Statistiek


BENJAMIN DISRAELI SAID THERE ARE THREE KINDS OF LIES:
LIES, DAMN LIES AND STATISTICS.
NINE OUT OF TEN PEOPLE SAY THAT WAS MARK TWAIN.
THERE ARE THREE KINDS OF PUNCH LINES: OBVIOUS, PAINFULLY OBVIOUS,
I THOUGHT THE TWAIN BIT ADDED AN EXTRA LAYER.
MALLETT
5-8

1) Inleiding

a) Wat is statistiek?

Beschrijvende statistiek

De beschrijvende statistiek verzamelt gegevens en beschrijft de toestand door die gegevens te ordenen in tabellen, te verwerken, samen te vatten en grafisch voor te stellen. Ook worden gemiddelden, standaardafwijkingen, vormcoëfficiënten en eventuele correlaties (statistische verbanden) berekend. De gegevens worden dus letterlijk beschreven aan de hand van een beperkt aantal typerende parameters. Dit is het onderwerp van dit eerste hoofdstuk en is eigenlijk een herhaling van wat vroeger reeds werd gezien.

Verklarende (inductieve) statistiek

De verklarende statistiek steunt op de resultaten uit de beschrijvende statistiek en op de kansrekening om op basis van steekproeven uitspraken te doen over de ganse populatie. Dit vormt het tweede deel van deze cursus en hier wordt dieper op in gegaan in universitaire cursussen.

b) Enkele definities

De groep individuen of objecten waarvan we één of meerdere kenmerken willen onderzoeken, noemen we de *populatie*. Een individueel lid van de populatie noemen we een element. Meestal is het onpraktisch of onmogelijk om de gehele populatie te onderwerpen aan een onderzoek. Vaak nemen we daarom een klein gedeelte van de populatie, een *steekproef*. Een kok eet immers ook niet de hele pan soep leeg om uitspraken te doen over de kwaliteit. Wel belangrijk is dat voor het proeven goed wordt geroerd. De eetlepel soep die beoordeeld wordt, moet representatief zijn voor het geheel.

De *variabele* is een eigenschap die bij de elementen van de populatie of steekproef varieert. Vaak worden er bij een steekproef verschillende variabelen gemeten. De *data(set)* is dan de verzameling van gegevens die wordt bekomen door de variabelen te meten. Met de *verdeling* van een variabele geven we aan welke waarden worden aangenomen en hoe vaak.

c) De meetschaal van variabelen

Nominale schaal

Bij de nominale schaal worden de waarden van de variabele gebruikt voor identificatie zonder dat ze een hoeveelheid aanduiden.

Voorbeelden:

- Variabele: Geslacht. Waarden: man, vrouw of andere.
- Variabele: Land van herkomst. Waarden: België, Frankrijk, Spanje, …

Bij deze voorbeelden is het duidelijk dat de waarde van de variabele geen numerieke betekenis heeft, het zijn zelfs geen getallen (man, vrouw, België, etc.). Echter ook nominale variabelen die numerieke waarden aannemen zonder dat het getal zelf een specifieke betekenis heeft (het wordt louter gebruikt ter identificatie).

Voorbeelden:

- Variabele: Rekeningnummer. Waarden: 37 0000 0000 2828 (Oxfam), 73 0000 0000 6060 (Artsen Zonder Grenzen), …
- Variabele: Rugnummer in het voetbal. Waarden: 1, 2, 10, 31, etc.

Ordinale schaal

De ordinale schaal erft alle eigenschappen van de nominale schaal samen met een extra eigenschap: de waarden duiden een volgorde aan. Behalve om een volgorde weer te geven, is de waarde van de variabele niet van belang.

Voorbeelden:

- Variabele: Uitslag wedstrijd. Waarden: goud, zilver of brons.
- Variabele: Officiersgraad. Waarden: Onderluitenant, Luitenant, Kapitein, …
- Variabele: Mate van instemming met een bepaalde uitspraak. Waarden: volledig mee oneens, mee oneens, neutraal, mee eens, volledig mee eens.

Intervalschaal

De intervalschaal erft alle eigenschappen van de ordinale schaal met de extra eigenschap dat verschillen tussen waarden een betekenis hebben. Er is echter geen absoluut nulpunt.

Voorbeelden:

- Variabele: Temperatuur in graden Celsius. Waarden: 0, 10, -30, ...
- Variabele: IQ. Waarden: 96, 100, 130, …

Niettegenstaande de temperatuur de waarde 0 kan aannemen, is het geen absoluut nulpunt omdat 0 °C niet willen zeggen dat er geen temperatuur aanwezig is. Het is de aanduiding van een bepaalde temperatuur. Een andere manier om te zien dat het nulpunt bij temperatuur niet absoluut is, is door een andere meeteenheid te gebruiken. Nul graden Celsius komt overeen met 32 graden Fahrenheit. Dit geeft aan dat het nulpunt hier relatief is: ze hangt af van de gebruikte meeteenheid (Celsius of Fahrenheit).

Variabelen gemeten op intervalschaal worden ook intervalvariabelen genoemd. De benaming interval drukt uit dat gelijke verschillen op de meetschaal (intervallen) duiden op gelijke verschillen in de variabele. Een stijging van 10 °C naar 20 °C is evenveel als een stijging van 20 °C naar 30 °C in termen van het warmteverschil. Bij een ordinale variabele is dit niet het geval: bv. het verschil in uitslag tussen de eerste (gouden medaille) en de tweede (zilveren medaille) hoeft niet gelijk te zijn aan het verschil in uitslag tussen de tweede en de derde (bronzen medaille).

Ratioschaal

De ratioschaal erft alle eigenschappen van de intervalschaal en heeft daarbij ook een absoluut nulpunt.

Voorbeelden:

- Variabele: Lengte in cm. Waarden: 0, 1, 354, ...
- Variabele: Geldbedrag in euro. Waarden: 0, 5, 2400, ...
- Variabele: Reactietijd tussen stimulus en gedrag in seconden. Waarden: 0, 5, 2, 58, …

Hier is het nulpunt absoluut: als je 0 euro hebt, wil dit zeggen dat je geen geld hebt. Een lengte van 0 centimeter blijft 0, ook als je de schaal omzet naar meter. Een reactietijd van 0 seconden wil zeggen dat er geen tijd zit tussen stimulus en gedrag.

Variabelen gemeten op ratioschaal worden ook ratiovariabelen genoemd. De benaming ratio drukt uit dat verhoudingen (ratio's) een betekenis hebben: 10 euro is bijvoorbeeld dubbel zoveel als 5 euro. Bij intervalvariabelen zijn verhoudingen niet zinnig: een temperatuur van 10 °C is niet dubbel zo warm 5 °C omdat deze uitspraak bij een omzetting naar Fahrenheit (5 °C = 41 °F en 10 °C = 50 °F) niet meer opgaat.

Discrete en continue variabelen

Naast de vier schaalfamilies kunnen we ook een onderscheid maken tussen *continue en discrete variabelen*.

Continue variabelen kunnen tussenwaarden aannemen. Met tussenwaarde bedoelen we dat er tussen elke twee willekeurige waarden een derde waarde ligt. Dit impliceert dat er tussen twee waarden (theoretisch gezien) oneindig veel andere waarden kunnen liggen.

We zeggen ook dat variabelen gemeten op continue schaal continu variëren.

