Les 2 Steekproeven

We zullen in deze les bekijken, hoe we gegevens van een populatie zoals het gemiddelde en de spreiding kunnen schatten, zonder naar elk individu van de populatie te kijken. Het idee hierbij is, in plaats van de volledige populatie slechts naar een deel van de populatie te kijken, dit noemt men een *steekproef*. Men gaat ervan uit dat de steekproef typisch (representatief) voor de hele populatie is en bepaalt de gegevens van de populatie op de steekproef.

De cruciale vraag is hoe dicht de schatting op de steekproef bij de ware waarde voor de hele populatie ligt, d.w.z. wat voor een afwijking we moeten verwachten omdat we niet naar de hele populatie hebben gekeken.

Voor dat we ons hiermee gaan bemoeien, moeten we een aantal feiten over de normale verdeling verzamelen (herhalen), omdat deze verdeling de basis voor de analyse van steekproeven vormt.

2.1 De normale verdeling

De meest belangrijke verdeling in de statistiek is de *normale verdeling*. Deze wordt volledig bepaald door de verwachtingswaarde μ en de variantie σ^2 (of de standaardafwijking σ) en heeft de dichtheidsfunctie

$$f_{\mu,\sigma}(x) := \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Een stochast X die een kansverdeling met deze dichtheidsfunctie heeft, heet *normaal verdeeld* en wordt vaak met $X \in \mathcal{N}(\mu, \sigma^2)$ genoteerd.

De verdelingsfunctie voor een normaal verdeelde stochast kan niet zonder integraal geschreven worden, er geldt

$$F(x) := P(X \leq x) = \int_{-\infty}^{x} f_{\mu,\sigma}(t)\, dt.$$

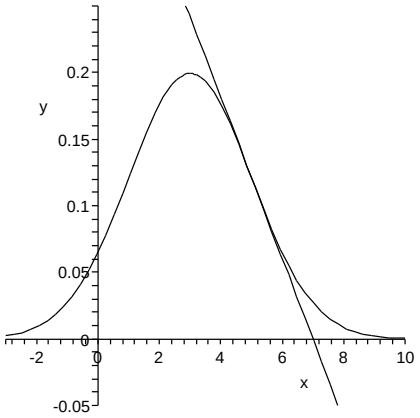
Voor een normaal verdeelde stochast X met verwachtingswaarde μ en variantie σ^2 heeft de *genormaliseerde stochast*

$$Z := \frac{X-\mu}{\sigma}$$

de verwachtingswaarde 0 en variantie 1. De stochast Z heet een *standaard-normaal verdeelde stochast*, zijn dichtheidsfunctie is de *standaard-normale verdeling* met de eenvoudigere dichtheidsfunctie

$$f(x) := f_{0,1}(x) := \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}.$$

De parameters μ en σ van een normale verdeling kunnen aan de grafiek van de dichtheidsfunctie $f(x)$ afgelezen worden zoals dit in Figuur 10 geïllustreerd is:



Figuur 10: Normale verdeling met $\mu = 3$ en $\sigma = 2$ en raaklijn aan de grafiek in $x = \mu + \sigma$.

- De verwachtingswaarde μ is het punt waar $f(x)$ zijn maximum heeft. Omdat de normale verdeling symmetrisch is, is dit ook de mediaan en de modus van de kansverdeling.
- De standaardafwijking σ vinden we op basis van het feit dat de grafiek van $f(x)$ juist in de punten $x = \mu - \sigma$ en $x = \mu + \sigma$ van kromming verandert. Op de punten waar en grafiek van kromming verandert is de stijging van de grafiek maximaal of minimaal en heeft de afgeleide van de functie dus een maximum of minimum (en dus de tweede afgeleide een nulpunt).

Omdat de verdelingsfunctie $F(x)$ van de normale verdeling niet makkelijk te berekenen is, worden de waarden vaak in tabellen aangegeven. Hierbij is het voldoende, de waarden voor de standaard-normale verdeling aan te geven, voor een willekeurige normale verdeling worden de waarden op de z-waarden van de standaard-normale verdeling genormaliseerd. Voor $z = \frac{x-\mu}{\sigma}$ en $Z = \frac{X-\mu}{\sigma}$ geldt immers:

$$P(X \leq x) = P(Z \leq z) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t} \, dt.$$

De tabellen voor de standaard-normale verdeling worden op twee manieren aangegeven:

(1) De waarden $P(Z \leq z)$ voor waarden van z in regelmatige afstanden, bijvoorbeeld afstanden van 0.05 tussen $z = -3$ en $z = 3$.

(2) Kritieke waarden van z zo dat $P(Z \leq z) = p$ voor zekere kansen p, bijvoorbeeld kansen in afstanden van 0.01 tussen 0 en 1.

Voorbeeld: Voor een normaal verdeelde stochast X met verwachtingswaarde 3 en standaardafwijking 2 willen we de kans $P(1 \leq X \leq 4)$ weten, dat een waarde tussen $x_1 = 1$ en $x_2 = 4$ ligt:

De genormaliseerde z-waarden zijn

$$z_1 = \frac{x_1 - 3}{2} = \frac{1-3}{2} = -1 \quad \text{en} \quad z_2 = \frac{x_2 - 3}{2} = \frac{4-3}{2} = 0.5.$$

De gezochte kans is dus $P(Z \leq 0.5) - P(Z \leq -1)$ voor de standaard-normaal verdeelde stochast Z. Voor deze twee kansen vinden we in een tabel de waarden

$$P(Z \leq -1) \approx 0.1587 \quad \text{en} \quad P(Z \leq 0.5) \approx 0.6915.$$

De gezochte kans is dus $0.6915 - 0.1587 = 0.5328$.

