerste keer ga trekken?

Hierop kan je alleen antwoorden met een “model”. En aangezien je uit die specifieke populatie X gaat trekken, zal je eerste trekking het volgende opleveren:



X_1

Inderdaad, dit is het toevalsmechanisme dat precies beschrijft wat je kan uitkomen en met welke kansen als je een eerste keer uit de populatie X zou gaan trekken. Dit toevalsmechanisme noteer je met een hoofdletter X_1 , en een toevallige waarde noteer je met de corresponderende kleine letter x_1 .

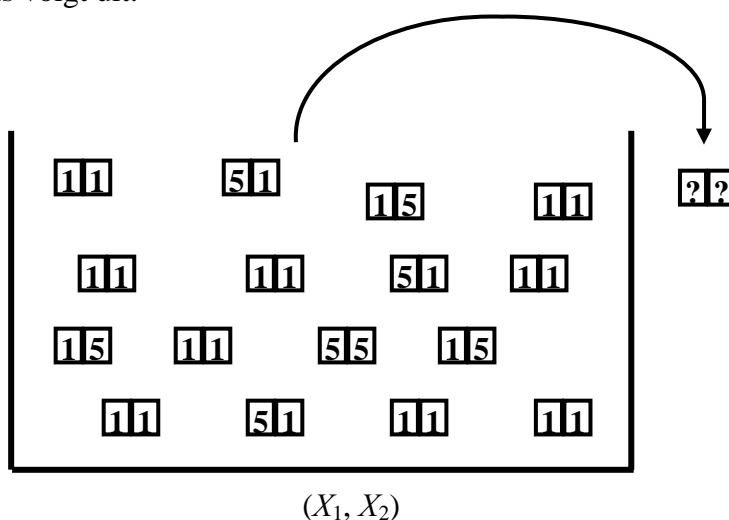
x_1	1	5
$P(X_1=x_1)$	$\frac{3}{4}$	$\frac{1}{4}$

Op dezelfde manier kan je het toevalsmechanisme van de tweede trekking opstellen:

x_2	1	5
$P(X_2=x_2)$	$\frac{3}{4}$	$\frac{1}{4}$

Mijn steekproef van grootte 2 is dus als model te schrijven als het geordende paar (X_1, X_2) . Dit geordend paar (X_1, X_2) beschrijft elk mogelijk geordend tweetal dat ik zou kunnen uitkomen samen met hun kansen als ik een steekproef van grootte twee zou gaan trekken uit deze populatie X .

Expliciet ziet dit er als volgt uit:



of ook:

(x_1, x_2)	$(1, 1)$	$(1, 5)$	$(5, 1)$	$(5, 5)$
$P((X_1, X_2) = (x_1, x_2))$	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$

Afspraak: het toevalsmechanisme (dat de resultaten van een steekproef van grootte n stuurt) wordt voorgesteld door hoofdletters (X_1, X_2, \dots, X_n) . De specifieke uitkomsten die ik in mijn toevallige steekproef vind stel ik voor door kleine letters (x_1, x_2, \dots, x_n) .

Grootheden gebaseerd op een steekproef:

- een eerste grootheid: het steekproefgemiddelde.

=====

Begin terug concreet. Als je uit de populatie X een steekproef van grootte twee trekt en je uitkomsten zijn 5 en 1, wat is dan het gemiddelde?

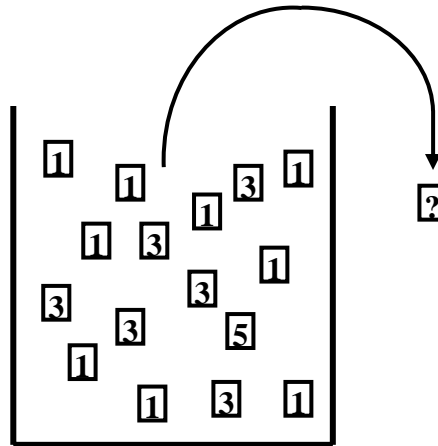
Wel, de formule en notatie zijn in deze situatie als volgt: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ zodat je hier $\bar{x} = 3$ vindt.

Realiseer je dan dat opnieuw een steekproef van grootte twee trekken uit diezelfde populatie, en dan terug het gemiddelde berekenen, waarschijnlijk niet hetzelfde zal opleveren. Om over “steekproefgemiddelde” te kunnen nadenken is er terug maar één wetenschappelijke weg (in de statistiek): je hebt de volledige specificatie nodig van het onderliggende toevalsmechanisme dat uw uitkomsten genereert (en in dit geval is dit: wat kan ik allemaal uitkomen, en met welke kans, als ik uit deze populatie X een steekproef van grootte twee zou gaan trekken en dan met de gevonden getallen het gemiddelde bereken).