Voorbeelden:

- Lengte in cm: tussen 2 en 3 cm liggen nog vele andere waarden: vb. 2,5 cm, 2,34 cm, 2,32348 cm, …
- Temperatuur in °C: tussen twee willekeurige temperaturen ligt altijd een derde.
- De tijd in seconden: tussen twee willekeurige tijdstippen ligt altijd een derde tijdstip.

Bij discrete variabelen bestaan steeds twee waarden waar geen derde waarde kan tussen liggen. Dit impliceert dat de variabele maar een eindig aantal waarden kan aannemen.

Voorbeelden:

- Het aantal kinderen. Tussen 0 en 1 kind ligt geen derde waarde: het is niet mogelijk om 1,5 kinderen te hebben.
- Het aantal keer dat `munt' wordt geworpen bij 4 worpen met een geldstuk. Het is niet mogelijk om 1,5 keer munt te werpen, ze kan slechts de waarden 0, 1, 2, 3, of 4 aannemen.
- Het aantal volgers op Twitter. Barack Obama kan bijvoorbeeld 63379938 of 63379939 volgers hebben, maar geen 63379938,5.

Het aantal volgers op Twitter is theoretisch gezien een discrete variabele, maar omdat deze variabele zéér veel verschillende waarden kan aannemen, wordt er ook gezegd dat ze bijna-continu is. In de praktijk zullen we variabelen die bijna-continu zijn, als continu beschouwen.

d) Steekproeven

Kenmerken van een (goede) steekproef:

- De steekproef moet *representatief* zijn.

  Dat wil zeggen dat de steekproef een correct beeld moet geven van de verscheidenheid binnen de populatie, dus dat in de steekproef alle deelverzamelingen van de populatie evenredig vertegenwoordigd moeten zijn (we noemen dit ook wel gestratificeerd).

  Ook moet de omvang van de steekproef voldoende groot zijn.

- De steekproef moet *aselect* zijn.

  Dat betekent dat elk element van de populatie dezelfde kans heeft om opgenomen te worden in de steekproef. Men spreekt vaak van *enkelvoudig aselecte steekproeven* (*EAS*) in deze context (enkelvoudig duidt dan op het feit dat elk individu of object maar één keer in de steekproef kan zitten).

Fouten die niet te wijten zijn aan de steekproef bestaan uiteraard ook. Zo bestaan er non-responsfouten als mensen niet willen meewerken aan een enquête. Ook responsfouten zijn mogelijk door bijvoorbeeld slechte communicatie of leugens.

Een typisch voorbeeld van een niet-representatieve steekproef treedt op bij het zogenaamde *convenience sampling*. Dit is je steekproef zo organiseren dat het gemak van de onderzoeker voorop staat. Voorbeelden zijn straatenquêtes, telefonische enquêtes, enz.

We noemen een steekproef *gestratificeerd* als de populatie eerst wordt onderverdeeld in unieke, homogene segmenten (strata) en vervolgens uit elk segment (stratum) een enkelvoudige aselecte steekproef wordt getrokken. Veelgebruikte strata zijn bijvoorbeeld gender en leeftijd. De geselecteerde monsters uit de verschillende strata worden samengevoegd tot één steekproef.

2) Univariate beschrijvende statistiek

Met univariate statistiek bedoelen we dat we per element in de steekproef (of populatie) maar één kenmerk of variabele onderzoeken. We beperken ons tot variabelen die gemeten worden op de ratio-schaal.

a) Frequentietabellen

Een dataset kan heel groot zijn en daardoor zeer onoverzichtelijk. Een eerste verwerking van de gegevens kan erin bestaan de gegevens voor te stellen in een frequentietabel. Daarbij moeten we de keuze maken of we gegevens gaan groeperen in klassen of niet.

We bekijken twee inleidende voorbeelden.

Niet-gegroepeerde gegevens

We tellen bij 25 gezinnen het aantal kinderen en we verkrijgen volgende data (gegevens):

0 1 2 0 1 0 4 3 3 5 1 1 2 2 2 2 1 3 0 4 2 1 5 2 0

Om meer overzicht te krijgen, kunnen we de waarnemingsgetallen ordenen:

0 0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 4 4 5 5

Nog overzichtelijker is een frequentietabel:

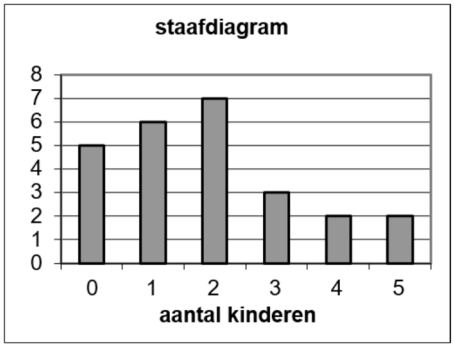
waarnemings-getallen	enkelvoudige frequenties			cumulatieve frequenties		
	absolute frequentie	relatieve frequentie	procentuele frequentie	absolute frequentie	relatieve frequentie	procentuele frequentie
x_i	n_i	f_i		cn_i	cf_i	
0	5	0,2	20%	5	0,2	20%
1	6	0,24	24%	11	0,44	44%
2	7	0,28	28%	18	0,72	72%
3	3	0,12	12%	21	0,84	84%
4	2	0,08	8%	23	0,92	92%
5	2	0,08	8%	25	1	100%

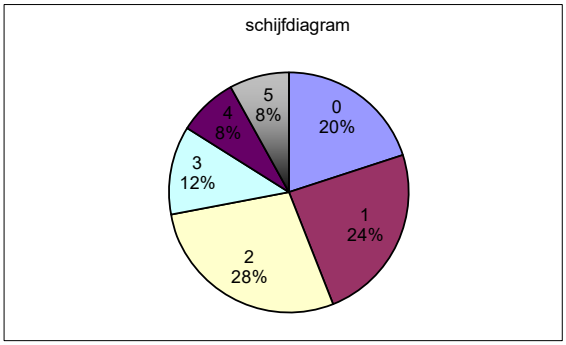
De *enkelvoudige absolute frequentie* n_i is het aantal keer dat een waarnemingsgetal x_i voorkomt.

De *cumulatieve absolute frequentie* cn_i is het aantal waarnemingsgetallen kleiner of gelijk aan x_i.

De *relatieve frequenties* geven de verhouding van de absolute frequenties tot de omvang n van de steekproef of populatie weer. Dus: $f_i = \frac{n_i}{n}$ en $cf_i = \frac{cn_i}{n}$. Dit worden ook vaak *proporties* genoemd.