Als we omgekeerd willen weten voor welke waarde van x de kans $P(X \leq x) = 0.8$ is, vinden we in een tabel dat dit voor de z-waarde 0.8416 het geval is, dus voor $x = \sigma \cdot z + \mu = 2 \cdot 0.8416 + 3 = 4.6832$.

> Inmiddels wordt het aflezen van waarden van de normale verdeling uit tabellen meestal vervangen door statistiek programma's, die de benodigde waarden berekenen, maar het doet geen kwaad om ook het principe van de tabellen goed te begrijpen.

De redenen voor de centrale stelling van de normale verdeling in de statistiek zijn veelvoudig, de volgende opmerkingen geven hier een idee van:

(1) Voor zekere parameters worden andere kansverdelingen zoals de binomiale verdeling of de Poisson-verdeling door de normale verdeling goed benaderd.

(2) De combinatie van een groot aantal resultaten met bijna willekeurige kansverdelingen levert (bij benadering) een normale verdeling.

(3) De frequentieverdelingen van de uitkomsten van veel experimenten worden goed weergegeven door een normale verdeling, bijvoorbeeld kenmerken van populaties (grootte, gewicht), herhaald meten van gegevens, resultaten van een grote groep mensen bij een test, enz. Dit is ten dele een consequentie uit het punt (2), want vaak is een grootheid bepaald door een aantal enigszins onafhankelijke factoren en de combinatie daarvan geeft een normale verdeling.

De punten (1) en (2) zullen we nu iets nader toelichten.

Normale benadering van andere kansverdelingen

Stel een toevalsexperiment levert met kans p een succes op, dan heeft de stochast X die het aantal successen in n pogingen telt een *binomiale verdeling* en er geldt

$$P(X = k) = b(n, p; k) = \binom{n}{k} p^k (1-p)^k.$$

Een binomiaal verdeelde stochast X heeft de verwachtingswaarde $E[X] = np$ en de variantie $Var(X) = np(1-p)$. We transformeren X met behulp van

$E[X]$ en $Var(X)$ op een stochast Z die verwachtingswaarde 0 en variantie (of standaardafwijking) 1 heeft. Hiervoor definiëren we:

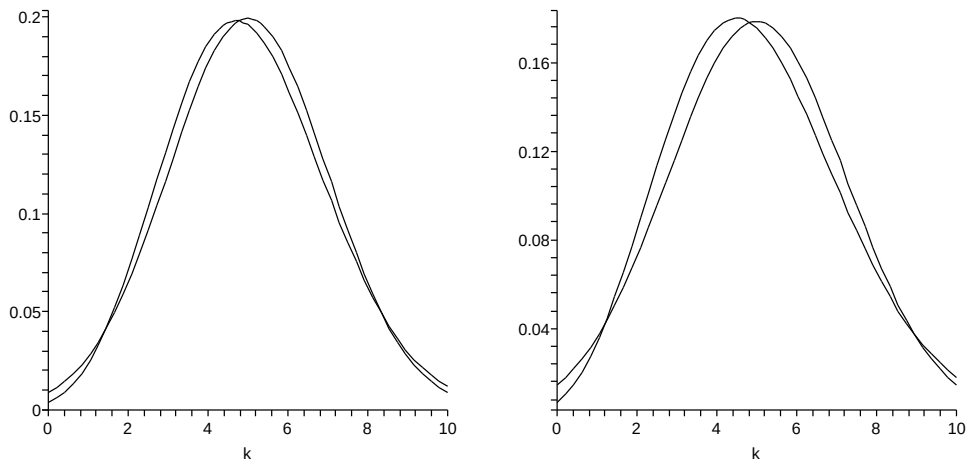
$$Z := \frac{X - np}{\sqrt{np(1-p)}}.$$

Als we n laten groeien, maakt de *stelling van De Moivre en Laplace* een belangrijke uitspraak over de stochast Z:

Stelling van *De Moivre* en *Laplace*: De limiet $\lim_{n\to\infty} \frac{X - np}{\sqrt{np(1-p)}}$ is een standaard-normaal verdeelde stochast.

Omgekeerd betekent dit, dat voor niet te kleine waarden van n de binomiale verdeling met parameters n en p door de normale verdeling met parameters $\mu = np$ en $\sigma^2 = np(1-p)$ benaderd kan worden. We noemen dit de *normale benadering* van de binomiale verdeling.

De benadering is beter als p in de buurt van $\frac{1}{2}$ ligt en slechter als p dicht bij 0 of 1 ligt. Als vuistregel wordt vaak gehanteerd, dat de normale benadering van de binomiale verdeling toegestaan is als $np \geq 5$ en $n(1-p) \geq 5$ (soms wordt ook $np \geq 10$ en $n(1-p) \geq 10$ geëist).



Figuur 11: Normale benadering van de binomiale verdeling met parameters $n = 25$ en $p = 0.2$ (links) en van de Poisson-verdeling met parameter $\lambda = 5$ (rechts).

We weten dat we voor een stochast X van zeldzame gebeurtenissen (dus met kleine p) de binomiale verdeling door de Poisson-verdeling met parameter $\lambda = np$ kunnen benaderen. Voor de kansen bij de Poisson-verdeling geldt

$$P(X = k) = po_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

en de stochast X heeft verwachtingswaarde $E[X] = \lambda$ en variantie $Var(X) = \lambda$.

Nadat we de binomiale verdeling behandeld hebben, zal het nu geen verrassing meer zijn, dat ook de Poisson-verdeling door de normale verdeling benaderd kan worden, als de parameter λ niet te klein is. Uit de stelling van De Moivre en Laplace volgt namelijk, dat voor een stochast X die Poisson-verdeeld met parameter λ is, de stochast

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

bij benadering standaard-normale verdeeld is.