Uit bovenstaand vaasmodel voor (X_1, X_2) kan je dat nu rechtstreeks aflezen. Op 9 van de 16 kaartjes staat (1,1) wat als som twee geeft en als gemiddelde één. Ik zal dus één vinden met kans 9/16. Een correcte

notatie voor het model van het steekproefgemiddelde is (hoofdletters !) $\bar{X} = \frac{1}{2} \sum_{i=1}^2 X_i$. Hierbij is \bar{X}

volledig vastgelegd door:



\bar{X}

of, equivalent, door:

\bar{x}	1	3	5
$P(\bar{X} = \bar{x})$	$\frac{9}{16}$	$\frac{6}{16}$	$\frac{1}{16}$

Afspraak: het toevalsmechanisme dat aan de basis ligt van de resultaten die je krijgt als je het gemiddelde

van n steekproefresultaten berekent wordt voorgesteld door een hoofdletter $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. De

specifieke waarden die je uitkomt bij het berekenen van het gemiddelde van uw toevallige

steekproefresultaten stel je voor door de corresponderende kleine letter $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Hou dus de “wereld van de onderliggende modellen” en de “wereld van de geobserveerde toevallige uitkomsten” duidelijk gescheiden, ook in notatie.

Grootheden gebaseerd op een steekproef:
 - een tweede grootheid: de steekproefvariantie.

=====

Begin terug concreet. Als je uit de populatie X een steekproef van grootte twee trekt en je uitkomsten zijn 5 en 1, wat is dan de variantie?

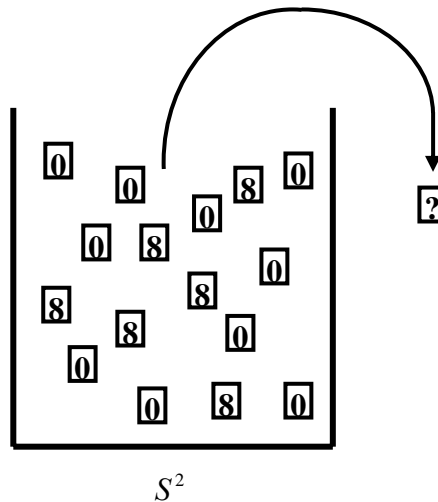
De formule en de notatie zijn als volgt: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ zodat je hier $s^2 = 8$ vindt.

Opnieuw een steekproef van grootte twee trekken levert waarschijnlijk twee andere getallen, en als je daar dan de variantie van berekent kan dat een andere uitkomst opleveren. Om “het gedrag” van de steekproefvariantie te begrijpen heb je terug de volledige specificatie nodig van het onderliggend toevalsmechanisme (en in dit geval is dit: wat kan ik allemaal uitkomen, en met welke kans, als ik uit deze populatie X een steekproef van grootte twee zou gaan trekken en dan de variantie zou berekenen van de gevonden getallen).

Uit bovenstaand vaasmodel voor (X_1, X_2) kan je dat nu rechtstreeks aflezen. Op 9 van de 16 kaartjes staat (1,1) en op één kaartje staat (5,5). Er zijn dus 10 van de 16 koppels die twee identieke getallen opleveren, en daarvan is de variantie gelijk aan nul. Verder zijn er nog 6 kaartjes die twee verschillende getallen opleveren, namelijk een één en een vijf. Het gemiddelde hiervan is 3 en de variantie is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = (1-3)^2 + (5-3)^2 = 8. \text{ Een correcte notatie voor het model van de}$$

steekproefvariantie is (hoofdletters !) $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Hierbij is S^2 volledig vastgelegd door:



of, equivalent, door:

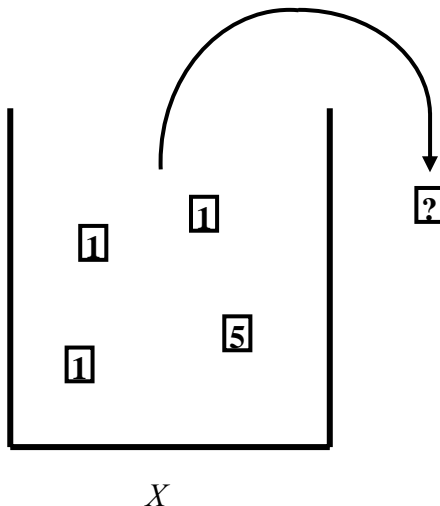
s^2	0	8
$P(S^2 = s^2)$	$\frac{10}{16}$	$\frac{6}{16}$

Afspraak: het toevalsmechanisme dat aan de basis ligt van de resultaten die je krijgt als je de variantie van n steekproefresultaten berekent wordt voorgesteld door een hoofdletter $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

De specifieke waarden die je uitkomt bij het berekenen van de variantie van uw toevallige steekproefresultaten stel je voor door de corresponderende kleine letter $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Eigenschappen van “ideale modellen” (toevalsmechanismen):
 -een eerste eigenschap: de verwachtingswaarde.