De relatieve frequenties worden ook vaak in procenten weergegeven. Dit zijn dan *procentuele frequenties*. Grafisch kunnen deze resultaten weergegeven worden in een staaf- en taartdiagram:





Gegroepeerde gegevens

We bepalen de lichaamslengte in cm van 100 16-jarige jongens, afgerond op de eenheid:

175	168	177	167	176	167	172	166
173	172	170	186	168	180	165	159
155	179	184	155	188	163	156	172
161	162	174	159	162	169	171	179
170	165	157	168	167	166	172	174
183	173	168	150	182	154	160	159
189	153	162	166	157	179	164	169
165	193	154	180	171	168	180	181
171	176	165	176	172	169	161	167
159	169	176	185	176	164	169	166
160	164	163	170	158	168	175	173
165	165	166	183	164	167	159	180
158	163	169	177				

Door de omvang van deze gegevens zou een niet-gegroepeerde frequentietabel zoals hierboven zeer onoverzichtelijk zijn. Daarom kiezen we er bij dit soort datasets vaak voor om de gegevens te groeperen in *klassen*. We doen dit op zo'n manier dat:

- elk waarnemingsgetal tot precies één klasse behoort;
- elke klasse vertegenwoordigd wordt door het *klassenmidden*;

Voor het bepalen van de *klassebreedte*, berekenen we de *variatiebreedte* van de steekproef:

$$VB = x_{\max} - x_{\min} = 193 - 150 = 43$$

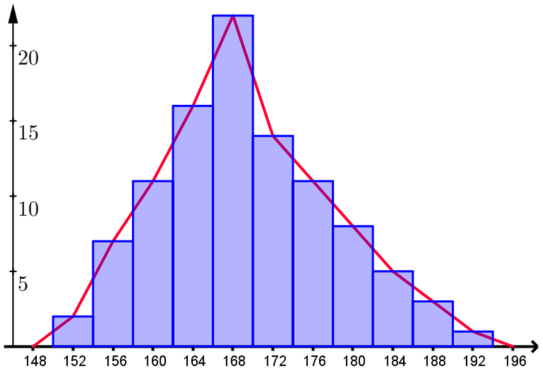
Deel de variatiebreedte door het gewenste aantal klassen (bvb. 10) en rond af: . $43/10 = 4{,}3 \approx 4$ ..

Een vuistregel die soms ook gehanteerd wordt om het aantal klassen te bepalen is de vierkantswortel te nemen van het aantal gegevens. Hier namen we 10 klassen omdat we 100 gegevens hebben.

Klasse	Midden x_i	n_i	cn_i	f_i	cf_i
[150 ,154 [	152	2	2	2%	2%
[154 ,158 [	156	7	9	7%	9%
[158 ,162 [	160	11	20	11%	20%
[162 ,166 [	164	16	36	16%	36%
[166 ,170 [	168	22	58	22%	58%
[170 ,174 [	172	14	72	14%	72%
[174 ,178 [	176	11	83	11%	83%
[178 ,182 [	180	8	91	8%	91%
[182 ,186 [	184	5	96	5%	96%
[186 ,190 [	188	3	99	3%	99%
[190 ,194 [	192	1	100	1%	100%

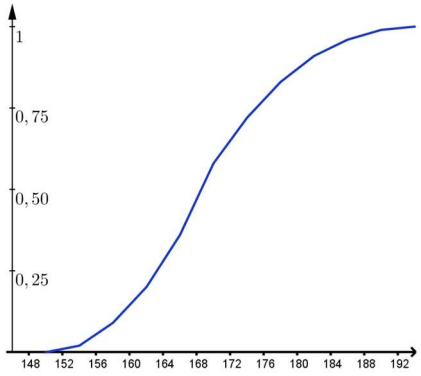
Deze tabel kan grafisch worden voorgesteld met een *histogram*. Dit is een speciaal soort staafdiagram waarbij de hoogte van de rechthoekjes zo is bepaald dat de oppervlaktes ervan recht evenredig zijn met de frequenties van de klasse. Als alle klassen even breed zijn dan kan voor de hoogte zowel de absolute als de relatieve frequentie genomen worden.

We noemen het histogram *genormaliseerd* als de hoogte van de rechthoekjes zo gekozen is dat de totale oppervlakte van het histogram 1 is (de hoogte van de balkjes is dan de relatieve frequentie gedeeld door de klassebreedte).



Op de figuur hiernaast zie je naast het histogram ook in het rood het enkelvoudig frequentiepolygoon getekend. Dit verbindt de punten met als x-waarde de klassemiddens en als y-waarde de hoogte van de rechthoekjes (frequentie).
Bij een genormaliseerd histogram zal ook de oppervlakte onder deze kromme 1 zijn. We noemen dit dan een *dichtheidskromme*.

Op de figuur hiernaast is het *ogief* getekend dat bij deze data hoort. Dit is het cumulatieve frequentiepolygoon. Dit verbindt de punten met als x-waarde de rechtergrenzen van de klassen en als y-waarden de cumulatieve frequentie (hier relatief, maar dat mag ook absoluut zijn).
Op het ogief lees je dus verticaal af hoeveel gegevens kleiner of gelijk zijn aan de gegeven horizontale waarde.



b) Centrummaten

Centrummaten zijn getallen die kenmerkend zijn voor de centrale ligging van de waarnemingsgetallen. De bekendste zijn het (rekenkundig) gemiddelde, de mediaan en de modus.

Het gemiddelde

Het (rekenkundig) gemiddelde $\overline{x}$ van een groep waarnemingsgetallen x_i is de som van alle waarnemingsgetallen gedeeld door hun aantal.

In formulevorm geeft dit $\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ of voor gegroepeerde gegevens: $\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{p} n_i x_i$.

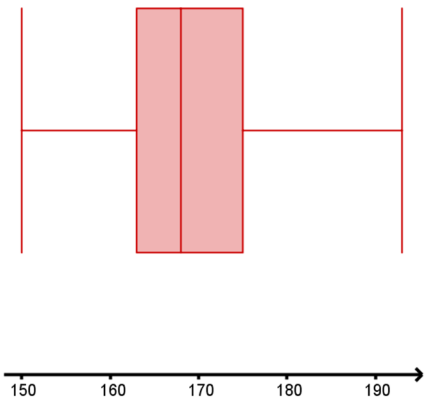
Het gemiddelde van een populatie wordt ook vaak met de Griekse letter μ aangeduid.

Alle gegevens zijn betrokken bij de berekening en hebben dus invloed op de grootte van het rekenkundig gemiddelde. Het nadeel is dat uitbijters en uitschieters (extreem kleine of grote waarnemingsgetallen) het rekenkundig gemiddelde enorm beïnvloeden.

Mediaan en kwartielen - boxplot

De mediaan Me van een groep getallen is het middelste getal als deze in volgorde van grootte zijn gerangschikt. Als het aantal even is dan neem je het gemiddelde van de twee middelste getallen.

Het eerste kwartiel Q_1 van een groep getallen is de mediaan van de eerste helft van de gegevens. Het derde kwartiel Q_3 van een groep getallen is de mediaan van de tweede helft van de gegevens. Bij een oneven aantal gegevens laat je in beide helften de mediaan weg.
Een boxplot is een grafische weergave waarop je het kleinste en het grootste gegeven kan aflezen naast de kwartielen en de mediaan.
Op de figuur hiernaast zie je de boxplot getekend die hoort bij het tweede voorbeeld uit de vorige paragraaf.



Modus

De *modus* Mo van een groep waarnemingsgetallen is het getal met de grootste absolute frequentie of het klassenmidden van de klasse met de grootste enkelvoudige frequentie.

c) Spreidingsmaten

Spreidingsmaten van een groep waarnemingsgetallen zijn getallen die aangeven of de waarnemingen dicht bij elkaar liggen of net eerder verspreid zijn. We bekijken er enkele:

De variatiebreedte en de interkwartielafstand

De *variatiebreedte* is het verschil tussen de grootste en de kleinste waarneming.

De *interkwartielafstand* (IKA) is het verschil tussen het derde en het eerste kwartiel $Q_3 - Q_1$. Dit is veel minder afhankelijk van extreme waarden dan de variatiebreedte.