Omgekeerd noemt men de normale verdeling met $\mu = \lambda$ en $\sigma^2 = \lambda$ de *normale benadering* van de Poisson-verdeling met parameter λ. Analoog met de binomiale verdeling wordt ook hier als vuistregel van de toepasbaarheid van de benadering meestal $\lambda \geq 5$ gehanteerd.

Dat de benaderingen voor de aangegeven grenzen inderdaad redelijk goed zijn, kunnen we aan de voorbeelden in Figuur 11 zien. Merk op dat de binomiale verdeling en de Poisson-verdeling scheef naar rechts zijn. Daarom ligt de modus van de twee in Figuur 11 aangegeven verdelingen links van 5 (bij 4.69 voor de binomiale verdeling en bij 4.49 voor de Poisson-verdeling) en is de normale verdeling dus telkens de verdeling met het maximum meer rechts.

Centrale limietstelling

De uitspraak van één van de meest belangrijke (en misschien ook meest verbazingwekkende) stellingen in de kansrekening en statistiek is ruwweg, dat de combinatie van min of meer willekeurige kansverdelingen bij benadering een normale verdeling geeft. Deze stelling heet de *Centrale limietstelling* en de precieze formulering luidt als volgt:

Stelling: Als $X_1, X_2, \ldots$ onafhankelijke stochasten zijn met verwachtingswaarde $E[X_i]$ en variantie $Var(X_i)$, dan is de limiet

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n}(X_i - E[X_i])}{\sqrt{\sum_{i=1}^{n} Var(X_i)}}$$

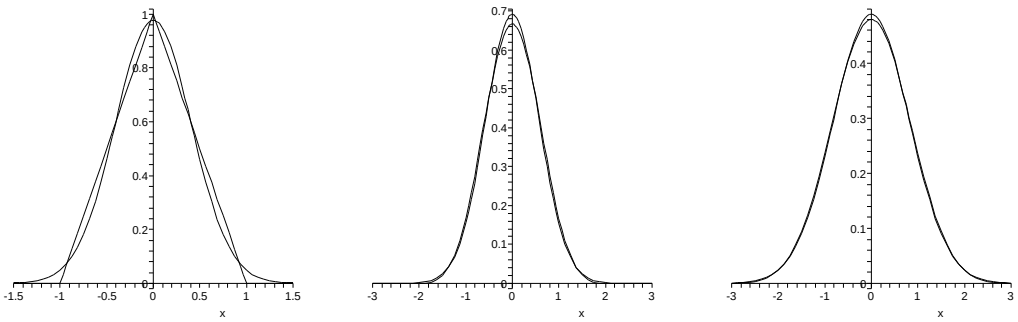
onder zwakke verdere voorwaarden aan de X_i een standaard-normaal verdeelde stochast. In het bijzonder wordt aan de voorwaarden voldaan als alle X_i dezelfde standaardafwijking σ hebben, in dit geval convergeert

$$\frac{1}{\sqrt{n}\,\sigma}\left(\sum_{i=1}^{n} X_i - E[X_i]\right)$$

tegen de standaard-normale verdeling.

Uit deze stelling kunnen we omgekeerd concluderen dat de normale verdeling met verwachtingswaarde $\mu = \sum_{i=1}^{n} E[X_i]$ en variantie $\sigma^2 = \sum_{i=1}^{n} Var(X_i)$ een benadering geeft voor de kansverdeling van de stochast $X := \sum_{i=1}^{n} X_i$. Hoe goed deze benadering is, hangt van de verdelingen van de enkele stochasten X_i en natuurlijk van n af.

Als voorbeeld kijken we naar de combinatie van n stochasten X_i met uniforme verdelingen op het interval $[-\frac{1}{2}, \frac{1}{2}]$. Omdat de verdelingen symmetrisch



Figuur 12: Benadering van de som van n uniforme verdeling door een normale verdeling voor $n = 2$, $n = 4$ en $n = 8$.

rond 0 liggen, is $E[X_i] = 0$ en voor de variantie geldt $Var(X_i) = \frac{1}{12}$. De som $X_1 + \ldots + X_n$ wordt dus benaderd door de normale verdeling met $\mu = 0$ en $\sigma^2 = \frac{n}{12}$. In Figuur 12 is de benadering voor $n = 2$, $n = 4$ en $n = 8$ te zien. Het is duidelijk, dat al voor $n = 4$ de normale verdeling een heel goede benadering geeft.

2.2 Aselecte steekproeven

We hebben in de eerste les gezien hoe we uit een verzameling gegevens uitspraken kunnen afleiden over typische waarden, spreiding, scheefheid, enz. van de gegevens. Hierbij hebben we altijd gebruik gemaakt van de kennis van *alle* gegevens. In de praktijk is dit vaak ondoenlijk of onwenselijk, omdat we uitspraken willen maken over een verzameling gegevens waarvan we niet ieder individu te pakken krijgen. In zo'n geval nemen we een deel van de gegevens - een *steekproef* - en proberen uit de resultaten op de steekproef conclusies over de volledige verzameling gegevens te trekken. Voorbeelden van deze situatie zijn:

- Verkiezingen: Om de percentages van de verschillende opties (verschillende partijen, ja/nee bij een referendum) bij een toekomstige verkiezing te schatten, wordt in een enquête een steekproef van typisch 1000 of 2000 mensen ondervraagd.

- Kwaliteitstoetsen: Om de percentage defecte stukken in een productie te schatten, nemen we een steekproef en testen de gekozen stukken. Het relatieve aantal defecte stukken in de steekproef nemen we als gok voor de percentage in de volledige productie.

- Gemiddelde waarden: Om de gemiddelde *intelligentiequotiënt* of *body-mass-index* in de bevolking te schatten, bepalen we deze voor een geselecteerde groep mensen.