Wat is het gemiddelde van een dobbelsteen? Of wat is het gemiddelde van volgend vaasmodel?



Dit zijn vragen over eigenschappen van (theoretische) onderliggende modellen, niet van een aantal geobserveerde toevallige getallen.

Met behulp van de kansdefinitie vanuit het begrip relatieve frequentie kan je inzicht krijgen in de wiskundige definitie van een modeleigenschap (zoals de verwachtingswaarde).

Stel je voor dat je uit bovenstaande vaas lukraak trekt, het getal noteert en dan het kaartje teruglegt, en dat je dit zeer veel keren doet. De regulariteit in je toevallige uitkomsten zal ertoe aanleiding geven dat je (ongeveer) drie keer op vier een “1” op je kaartje hebt, en ongeveer één keer op vier een “5”. Als je nu met die “tienduizenden getallen” die je zo gevonden hebt het gemiddelde zou berekenen, dan zou je bij benadering iets hebben van de vorm $\frac{1}{n}(\frac{3n}{4} \text{ keer het cijfer } 1 \text{ plus } \frac{n}{4} \text{ keer het cijfer } 5)$ wat zeer goed lijkt op de gewogen som van alle mogelijke verschillende uitkomsten (als gewicht neem je de “belangrijkheid” van een uitkomst, namelijk de kans dat je precies die uitkomst vindt).

Inderdaad, de exacte formule voor wat je gemiddeld verwacht van een model, en die dus over een “intrinsieke modeleigenschap” gaat, is:

gemiddelde (of verwachtingswaarde) van een discreet toevalsmechanisme : $E(X) = \sum_i x_i P(X=x_i)$

Hierbij staat E voor verwachtingswaarde (= Expectation) en staat er na de E een hoofdletter X tussen de haakjes. Het gaat over wat je gemiddeld verwacht te vinden als je met dat model X werkt.

$E(X)$ is een eigenschap (of een karakteristiek) van het model, het is een getal (en dus geen model !!!!).

Als het toevalsmechanisme X het model is waarmee je de onderliggende populatie beschrijft, dan geef je $E(X)$ een speciale notatie. Het is immers gebruikelijk dat karakteristieken van een populatie aangeduid worden met Griekse letters, en de verwachtingswaarde van een populatiemodel (of het gemiddelde van een populatie) wordt genoteerd door μ .

In ons voorbeeld hebben we: $\mu = \sum_{i=1}^2 x_i P(X=x_i) = (1)\left(\frac{3}{4}\right) + (5)\left(\frac{1}{4}\right) = 2$ zodat het populatiegemiddelde gelijk is aan 2.

Voor elk discreet toevalsmechanisme kan je de verwachtingswaarde berekenen. De formule hiervoor is telkens dezelfde, namelijk de gewogen som van “alle mogelijke verschillende uitkomsten vermenigvuldigd met hun kansen”.

Laten we dat even bekijken voor het toevalsmechanisme dat het “steekproefgemiddelde” aanstuurt. Wat komt daar gemiddeld uit?

In ons voorbeeld hebben we de volledige specificatie opgesteld voor het toevalsmechanisme van ons steekproefgemiddelde. Dat was:

\bar{x}	1	3	5
$P(\bar{X} = \bar{x})$	$\frac{9}{16}$	$\frac{6}{16}$	$\frac{1}{16}$

Met de formule voor de verwachtingswaarde krijgen we hier:

$$E(\bar{X}) = \sum_i \bar{x}_i P(\bar{X} = \bar{x}_i) = (1)\left(\frac{9}{16}\right) + (3)\left(\frac{6}{16}\right) + (5)\left(\frac{1}{16}\right) = 2$$