De interkwartielafstand wordt vaak gebruikt om uitbijters en uitschieters preciezer te definiëren:

Een *uitschieter* x_i is een waarnemingsgetal dat groter is dan het derde kwartiel vermeerderd met anderhalf keer de interkwartielafstand. Dus x_i is een uitschieter $\Leftrightarrow x_i > Q_3 + 1{,}5 \cdot (Q_3 - Q_1)$.

Een *uitbijter* x_i is een waarnemingsgetal dat kleiner is dan het eerste kwartiel verminderd met anderhalf keer de interkwartielafstand. Dus x_i is een uitbijter $\Leftrightarrow x_i < Q_1 - 1{,}5 \cdot (Q_3 - Q_1)$.

In het Engels worden beide extreme waarden samengevat onder de noemer *outliers*. Deze kunnen perfect te verklaren zijn en realistisch zijn, maar kunnen ook het gevolg zijn van meetfouten of schrijffouten.

Het gemiddelde en de variantie (en standaardafwijking – zie hieronder) zijn maten die heel gevoelig zijn voor outliers. De mediaan en de interkwartielafstand daarentegen zijn hieraan niet gevoelig, we noemen ze *resistente maten*.

Variantie en standaardafwijking

De *variantie* σ^2 van een groep gegevens is de gemiddelde kwadratische afwijking van de gegevens t.o.v. het rekenkundig gemiddelde. De *standaardafwijking* is de positieve vierkantswortel hieruit.

In formulevorm is dit: $\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$ of voor gegroepeerde gegevens: $\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{p} n_i (x_i - \overline{x})^2$.

De *standaardafwijking* σ is dan uiteraard de vierkantswortel van de variantie: $\sigma = \sqrt{\sigma^2}$.

Deze twee spreidingsmaten zijn enorm afhankelijk van extreme meetwaarden.

Opmerking: Als het om populaties gaat worden het gemiddelde en de standaardafwijking met de Griekse letters μ en σ (mu en sigma) genoteerd.

Als het om steekproeven gaat noteren we het zoals gedefinieerd met $\overline{x}$ en s. In de formule wordt er bij s^2 dan soms gedeeld door $n-1$ in plaats van n omdat we later zullen zien dat dit een betere schatter is voor de standaardafwijking σ van de populatie waaruit de steekproef werd genomen.

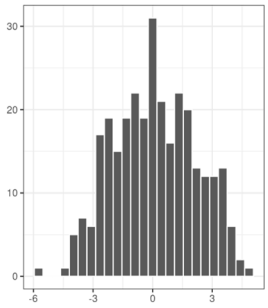
d) Kenmerken van verdelingen

Modaliteit

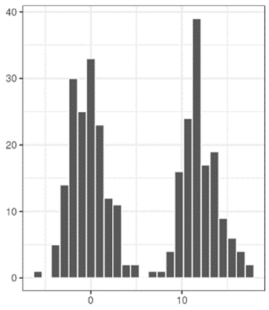
Modaliteit of *toppigheid* beschrijft het aantal toppen van een verdeling. Het begrip is uiteraard afgeleid van de modus, de meest voorkomende waarde in een datareeks. De modus vormt een top in de verdeling van de data. Verdelingen kunnen één of meerdere toppen hebben.

Een verdeling met één top wordt *unimodaal* of *eentoppig* genoemd, en een verdeling met meer toppen wordt *multimodaal* of *meertoppig* genoemd. Hier zijn specifieke varianten van zoals de *bimodale* of *tweetoppige* verdeling.

Een voorbeeld van een bimodale verdeling is de lengte zijn van de Belgische populatie, omdat mannen gemiddeld gezien groter zijn dan vrouwen. De verdeling vertoont twee toppen.



Een unimodale verdeling

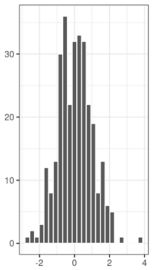


Een bimodale verdeling

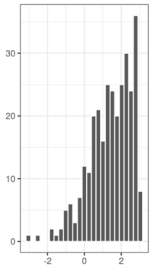
Scheefheid

Scheefheid (Engels: *skewness*), beschrijft of een verdeling symmetrisch of asymmetrisch is. Een scheve verdeling is asymmetrisch. Dit betekent dat de meeste datapunten aan één kant van de schaal liggen.

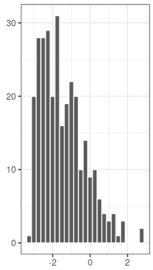
Een (eentoppige) verdeling kan *symmetrisch, linksscheef (negatief scheef)* of *rechtsscheef (positief scheef)* zijn.



Een symmetrische verdeling



Een linksscheve verdeling



Een rechtsscheve verdeling

In een symmetrische (unimodale) verdeling liggen de meeste datapunten rondom het gemiddelde en zijn er steeds minder datapunten naarmate de afstand tot het gemiddelde toeneemt. De verdeling heeft de vorm van een antieke mantelklok en wordt in het Engels vaak aangeduid met *bell curve*.

Bij een linksscheve verdeling liggen er minder datapunten aan de linkerkant van het gemiddelde. De meeste datapunten liggen dus aan de rechterkant en er is een staart met datapunten relatief ver weg van het gemiddelde aan de linkerkant. Bij een rechtsscheve verdeling liggen juist de meeste datapunten links van het gemiddelde. Rechts van het gemiddelde liggen minder datapunten in een staart.

Bij linksscheve verdelingen is het gemiddelde kleiner dan de mediaan, bij rechtsscheve verdeling is het gemiddelde groter dan de mediaan.

3) Bivariate beschrijvende statistiek

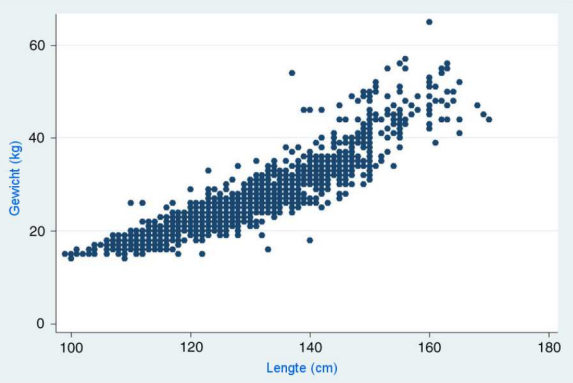
In veel wetenschappelijke studies beperkt men zich niet tot één variabel kenmerk van een populatie, maar meerdere. In dit hoofdstuk gaan we dieper in op hoe we associaties tussen 2 variabelen kunnen beschrijven.

a) Samenvattingsmaten voor ratio variabelen

Spreidingsdiagrammen

Een eenvoudige manier om data die van een populatie twee kenmerken meet grafisch weer te geven is een spreidingsdiagram. Elk meting wordt weergegeven als een punt in een assenstelsel. Dit wordt ook wel eens een puntenwolk of scatterplot genoemd.

Op de figuur hiernaast zie je van een groot aantal kinderen uit een lagere school hun lengte en gewicht tegen elkaar uitgezet. Wat je hier direct kan uit afleiden is een positief verband tussen beide, dus dat kinderen die langer zijn waarschijnlijk ook meer wegen.