Het idee achter het nemen van een steekproef zit in de veronderstelling, dat de steekproef *representatief* voor de volledige verzameling is. De manier hoe een

steekproef wordt genomen, heeft natuurlijk een grote invloed erop of dit inderdaad klopt. Het is bijvoorbeeld bekend dat verkiezingsresultaten tussen zekere groepen in de bevolking duidelijk verschillen, afhankelijk van inkomen, leeftijd of burgerlijke staat van de mensen in een groep. Men moet daarom ervoor zorgen, dat deze factoren in de steekproef met de juiste relatieve frequenties gerepresenteerd zijn.

Een voorbeeld van een slechte steekproef is, bij een enquête gewoon de eerste 100 mensen te vragen die je tegenkomt. Dit zou bijna nooit representatief zijn, omdat je op zekere plekken vooral mensen met gemeenschappelijke eigenschappen tegenkomt, op het station bijvoorbeeld mensen die naar hun werkplek reizen en op de campus van de universiteit studenten. Ook als je in de telefoongids willekeurig nummers kiest, is dit meestal niet representatief, omdat je mensen zonder telefoon buiten beschouwing laat en afhankelijk van de tijd verschillende bewoners van een woning bereikt.

> Het juiste kiezen van een steekproef is een moeilijke taak waarmee zich een belangrijk speciaal gebied van de statistiek bezig houdt.
>
> We zullen ons echter in dit college niet verder met de vraag van het juiste opzetten van steekproeven bemoeien, we gaan er vanaf nu van uit dat we het goed hebben gedaan en het met een *aselecte steekproef* te maken hebben.

Een *aselecte steekproef* (zoals we die vanaf nu als gegeven veronderstellen) is een steekproef die aan de volgende twee eisen voldoet:

(1) De steekproef is *onbevooroordeeld* (unbiased): Elk individu heeft dezelfde kans om gekozen te worden.

(2) De steekproef is *onafhankelijk*: De keuze van één individu voor de steekproef heeft geen invloed op de kansen van de andere individuen om in de steekproef te komen.

2.3 Het gemiddelde van een steekproef

Vaak berekenen we het gemiddelde van een steekproef en gebruiken dit als schatting voor het gemiddelde (of de verwachtingswaarde) van de volledige populatie. Als we bijvoorbeeld bij een kwaliteitstoets de kans op een foutief stuk in een productieproces willen bepalen, nemen we hiervoor als *schatting* de relatieve frequentie van foutieve stukken in een (aselecte) steekproef. De vraag is nu, hoe goed de schatting vanuit de steekproef voor de echte kans is, dus hoe sterk het gemiddelde van de steekproef van het gemiddelde van de populatie afwijkt.

Het cruciale idee, om bij deze vraag verder te komen, is dat we ons voorstellen, het nemen van de steekproef vaak te herhalen en de uitslagen van de enkele steekproeven als toevalsexperiment, dus als stochast te beschouwen.

Stel we hebben een steekproef $x_1, \ldots, x_n$. Dan kunnen we ieder element x_i in de steekproef als resultaat van een stochast X_i beschouwen en als we veronderstellen dat de elementen in de steekproef op grond van hetzelfde proces geproduceerd worden, hebben de stochasten X_i alle dezelfde kansverdeling. Merk op dat we bij deze aanpak iets over het onderliggende proces veronderstellen, bijvoorbeeld dat bij de productie van de gecontroleerde stukken inderdaad elk stuk met kans p defect is en dat dit bij de verschillende stukken onafhankelijk gebeurt.

Als we nu naar *alle* mogelijke steekproeven $x_1, \ldots, x_n$ willen kijken, kunnen we dit met behulp van de stochasten $X_1, \ldots, X_n$ beschrijven, want X_i geeft juist de kans aan waarmee het resultaat x_i voorkomt. Op deze manier krijgen we in het bijzonder voor het *steekproefgemiddelde*

$$\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$$

de stochast

$$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

die de verdeling van de steekproefgemiddelden over alle mogelijke steekproeven aangeeft.

> **Merk op:** Het is in de literatuur gebruikelijk, een concrete steekproef met kleine letters (zoals x_1, x_2, y) aan te geven, terwijl hoofdletters (zoals X_1, X_2, Y) de stochasten voor de verdeling over alle steekproeven aangeven.

Voorbeeld: Zij X de stochast van een Bernoulli-experiment met parameter p, d.w.z. er geldt $P(X = 1) = p$ en $P(X = 0) = 1 - p$. De verwachtingswaarde $E[X]$ is dan

$$E[X] = p \cdot 1 + (1 - p) \cdot 0 = p$$

en de variantie $Var(X)$ is

$$Var(X) = p \cdot (1 - p)^2 + (1 - p) \cdot p^2 = p(1 - p).$$

Als we een steekproef van grootte n nemen, herhalen we het Bernoulli-experiment n keer onafhankelijk en hebben hierbij n stochasten $X_1, \ldots, X_n$ met dezelfde verdeling als X.

Voor de stochast $\overline{X} := \frac{1}{n}(X_1 + \ldots + X_n)$ die de relatieve frequentie van 1en bij n pogingen aangeeft, hebben we

$$E[\overline{X}] = \frac{1}{n}(p + \ldots + p) = \frac{1}{n}\, np = p$$

dus is de verwachtingswaarde van de steekproefgemiddelden inderdaad de juiste parameter p. Als we dus meerdere steekproeven nemen, kunnen we ervan uitgaan dat de *ware* waarde van p ongeveer het gemiddelde van de steekproefgemiddelden is.