Het feit dat we hier terug een waarde 2 vinden is geen toeval. Dit is gewoon een illustratie van een algemene eigenschap die zegt dat het toevalsmechanisme dat het steekproefgemiddelde aanstuurt een intrinsiek gemiddelde heeft dat gelijk is aan het intrinsiek gemiddelde van de populatie waaruit je de steekproeven trekt. Of verkort: **“het gemiddelde van het steekproefgemiddelde is het populatiegemiddelde”**. Je kan dit als volgt voorstellen. Trek een steekproef van grootte n en bereken het gemiddelde. Dan kom je ergens terecht. Trek uit dezelfde populatie terug een steekproef van grootte n en bereken het gemiddelde. Dan kom je ook ergens terecht, en dat zal wel niet op exact dezelfde plaats zijn als zopas. Blijf dit doen en bereken telkens het steekproefgemiddelde. Je zal dan getallen vinden die van elkaar afwijken. Maar als je het gemiddelde van deze getallen nu berekent, wel dan valt dat gemiddelde (in de long run) exact samen met het populatiegemiddelde. Je hebt dus de algemene (wiskundig bewijsbare) eigenschap dat:

$$E(\bar{X}) = \mu$$

Deze eigenschap gebruik je bij het opstellen van betrouwbaarheidsintervallen en bij het toetsen van hypothesen.

Wat is het gemiddelde van de steekproefvariantie S^2 ?

Ook hiervoor heb je eerst het onderliggende model van S^2 nodig. Dat was:

s^2	0	8
$P(S^2 = s^2)$	$\frac{10}{16}$	$\frac{6}{16}$

Met de formule voor de verwachtingswaarde krijg je hier:

$$E(S^2) = \sum_{i=1}^2 s_i^2 P(S^2 = s_i^2) = (0)\left(\frac{10}{16}\right) + (8)\left(\frac{6}{16}\right) = 3$$

Verder in deze tekst wordt aangetoond dat de populatievariantie σ^2 ook gelijk is aan 3. Dit is terug geen toeval maar een illustratie van de algemene (wiskundig bewijsbare) eigenschap dat $E(S^2) = \sigma^2$. Er geldt dus **dat het gemiddelde van de steekproefvariantie gelijk is aan de populatievariantie**.

Bemerk dat $E(S^2) = \sigma^2$ slechts waar is wanneer de steekproefvariantie gedefinieerd wordt als

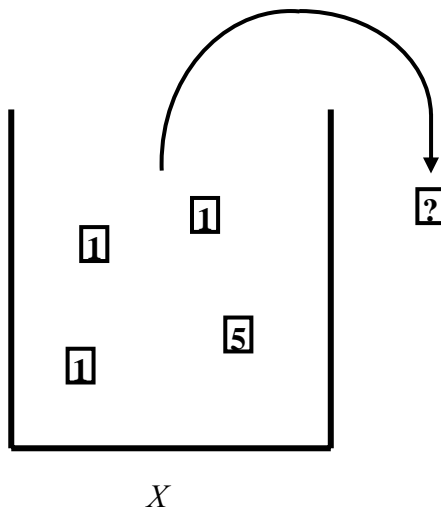
$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \text{ waarbij er gedeeld wordt door } (n-1).$$

De afspraken en eigenschappen die betrekking hebben op het gemiddelde (de verwachtingswaarde) van een "ideaal model" (toevalsmechanisme) kunnen we in onderstaande tabel samenvatten.

Afspraken en eigenschappen voor de verwachtingswaarde $E(\cdot)$		
Gemiddelde van de populatie X	$E(X) = \mu$	Dit is een notatieafpraak: de verwachtingswaarde van de populatie X geven we de naam μ
Gemiddelde van het steekproefgemiddelde \bar{X}	$E(\bar{X}) = \mu$	Dit is een algemene eigenschap. Het steekproefgemiddelde heeft een gemiddelde, en de populatie heeft een gemiddelde, en deze twee gemiddelden vallen samen.
Gemiddelde van de steekproefvariantie S^2	$E(S^2) = \sigma^2$	Dit is een algemene eigenschap. De steekproefvariantie heeft een gemiddelde, en dat gemiddelde valt samen met de populatievariantie.

Eigenschappen van “ideale modellen” (toevalsmechanismen):
 -een tweede eigenschap: de variantie.

Wat is de variantie van volgend vaasmodel?



Bemerk terug dat dit een vraag is over een eigenschap van een onderliggende model, niet van een aantal geobserveerde getallen.

Je kan hier volledig analoog redeneren zoals bij de verwachtingswaarde. De exacte formule voor de variantie van een model (wat dus een “intrinsieke modeleigenschap” is) ziet er als volgt uit:

variantie van een discreet toevalsmechanisme: $\text{var}(X) = \sum_i (x_i - E(X))^2 P(X=x_i)$

Als je naar de structuur van de formule kijkt zie je dat je terug te maken hebt met een gewogen som. Deze keer wordt, voor alle mogelijke verschillende uitkomsten, de som gemaakt van hun kwadratische afstand tot het modelgemiddelde, gewogen met hun eigen kans.