Covariantie

De *covariantie* is een maat voor de spreiding van twee variabelen die aangeeft in welke mate twee variabelen met elkaar samenhangen, ze geeft meer specifiek aan in welke richting ze elkaar beïnvloeden.

In een populatie van grootte n waarin van elk element twee kenmerken x en y worden gemeten (elk element geeft dus een koppel waarnemingsgetallen (x_i, y_i)), wordt de covariantie berekend als volgt:

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Net als bij de variantie berekenen we de covariantie bij een steekproef door te delen door $n-1$.

Enkele eigenschappen die onmiddellijk duidelijk zijn:

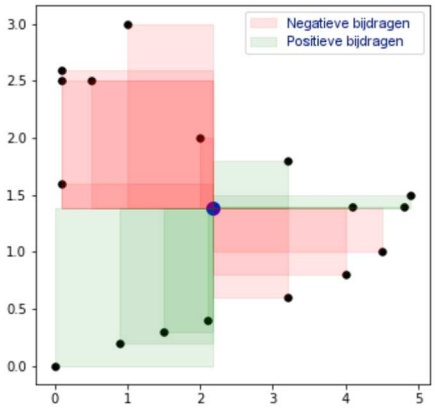
- $Cov(x,y) = Cov(y,x)$
- $Cov(x,x) = \sigma_x^2$

We kunnen de variantie van een variabele dus beschouwen als de covariantie ervan met zichzelf.

Grafische interpretatie:

Hiernaast staat een spreidingsdiagram met van een aantal drugsgebruikers hoeveel cocaïne (x) en heroïne (y) ze gebruiken in gram per week.

De dikke(re) stip in het midden geeft het gemiddelde $M(\overline{x}, \overline{y})$ aan. Als we dit punt verbinden met een ander punt waarnemingsgetallen $P(x_i, y_i)$ dan geeft $(x_i - \overline{x})(y_i - \overline{y})$ de georiënteerde oppervlakte aan van een rechthoek met als diagonaal het lijnstuk dat de twee punten M en P verbindt. De oppervlakte wordt negatief gerekend als het lijnstuk dalend is, en positief als het lijnstuk stijgend is.



Bij een gemiddeld positief verband tussen variabelen zullen er meer positieve bijdragen zijn in de formule dan negatieve en omgekeerd. De covariantie is dus wel degelijk een maat voor de richting waarin twee variabelen elkaar beïnvloeden.

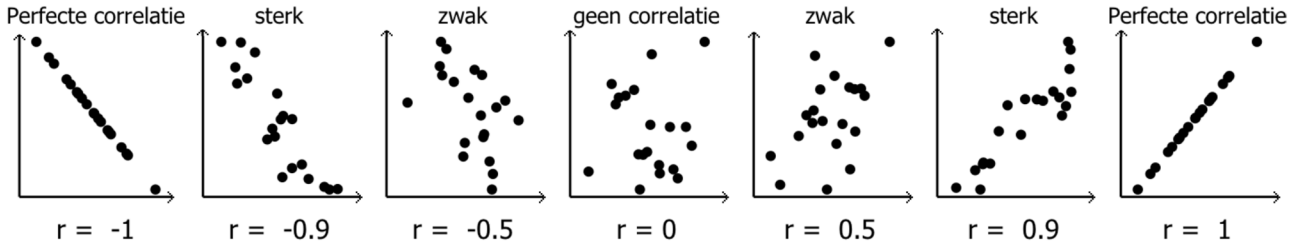
Correlatie

Het probleem van de covariantie is dat ze afhankelijk is van de gekozen eenheden. Als we in het voorbeeld hierboven het drugsgebruik in milligram zouden weergegeven hebben dan zou de covariantie een miljoen keer groter geweest zijn. Om dit probleem te verhelpen definiëren we de *correlatie*: dit is de covariantie gedeeld door de standaardafwijkingen van de beide variabelen.

In formulevorm $r_{xy} = \dfrac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$.

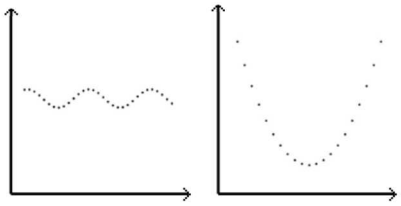
De correlatie is een maat voor lineaire samenhang tussen twee variabelen die altijd tussen -1 (perfect negatief lineair verband) en 1 (perfect positief lineair verband) ligt. Het bewijs hiervan ligt echter buiten het bestek van deze cursus. Het is nu dan ook veel belangrijker om correlatie juist te interpreteren.

Enkele voorbeelden van spreidingsdiagrammen en hun bijhorende correlaties:



Vanaf $|r_{xy}| \geq 0{,}85$ spreken we van een sterke correlatie, $|r_{xy}| \leq 0{,}5$ noemen we een zwakke correlatie. De waarden daartussen noemen we matige correlatie.

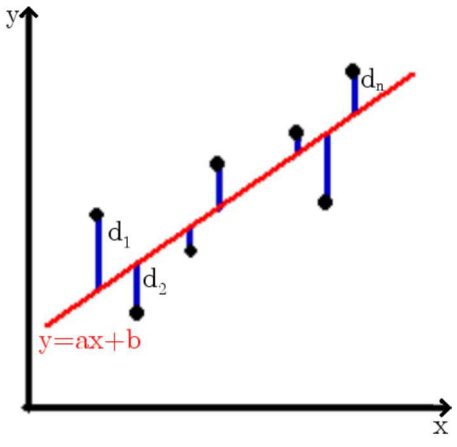
Je mag echter niet vergeten dat correlatie een maat is voor lineair verband. Het is niet omdat de correlatie dicht bij 0 ligt dat er geen sprake is van samenhang. Dat kan je zien aan de hiernaast staande spreidingsdiagrammen die beide correlatie 0 hebben, terwijl er wel duidelijke verbanden zijn (links sinusoïdaal en rechts parabolisch).



Lineaire regressie

Bij een lineaire regressie wordt er gezocht naar de vergelijking van de best passende rechte bij de puntenwolk. We noemen dit dan de regressierechte. Dit kan op verschillende manieren geïnterpreteerd worden.

Stel dat de best passende rechte als vergelijking $y = ax + b$ heeft. Voor elk koppel waarnemingsgetallen (x_i, y_i) kunnen we dan het residu d_i berekenen, de verticale afstand van het waargenomen punt tot de regressierechte. Dus $d_i = |y_i - (ax_i + b)|$. Eén methode om de best passende rechte te vinden bestaat erin de som van de kwadraten van deze residuen zo klein mogelijk te maken.



Vandaar dat we deze methode ook wel eens de *kleinste kwadraten regressie* noemen. We doen nu een klein uitstapje langs de wereld van de afgeleiden om de parameters a en b te berekenen.

We willen de som van de kwadraten van de residuen minimaliseren, we noemen dit $SSR = \sum_{i=1}^{n} d_i^2$.

(SSR staat voor de sum of squares of residues).

Uit de analyse weten we dat bij een minimum de afgeleiden nul zijn, waarbij we afleiden naar a en b.

$$\frac{dSSR}{db}=0 \Leftrightarrow \frac{d\sum_{i=1}^{n} d_i^2}{db}=0 \Leftrightarrow \frac{d\sum_{i=1}^{n}\left(y_i-(ax_i+b)\right)^2}{db}=0 \Leftrightarrow -2\cdot\sum_{i=1}^{n}\left(y_i-(ax_i+b)\right)=0$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b = 0 \Leftrightarrow n\cdot b = \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i \Leftrightarrow b=\frac{1}{n}\sum_{i=1}^{n} y_i - a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\Leftrightarrow b=\overline{y}-a\cdot\overline{x}$$

Let erop dat dit eigenlijk gewoon impliceert dat $\overline{y}=a\cdot\overline{x}+b$ dus dat het "gemiddelde punt" $(\overline{x},\overline{y})$ op de regressierechte ligt.