Het feit dat we in plaats van (bijvoorbeeld) 10 steekproeven met grootte n apart te nemen ook meteen een grotere steekproef van grootte $10n$ hadden kunnen nemen om de waarde van p te schatten, leidt tot de interessante vraag hoe ver het steekproefgemiddelde van de juiste waarde van p afwijkt.

Maar hierover maakt juist de variantie $Var(\overline{X})$ van de stochast $\overline{X}$ een uitspraak, we kunnen namelijk verwachten dat het steekproefgemiddelde 'meestal' niet meer dan één standaardafwijking $\sigma_{\overline{X}}$ van p afwijkt, en de standaardafwijking $\sigma_{\overline{X}}$ is gegeven door $\sigma_{\overline{X}} = \sqrt{Var(\overline{X})}$. De variantie van $\overline{X}$ laat zich berekenen door

$$Var(\overline{X}) = \frac{1}{n^2}(p(1-p) + \ldots + p(1-p)) = \frac{1}{n^2}\, np(1-p) = \frac{1}{n}\, p(1-p).$$

Dit betekent dat het steekproefgemiddelde een standaardafwijking van $\sqrt{\frac{p(1-p)}{n}}$ heeft. In het bijzonder neemt de onzekerheid van de schatting van p met de wortel uit de grootte van de steekproef af.

Omdat we steeds van een aselecte steekproef uitgaan, is voor het n keer herhalen van een Bernoulli-experiment de Centrale limietstelling van toepassing en we krijgen voor niet te kleine n als verdeling voor de waarde van $\overline{X}$ (bij benadering) een normale verdeling. Dit betekent dat het steekproefgemiddelde met een kans van ongeveer 68% in het interval

$$\left[p - \sqrt{\frac{p(1-p)}{n}},\ p + \sqrt{\frac{p(1-p)}{n}}\right]$$

ligt, want dit is juist de kansmassa die bij de normale verdeling tussen $\mu - \sigma$ en $\mu + \sigma$ ligt.

> Merk op dat we in het voorbeeld een alternatieve verdeling met parameter p verondersteld hebben, en hiermee iets over de verdeling van $\overline{X}$ konden zeggen. Dit is de situatie van een *hypothese* die we over de onderliggende kansverdeling hebben en die we met de realisaties $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ van $\overline{X}$ op concrete steekproeven kunnen toetsen. Het probleem van het toetsen van hypothesen zullen we later in deze cursus behandelen.

Het resultaat van het voorbeeld met het Bernoulli-experiment geldt inderdaad algemeen voor het bepalen van het gemiddelde van gegevens:

Stel we willen het gemiddelde van een zekere grootheid bepalen, dan zien we elke meting als het resultaat van een kansexperiment met een stochast X die een zekere kansverdeling heeft. We *veronderstellen* dus een stochast X met verwachtingswaarde $E[X]$ en standaardafwijking $\sigma = \sigma_X = \sqrt{Var(X)}$.

Bij een steekproef van n metingen beschouwen we het *steekproefgemiddelde* $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$ als uitkomst voor de nieuwe stochast $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$, waarbij de stochasten X_i dezelfde kansverdeling als de veronderstelde stochast X hebben. Voor de stochast $\overline{X}$ van het steekproefgemiddelde geldt nu:

$$E[\overline{X}] = \frac{1}{n}(E[X_1] + \ldots + E[X_n]) = \frac{1}{n}\, n \cdot E[X] = E[X]$$

en

$$Var(\overline{X}) = \frac{1}{n^2}(Var(X_1) + \ldots + Var(X_n)) = \frac{1}{n^2}\, n \cdot Var(X) = \frac{1}{n}\, \sigma_X^2$$

dus geldt voor de variantie $\sigma^2_{\overline{X}}$ en de standaardafwijking $\sigma_{\overline{X}}$ van $\overline{X}$:

$$\sigma^2_{\overline{X}} = \frac{1}{n}\, \sigma_X^2 \quad \text{en} \quad \sigma_{\overline{X}} = \frac{1}{\sqrt{n}}\, \sigma_X.$$

De verdeling van het steekproefgemiddelde heeft dus dezelfde verwachtingswaarde als de onderliggende kansverdeling en de standaardafwijking van de steekproefgemiddelden neemt met de wortel uit de grootte van de steekproef af. Merk op dat we bij het berekenen van de variantie van $\overline{X}$ weer gebruik ervan hebben gemaakt dat de X_i onafhankelijk zijn, dus dat we het met een aselecte steekproef te maken hebben.

> Strikt genomen geldt $\sigma^2_{\overline{X}} = \frac{1}{n}\, \sigma_X^2$ voor de variantie van $\overline{X}$ alleen maar als we een steekproef uit een oneindige populatie nemen of als we de steekproef door trekken met terugleggen verkrijgen. Dit is bijvoorbeeld bij herhaalde metingen van een waarde van toepassing, want in principe kunnen we oneindig lang doorgaan met de metingen en de populatie is dus oneindig.
>
> Als een steekproef van grootte n uit een eindige populatie met N elementen door trekken *zonder terugleggen* genomen wordt, geldt voor de variantie van het steekproefgemiddelde
>
> $$\sigma^2_{\overline{X}} = \frac{1}{n}\, \sigma_X^2 \left(\frac{N-n}{N-1}\right).$$
>
> Maar deze correctie kunnen we in de praktijk bijna altijd verwaarlozen, omdat N veel groter is dan n (anders zouden we geen steekproef nemen, maar de hele populatie bekijken) en dus $\frac{N-n}{N-1}$ heel dicht bij 1 ligt.

Het probleem is nu, dat we over de kwaliteit van onze schatting voor het gemiddelde $E[X]$ alleen iets kunnen zeggen als we de standaardafwijking σ_X van X kennen.