De standaardafwijking (of standaarddeviatie) kan je noteren als $sd(\cdot)$:

standaardafwijking van een discreet toevalsmechanisme: $sd(X) = \sqrt{\sum_i (x_i - E(X))^2 P(X=x_i)}$

Wanneer het toevalsmechanisme de beschrijving is van de onderliggende populatie, dan krijgt $\text{var}(X)$ een speciale notatie, namelijk σ^2 (Griekse letter). En daar in dit geval $E(X)$ genoteerd wordt door μ heb je

dat $\sigma^2 = \sum_i (x_i - \mu)^2 P(X=x_i)$.

Voor de populatie X beschreven in het vaasmodel, vind je dat

$\sigma^2 = \sum_i (x_i - \mu)^2 P(X=x_i) = (1-2)^2 \left(\frac{3}{4}\right) + (5-2)^2 \left(\frac{1}{4}\right) = 3$ zodat $\sigma = \sqrt{3}$.

Voor elk discreet toevalsmechanisme kan je de variantie en de standaardafwijking berekenen. De formule hiervoor is telkens dezelfde. De standaardafwijking van een model gebaseerd op een steekproef (zoals het steekproefgemiddelde en de steekproefvariantie) wordt meestal **standaardfout** genoemd. De afkorting hiervoor is $se(\cdot)$ wat verwijst naar de Engelse benaming: “standard error”.

Wat is de variantie van het steekproefgemiddelde?

Zoals altijd vertrek je vanuit de modelspecificatie van het toevalsmechanisme, die hier gegeven wordt door:

\bar{x}	1	3	5
$P(\bar{X} = \bar{x})$	$\frac{9}{16}$	$\frac{6}{16}$	$\frac{1}{16}$

Een rechtstreekse toepassing van de formule geeft (herinner u dat $E(\bar{X}) = 2$):

$$\text{var}(\bar{X}) = \sum_i (\bar{x}_i - E(\bar{X}))^2 P(\bar{X} = \bar{x}_i) = (1-2)^2 \left(\frac{9}{16}\right) + (3-2)^2 \left(\frac{6}{16}\right) + (5-2)^2 \left(\frac{1}{16}\right) = \frac{3}{2}$$

zodat de standaardfout van het steekproefgemiddelde gelijk is aan: $se(\bar{X}) = \sqrt{\frac{3}{2}}$.

Bemerk dat $\sqrt{3}$ ook gelijk is aan de standaardafwijking van de populatie X , en dat $\sqrt{2}$ de vierkantswortel is uit de steekproefgrootte (want hier is $n=2$). Dit is geen toeval. Je hebt hier een voorbeeld van een algemene (wiskundig bewijsbare) eigenschap die zegt dat: $se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

De standaardfout van het steekproefgemiddelde is gelijk aan de standaardafwijking van de populatie gedeeld door de vierkantswortel van de steekproefgrootte.

Deze eigenschap gebruik je bij het opstellen van betrouwbaarheidsintervallen en bij het toetsen van hypothesen. Hierbij is de standaardafwijking σ van de populatie meestal niet gekend, en moet die vervangen worden door een goede schatter. Een goed model, dat “gemiddeld” exact op σ valt, wordt

gegeven door $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, want $E(S^2) = \sigma^2$.

Dit verklaart waarom er gedeeld wordt door $(n-1)$.

=====

(*) Strikt genomen kan je de werelden van de statistiek herleiden tot twee werelden:

- De wereld van de “toevalsmechanismen” (die observatiegetallen genereren)
- De wereld van de concrete observatiegetallen zelf

Voor toevalsmechanismen heb je algemene notaties die hun eigenschappen aanduiden, zoals $E(\cdot)$ voor de verwachtingswaarde. Je hebt dus $E(X)$ voor de verwachtingswaarde van *het toevalsmechanisme dat de populatie aanstuurt*, $E(\bar{X})$ voor de verwachtingswaarde van *het toevalsmechanisme dat het steekproefgemiddelde aanstuurt*, enz..

De populatie, als onderliggend model waarover de primaire onderzoeksvraag gaat, en waarop ook alle verdere toevalsmechanismen steunen via de steekproef, krijgt een speciale plaats (en notatie) binnen al die “toevalsmechanismen” (maar de formules blijven wel dezelfde, namelijk de algemene formules voor toevalsmechanismen !). De notatieafpraak echter vraagt dat we eigenschappen van een populatie de naam “populatieparameters” geven, en dat we ze aanduiden met Griekse letters.

=====