Voor de residuen geldt dan: $d_i=y_i-(ax_i+b)=y_i-(ax_i+\overline{y}-a\cdot\overline{x})=y_i-\overline{y}-a\cdot(x_i-\overline{x})$.

We leiden nu de som van de kwadraten van de residuen ook nog eens af naar a.

$$\frac{dSSR}{da}=0 \Leftrightarrow \frac{d\sum_{i=1}^{n} d_i^2}{da}=0 \Leftrightarrow \frac{d\sum_{i=1}^{n}\left(y_i-\overline{y}-a\cdot(x_i-\overline{x})\right)^2}{da}=0 \Leftrightarrow -2\cdot\sum_{i=1}^{n}\left(y_i-\overline{y}-a\cdot(x_i-\overline{x})\right)(x_i-\overline{x})=0$$

$$\Leftrightarrow \sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})-a\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2 \Leftrightarrow a=\frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\cdot\frac{\frac{1}{n-1}}{\frac{1}{n-1}}=\frac{Cov(x,y)}{s_x^2}$$

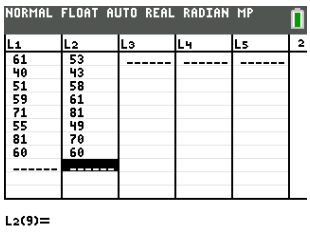
Merk op dat we dit ook kunnen schrijven als $a=\frac{r_{xy}\cdot s_y}{s_x}$. We gebruiken hier $n-1$ i.p.v. n (en s i.p.v. σ) omdat regressies meestal gebeuren bij steekproeven en minder bij populaties. De formules kunnen eenvoudig aangepast naar populaties.

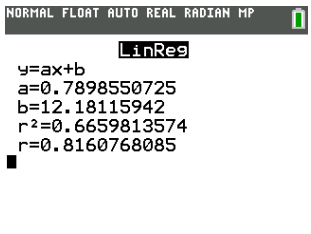
Een uitgewerkt voorbeeld (met de grafische rekenmachine)

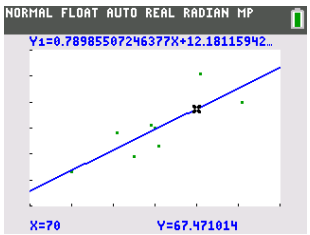
In een klas met 8 leerlingen krijg je de punten op het examen wiskunde en fysica gegeven.

Wiskunde	61	40	51	59	71	55	81	60
Fysica	53	43	58	61	81	49	70	60

a) Teken de puntenwolk, bereken de correlatie en bereken en teken de regressierechte.
b) Een nieuwe leerling in de klas scoort 70% op zijn examen van wiskunde. Hoeveel verwacht je volgens dit model dat hij zal scoren op zijn examen van fysica?







We lezen af dat de correlatie $r_{xy}=0{,}816$ (matige correlatie).

Als iemand 70% scoort op zijn examen wiskunde verwachten we (volgend dit model) 67,5% op fysica.

De waarde van r^2 (hier $r^2\approx 0{,}666$) noemen we de *determinatiecoëfficiënt* van het model. Het is een maatstaf voor de aansluiting van het model bij de daadwerkelijke uitkomst (*goodness of fit*). Het is de proportie van variantie in de afhankelijke variabele die wordt verklaard door het model. De verklaring hiervoor valt buiten het bestek van deze cursus.

b) Samenvattingsmaten voor categorische variabelen

De samenvattingsmaten uit de vorige hoofdstukken (covariantie en correlatie, maar ook gemiddelde, mediaan, standaardafwijking, …) kunnen niet zomaar toegepast worden voor de beschrijving van categorische variabelen. Daarvoor zijn andere maten nodig. We bespreken er hier enkele.

Absolute en relatieve risicoreductie

Het *relatieve risico*, ook wel *risk ratio* genoemd, is een bekende associatiemaat voor variabelen met twee uitkomstcategorieën. De maat drukt de sterkte uit van het verband tussen een determinant (een beïnvloedende factor) en een uitkomst, door de kansen op de uitkomst met en zonder de determinant te delen door elkaar.

De *absolute risicoreductie* is het verschil tussen de kansen op de uitkomst met en zonder determinant.

Voorbeeld: het universitair ziekenhuis in Gent doet een grootschalig, placebogecontroleerd klinisch onderzoek naar het effect van langdurig statine gebruik op de reductie van sterfte bij mensen die lijden aan angina pectoris. Het onderzoek laat de hiernaast staande resultaten zien.

Data

	gestorven	overleefd	
statine	**182**	**2039**	*2221*
placebo	**256**	**1967**	*2223*
	438	*4006*	*4444*

De sterfte in de statinegroep wordt gegeven door $\frac{182}{2221} = 8,2\%$, in de placebogroep is dit $\frac{256}{2223} = 11,5\%$.

De absolute risicoreductie is dus 3,3%. Het relatieve risico is $\frac{8,2\%}{11,5\%} = 0,71$, zodat de relatieve risicoreductie gelijk is aan 29%.

De interpretatie van deze maten gaat als volgt: het gebruik van statine reduceert het sterfterisico met een factor $0,71$, of dus met 29%. Door 100 patiënten te behandelen met statine spaar je 3,3 levens.

De waarden van het relatieve risico liggen tussen 0 (maximale negatieve associatie) en +oneindig (maximale positieve associatie), terwijl bij het ontbreken van associatie het relatieve risico gelijk is aan 1.

Deze maten worden veel gebruikt bij onderzoeken waar er sprake is van een prospectief studieopzet. Dit wil zeggen dat je op voorhand kan kiezen hoeveel mensen je laat deelnemen en wie je laat beïnvloeden door de determinant of niet.

Odds en oddsratio

Bij retrospectieve onderzoeken wordt in plaats van de voorgaande associatiemaat vaker de oddsratio gebruikt.

Net als bij het relatieve risico is de oddsratio ook een quotient van verhoudingen. Je gaat nu echter geen proporties delen maar odds. Bij odds deel je de kans dat een gebeurtenis optreedt door de kans dat ze niet optreedt. Je ziet dit bijvoorbeeld vaak bij weddenschappen. Een kans van 2/3 om te winnen vertaalt zich naar een odds van 2:1.

Voorbeeld: na een restaurantbezoek blijken veel klanten ziek te worden. Er is een vermoeden dat dit ligt aan het ijsdessert dat niet voldoende gekoeld was. Het restaurant contacteerde al zijn klanten en vond de hiernaast staande data.

Data

	ziek	gezond	
ijs	**17**	**23**	*40*
geen ijs	**13**	**32**	*45*
	30	*55*	*85*

De odds bij mensen die ijs aten om ziek te zijn is $17/23$, bij mensen die geen ijs aten is dit $13/32$. De oddsratio is dus $\frac{17/23}{13/32} \approx 1,82$. De odds om ziek te zijn na het eten van het ijsdessert is dus bijna twee keer zo groot als wanneer je geen ijs at.