2.4 De standaardafwijking van een steekproef

Net zo als we het steekproefgemiddelde als het gemiddelde $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$ van de waarden in een steekproef hebben gedefinieerd, kunnen we ook een *steekproefvariantie* en een *steekproefstandaardafwijking* definiëren. De voor de hand liggende gedachte zou zijn, de steekproefvariantie door $\frac{1}{n}((x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2)$ te definiëren. Maar met het steekproefgemiddelde is al een afhankelijkheid tussen de x_i gegeven, als we namelijk $x_1, \ldots, x_{n-1}$ en $\overline{x}$ kennen, ligt x_n vast. Men zegt daarom, dat we slechts nog $n-1$ *vrijheidsgraden* hebben, omdat we met $\overline{x}$ een afhankelijkheid tussen de x_i ingevoerd hebben. In plaats van de som van de kwadratische afstanden door n te delen, delen we door

het aantal $n-1$ van *onafhankelijke waarden* in de steekproef en definiëren de steekproefvariantie s^2 en de steekproefstandaardafwijking s als volgt:

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{en} \quad s := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Er is ook een minder heuristische verklaring voor het gebruiken van $n-1$ in plaats van n in de noemer. Dit hangt samen met de theorie van *schatters* die we in de volgende les gaan bediscussiëren. Het cruciale punt is, dat we graag willen dat de verwachtingswaarde van de steekproefvariantie de ware variantie σ^2 van de onderliggende verdeling geeft, net zo als de verwachtingswaarde $E[\overline{X}]$ van het steekproefgemiddelde de ware verwachtingswaarde $E[X]$ is.

Om de verdeling van de steekproefvariantie over verschillende steekproeven te analyseren, definiëren we weer een stochast X met de onderliggende kansverdeling en nemen aan dat alle mogelijke steekproeven door onafhankelijke stochasten $X_1, \ldots, X_n$ met dezelfde kansverdeling als X worden beschreven.

De verwachtingswaarde en variantie van X noteren we met $\mu := E[X]$ en $\sigma^2 := Var(X)$. We weten dat $\sigma^2 = E[X^2] - E[X]^2$, dus is $E[X^2] = \sigma^2 + \mu^2$.

De stochast $\overline{X}$ voor het steekproefgemiddelde is weer gedefinieerd door

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + \ldots + X_n).$$

Er geldt

$$(X_i - \overline{X})^2 = (X_i - \frac{1}{n}(\sum_j X_j))^2 = X_i^2 - \frac{2}{n}X_i(\sum_j X_j) + \frac{1}{n^2}\sum_{j,k} X_j X_k.$$

Als we dit over alle indices i optellen, krijgen we

$$\begin{aligned}\sum_i (X_i - \overline{X})^2 &= \sum_i X_i^2 - \frac{2}{n}\sum_{i,j} X_i X_j + n\frac{1}{n^2}\sum_{j,k} X_j X_k \\ &= \sum_i X_i^2 - \frac{1}{n}\sum_{j,k} X_j X_k = \sum_i X_i^2 - \frac{1}{n}(\sum_i X_i)^2.\end{aligned}$$

Er geldt $E[X_i^2] = \sigma^2 + \mu^2$, $E[\sum_i X_i] = n\mu$ en $Var(\sum_i X_i) = n\sigma^2$. Hieruit volgt $E[(\sum_i X_i)^2] = Var(\sum_i X_i) + E[\sum_i X_i]^2 = n\sigma^2 + n^2\mu^2$ en hiermee krijgen we

$$\begin{aligned}E[\sum_i (X_i - \overline{X})^2] &= E[\sum_i X_i^2] - \frac{1}{n}E[(\sum_i X_i)^2] \\ &= n(\sigma^2 + \mu^2) - \frac{1}{n}(n\sigma^2 + n^2\mu^2) = n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\ &= (n-1)\sigma^2.\end{aligned}$$

We moeten dus de steekproefvariantie als $s^2 := \frac{1}{n-1}(\sum_i (x_i - \overline{x})^2)$ definiëren, om als verwachtingswaarde van de steekproefvariantie over alle steekproeven de variantie σ^2 te krijgen. De stochast die de verdeling van de steekproefvarianties beschrijft noemen we S^2 en definiëren deze door

$$S^2 := \frac{1}{n-1}(\sum_i (X_i - \overline{X})^2).$$

2.5 Student t-verdeling en χ^2-verdeling

Student t-verdeling

Bij een stochast X krijgen we de verdeling van de z-waarden door $Z := \frac{X-\mu}{\sigma}$ en analoog krijgen we bij een steekproef van n waarden de z-waarde van het steekproefgemiddelde als

$$z := \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\overline{x} - \mu}{s}\sqrt{n}$$

waarbij we de onbekende standaardafwijking σ door de steekproefstandaardafwijking s vervangen.

Om de verdeling van de z-waarden van het steekproefgemiddelde te beschrijven, interpreteren we de elementen x_i van een steekproef weer als realisaties van stochasten X_i, dan wordt de verdeling van de z-waarden beschreven door de stochast

$$T := \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\overline{X} - \mu}{S}\sqrt{n} \text{ met } \overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \text{ en } S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Voor een normaal verdeelde stochast X heet de kansverdeling van T de *Student t-verdeling* met $n-1$ vrijheidsgraden. De Student t-verdeling is platter dan de standaard-normale verdeling maar komt voor groeiende n steeds dichter bij de standaard-normale verdeling. De oorzaak hiervoor is de onzekerheid over de variantie die de steekproefgemiddelden sterker om de ware waarde van het gemiddelde verspreidt.