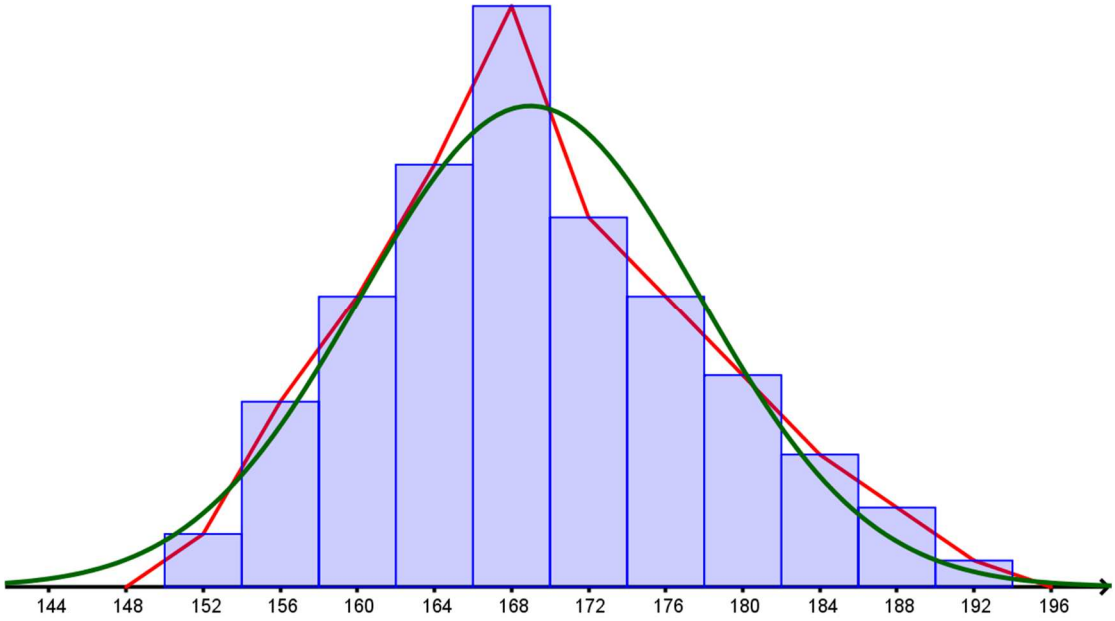
4) De normale verdeling

a) Voorschrift

Als we het voorbeeld van de lichaamslengte uit hoofdstuk 2 herbekijken dan zien we dat de gegevens zich vrij symmetrisch verdelen rond de gemiddelde waarde van $\overline{x} = 168{,}97$ (cm). Het (enkelvoudig) frequentiepolygoon ziet er een beetje uit als een klok. De oppervlakte onder deze kromme is 1. We noemden dat dan reeds een dichtheidskromme. Je kan de oppervlakte binnen een interval onder deze kromme dus interpreteren als de kans dat een waarnemingsgetal binnen dit interval zal liggen.

Carl Friedrich Gauss bewees dat het voorschrift van zo een symmetrische klokvormige dichtheidskromme (met gemiddelde μ en standaardafwijking σ) kan geschreven worden als:

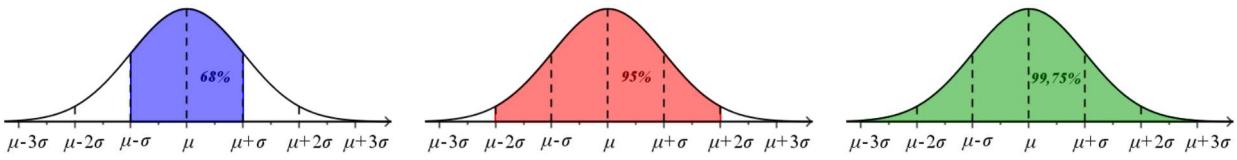
$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{\frac{(x-\mu)^2}{2.\sigma^2}}.$$



Van heel veel gegevens uit de realiteit is geweten dat ze min of meer normaal verdeeld zijn: het geboortegewicht, het IQ, de hoofdomtrek, ... Maar je mag zeker niet denken dat alles normaal verdeeld is. Zo zijn bijvoorbeeld de snelheid bij overtredingen in de bebouwde kom of het gezinsinkomen zeker niet normaal verdeeld.

b) De vuistregel

Bij een normale verdeling zal altijd 68% van de gegevens in het interval $[\mu - \sigma, \mu + \sigma]$ liggen, 95% zal in het interval $[\mu - 2\sigma, \mu + 2\sigma]$ liggen en 99,75% zal in het interval $[\mu - 3\sigma, \mu + 3\sigma]$ liggen. We noemen dit de vuistregel van een normale verdeling. Het geeft ons een criterium om na te rekenen of gegevens al dan niet normaal verdeeld zijn. Grafisch geeft dit het volgende:

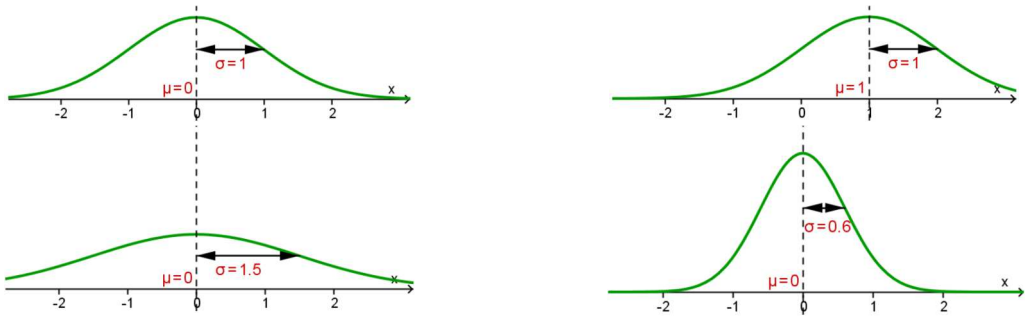


c) De grafiek van de normale verdelingsfunctie

Een normale verdeling hangt dus enkel af van twee parameters: het gemiddelde μ en de standaardafwijking σ. We noteren normale verdelingen vaak korter als $N(\mu,\sigma)$.

Grafisch kunnen we de parameters aflezen als volgt:

- Het gemiddelde μ is de x-waarde van de top.
- De standaardafwijking σ is de horizontale afstand tussen top en buigpunt.



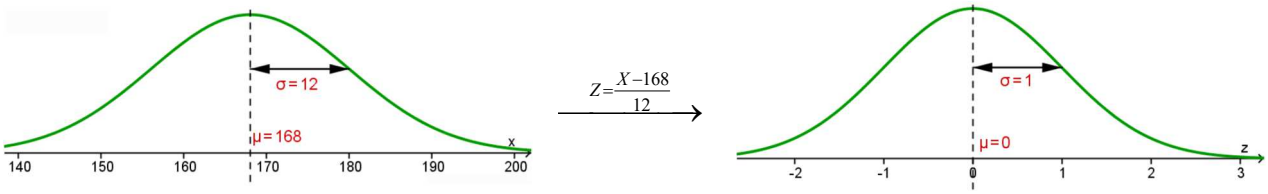
Het gemiddelde μ bepaalt het midden van de grafiek. De standaardafwijking σ bepaalt hoe steil de grafiek is, dus in welke mate de gegevens in de buurt van het gemiddelde liggen.

d) De standaard normale verdeling

De normale verdeling met gemiddelde $\mu=0$ en standaardafwijking $\sigma=1$ noemen we de standaard normale verdeling. De standaard normale verdeling noteren we met Z, dus $Z \sim N(0,1)$.