> De rare naam van deze verdeling gaat terug op William Sealey Gosset (1876-1937), die 1908 een artikel hierover gepubliceerd heeft. Omdat hij als medewerker van de *Guinness* brouwerij niet onder zijn eigen naam mocht publiceren, koos hij het pseudoniem *Student* voor zijn wetenschappelijke artikelen. Een beschrijving van hem zegt: *To many in the statistical world "Student"was regarded as a statistical advisor to Guinness's brewery, to others he appeared to be a brewer devoting his spare time to statistics.*

De dichtheidsfunctie van de Student t-verdeling met n vrijheidsgraden is

$$f_n(x) := C_n\,(1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$$

waarbij de normaliseringsconstante C_n gegeven is door

$$C_n := \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \frac{1}{\sqrt{\pi n}}.$$

De hierbij optredende *Gamma-functie* $\Gamma(t)$ is gedefinieerd door

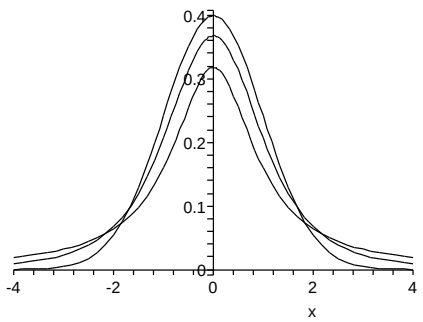
$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dt.$$

Ook dit is (net als de verdelingsfunctie van de normale verdeling) een functie die niet zonder integraal te schrijven is. Uit de eigenschappen $\Gamma(t+1) = t\Gamma(t)$ en $\Gamma(1) = 1$ volgt dat $\Gamma(n+1) = n!$ voor natuurlijke getallen n. De Gamma-functie is dus een soort interpolatie van de faculteit en speelt daarom in veel gebieden van de wiskunde een belangrijke rol.

Omdat de Student t-verdeling symmetrisch is, heeft een stochast T met deze verdeling de verwachtingswaarde $E[T] = 0$. Heeft T een verdeling met $n \geq 3$ vrijheidsgraden, dan geldt

$$Var(T) = \frac{n}{n-2},$$

de variantie is dus inderdaad groter dan bij de standaard-normale verdeling.



Figuur 13: Student t-verdeling voor $n = 1$ en $n = 3$ in relatie tot standaard-normale verdeling.

χ^2-verdeling

Met de Student t-verdeling wordt de verdeling van de steekproefgemiddelden bij onbekende onderliggende variantie beschrijven. Een andere klasse van functies is geschikt om de verdeling van de steekproefvarianties te beschrijven.

Voor n standaard-normaal verdeelde stochasten $X_1, \ldots, X_n$ heet de verdeling van de stochast $Y = X_1^2 + \ldots + X_n^2$ een χ^2-*verdeling* met n vrijheidsgraden. Het betekenis van deze verdeling ligt in het verband met de verdeling van de steekproefvarianties:

Voor de stochast S^2 van de steekproefvarianties geldt

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2 = \frac{\sigma^2}{n-1}\sum_i \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 .$$

Nu is $\frac{X_i - \overline{X}}{\sigma}$ zelf niet standaard-normaal verdeeld, maar voor de stochast $\frac{X_i - \mu}{\sigma}$ geldt dit wel, dus is $\sum_i (\frac{X_i - \mu}{\sigma})^2$ een χ^2-verdeling met n vrijheidsgraden.

Met behulp van de relatie

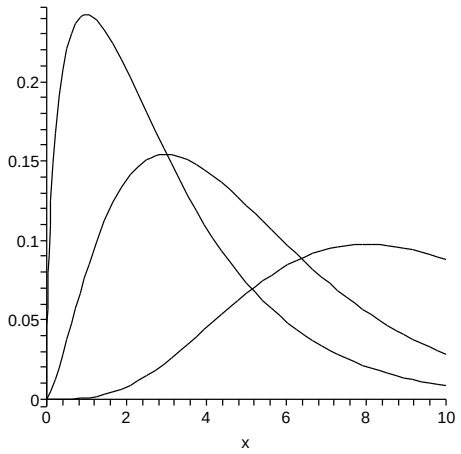
$$\sum_i (X_i - \overline{X})^2 = \sum_i (X_i - \mu)^2 - n(\overline{X} - \mu)^2$$

laat zich aantonen dat $\sum_i (\frac{X_i - \overline{X}}{\sigma})^2$ inderdaad wel een χ^2-verdeling met $n-1$ vrijheidsgraden is, dus geldt samengevat:

$$\frac{n-1}{\sigma^2} S^2 = \sum_i (\frac{X_i - \overline{X}}{\sigma})^2 \text{ heeft een } \chi^2\text{-verdeling met } n-1 \text{ vrijheidsgraden.}$$

Ook de χ^2-verdelingen kunnen we expliciet aangeven, de χ^2-verdeling met n vrijheidsgraden heeft de dichtheidsfunctie

$$f_n(x) = \begin{cases} C_n\, x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{voor } x > 0 \\ 0 & \text{voor } x \leq 0, \end{cases} \quad \text{waarbij } C_n = (2^{\frac{n}{2}} \cdot \Gamma(\frac{n}{2}))^{-1}.$$



Figuur 14: χ^2-verdelingen voor $n = 3$, $n = 5$ en $n = 10$.

Voor een stochast Y met χ^2-verdeling met n vrijheidsgraden geldt

$$E[Y] = n \text{ en } Var(X) = 2n$$

en voor $n \to \infty$ wordt de χ^2-verdeling steeds beter benaderd door een normale verdeling met $\mu = n$ en $\sigma^2 = 2n$.

We zullen de χ^2-verdeling in het kader van betrouwbaarheidsintervallen en het toetsen van hypothesen in dit cursus nog vaker tegen komen.

Belangrijke begrippen in deze les

- normale verdeling
- normale benadering
- Centrale limietstelling
- steekproef, aselecte steekproef
- steekproefgemiddelde, -variantie, -standaardafwijking
- Student t-verdeling
- χ^2-verdeling

Opgaven

6. Laten $X_1, \ldots, X_n$ onafhankelijke normaal verdeelde stochasten zijn met $E[X_i] = \mu_i$ en $Var(X_i) = \sigma_i^2$. Er geldt dat ook de lineaire combinatie $Y = a_1X_1 + \ldots + a_nX_n$ een normaal verdeelde stochast is.

   Bereken de verwachtingswaarde $E[Y]$ en de variantie $Var(Y)$ van Y.

7. Een populatie bestaat uit de vier waarden 3, 7, 11 en 13. Een mogelijke methode om het gemiddelde van de populatie te schatten, is steekproeven van 2 elementen *met* terugleggen te nemen en hiervan het gemiddelde te bepalen. Algemeen noemt men een methode om een parameter van een populatie te schatten ook een *schatter*.

   (i) Bereken het gemiddelde van de schattingen over alle mogelijke steekproeven (dus de verwachtingswaarde van de schatter). Vergelijk dit met het echte gemiddelde van de populatie.

   (ii) Bepaal de standaardafwijking van deze schatter voor het gemiddelde van de populatie.

   (iii) Bij een alternatieve schatter neem je steekproeven van 2 elementen *zonder* terugleggen. Bepaal weer de verwachtingswaarde en de standaardafwijking van deze schatter, dus het gemiddelde van de steekproefgemiddelden over alle mogelijke steekproeven en de standaardafwijking van de verzameling van alle steekproefgemiddelden.

8. Bij een steekproef van n stukken worden s defecte stukken gevonden, de schatting voor de kans p op een defect stuk is dus $\overline{p} = \frac{s}{n}$. Voor een *gegeven* waarde van p laat zich de kwaliteit van de schatting makkelijk toetsen, omdat in dit geval de standaardafwijking van de verdeling van schattingen (dus de standaardafwijking van de schatter) gegeven is door $\sqrt{\frac{p(1-p)}{n}}$. Maar in veel gevallen is de ware waarde van p onbekend en we moeten onze conclusies alleen uit de steekproef trekken.

(i) Bij een steekproef van 100 stukken werden 20 defecte stukken gevonden. Bepaal de minimale en de maximale waarde van p zo dat de schatting $\overline{p} = 0.2$ binnen één standaardafwijking (van de schatter) van p ligt.

(ii) We noteren de grootste waarde van p waarvoor de schatting $\overline{p}$ nog net binnen één standaardafwijking van p ligt met p_{max}. Geef een formule afhankelijk van p_{max}, $\overline{p}$ en n aan, waar p_{max} aan voldoet.

(Hint: Bepaal een functie van p die p_{max} als nulpunt heeft. Het nulpunt van deze functie kan niet expliciet bepaald worden, maar moet numeriek benaderd worden.)

Geef ook een formule voor de kleinste waarde p_{min} van p aan, waarvoor $\overline{p}$ nog binnen één standaardafwijking van p ligt.

(iii) Stel iemand beweert dat zijn schatting van $\overline{p} = 0.2$ binnen één standaardafwijking van 0.01 van de ware waarde van p ligt. Hoe groot moet zijn steekproef voor deze bewering minstens zijn?

9. Zij X een stochast met de drie mogelijke uitkomsten -1, 0 en 1 en met de kansverdeling $P(X = -1) = P(X = 1) = \frac{1}{2}p$ en $P(X = 0) = 1 - p$ die van een parameter $0 \leq p \leq 1$ afhangt. Zij T_0 de stochast die het aantal 0en in een steekproef van grote n aangeeft, en T_1 de stochast die het aantal 1en aangeeft.

   Laat zien dat de verwachtingswaarden van $\frac{1}{n}(n - T_0)$ en van $\frac{2}{n}T_1$ gelijk aan p zijn.

10. Bij een zeker chemisch proces wordt de afgegeven energie (warmte) gemeten en er wordt verondersteld dat de afgegeven energie door een stochast X met verwachtingswaarde μ en variantie σ^2 wordt beschreven. Bij 10 metingen zijn de volgende resultaten verkregen:

$$\begin{array}{lllll} x_1 = 1244, & x_2 = 1198, & x_3 = 1212, & x_4 = 1235, & x_5 = 1245, \\ x_6 = 1190, & x_7 = 1202, & x_8 = 1220, & x_9 = 1233, & x_{10} = 1208. \end{array}$$

(i) Bepaal het steekproefgemiddelde $\overline{x}$ en de steekproefvariantie s^2 van de metingen.

(ii) In plaats van over alle steekproefwaarden te middelen, zou men ook het gemiddelde van de eerste en de laatste waarde, of het gemiddelde van de waarden x_3 t/m x_8 kunnen nemen. Dit geeft aanleiding tot de schatters $Y := \frac{1}{2}(X_1 + X_{10})$ en $Z := \frac{1}{6}(X_3 + X_4 + X_5 + X_6 + X_7 + X_8)$. Bepaal de schattingen voor het gemiddelde van de aangegeven steekproef met deze twee schatters.

(iii) Laat zien dat voor de schatters Y en Z uit (ii) geldt dat $E[Y] = E[Z] = E[X]$. Bepaal ook de varianties van deze schatters.

(iv) De schatter Y voor de verwachtingswaarde μ van X kunnen we ook voor een algemeen steekproef van grote n definiëren door $Y := \frac{1}{2}(X_1 + X_n)$. Laat zien dat deze schatter Y verwachtingswaarde $E[Y] = \mu$ en variantie $Var(Y) = \frac{\sigma^2}{2}$ heeft.