Elke normale verdeling kunnen we via een eenvoudige transformatie herleiden tot de standaard normale verdeling. We zeggen dan dat we de verdeling standaardiseren. Dit gebeurt aan de hand van de formule: $Z=\dfrac{X-\mu}{\sigma}$, of omgekeerd $X=\mu+\sigma\cdot Z$.

We standaardiseren als voorbeeld de verdeling $X \sim N(168,12)$:



Je ziet dat in de praktijk enkel de ijk op de assen verandert (de x-as wordt de z-as). In oefeningen zal het vaak nuttig zijn onder de grafiek zowel x-waarden als z-waarden te plaatsen.

De z-score

Het feit dat elke normale verdeling gestandaardiseerd kan worden laat ons ook toe gegevens objectief met elkaar te vergelijken. We illustreren dit met een eenvoudig voorbeeld:

Jan scoort 45/60 op een toets waar het gemiddelde 36 was en de standaarddeviatie 4. Eva scoort op een andere toets 48/60 waar het gemiddelde 42 was en de standaardafwijking 6.

We standaardiseren beide punten (we noemen dit hun z-scores berekenen):

$z_{Jan}=\dfrac{45-36}{4}=\dfrac{9}{4}=2,25$ en $z_{Eva}=\dfrac{48-42}{6}=\dfrac{6}{6}=1$. Relatief gezien scoort Jan dus (veel) beter!

In woorden geeft de z-score aan hoeveel standaardafwijkingen je (positief of negatief) van het gemiddelde afwijkt.

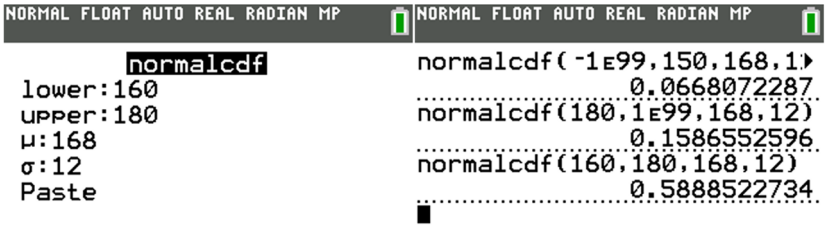
e) Rekenen met de normale verdeling

Percentages berekenen met de GRM

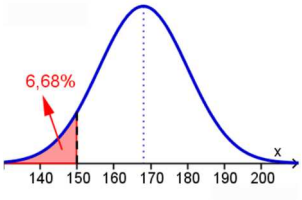
Je rekenmachine kan voor elke normale verdeling berekenen hoeveel % van de gegevens er tussen twee waarden liggen. Je gebruikt daarvoor de functie **normalcdf**. Deze functie heeft 4 parameters nodig.

Voorbeeld: De lichaamslengte van de vrouwen uit een bevolkingsgroep is normaal verdeeld met een gemiddelde van 168 cm en een standaardafwijking van 12 cm.

a) Wat is de kans dat een vrouw kleiner is dan 150cm?
b) Wat is de kans dat een vrouw groter is dan 180cm?
c) Wat is de kans dat een vrouw een lengte heeft tussen 160cm en 180cm?

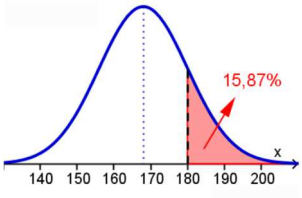


a)



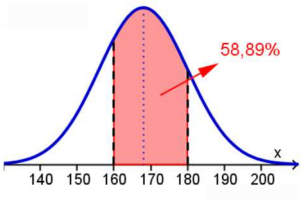
We krijgen enkel een bovengrens gegeven. Alle gegevens kleiner dan 150 voldoen dus aan het gegeven. Als ondergrens gebruiken we dan $-\infty$ (in de rekenmachine typ je in: $-1E99$). We berekenen:
normalcdf(-1E99, 150, 168, 12) = 0.0668072287
→ ongeveer 6,68% van de vrouwen is kleiner dan 150 cm.

b)



Hier krijgen we enkel een ondergrens. Analoog aan de voorgaande oefening berekenen we:
normalcdf(180, 1E99, 168, 12) = 0.1586552596
→ ongeveer 15,87% van de vrouwen is groter dan 180 cm.

c)



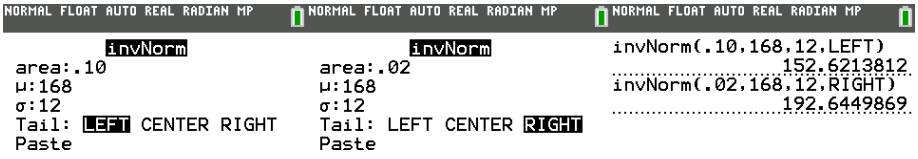
Hier zijn beide grenzen gegeven. We berekenen:
normalcdf(160, 180, 168, 12) = 0.5888522734
→ ongeveer 58,89% van de vrouwen is tussen 160 cm en 180 cm groot.

Grenzen berekenen met de GRM

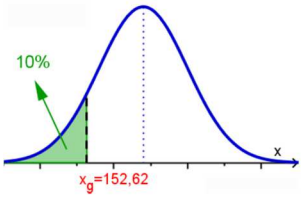
Anderzijds kan de rekenmachine ook de grens berekenen waarvoor een gegegeven percentage van de gegevens kleiner zal zijn. We gebruiken hiervoor de functie **invNorm**. Deze functie heeft 3 parameters nodig.

Voorbeeld: De lichaamslengte van de vrouwen uit een bevolkingsgroep is normaal verdeeld met een gemiddelde van 168 cm en een standaardafwijking van 12 cm.

a) De kortste 10% vrouwen noemen we zeer klein. Tot welke grootte is dit?
b) De grootste 2% van de vrouwen noemen we reuzen. Vanaf welke grootte is dit?

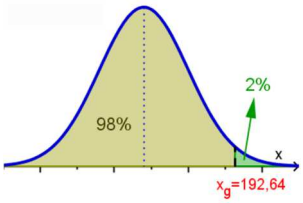


a)



Dit kunnen we berekenen met:
invNorm(0.10, 168, 12,LEFT) = 152.6213812
→ 10% van de vrouwen is korter dan ongeveer 152,62 cm.

b)

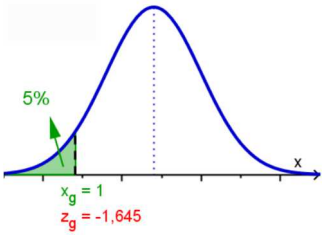


Denk eraan dat de functie **invNorm** de grens berekent waarvoor een bepaald percentage kleiner is. Vermits de totale oppervlakte onder de kromme 100% is, zoeken we dus hier de grens waarvoor 98% van de vrouwen kleiner is:
invNorm(0.02, 168, 12,RIGHT) = 192.6449869
→ 2% van de vrouwen is groter dan ongeveer 192,64 cm.

Wat als het gemiddelde of de standaardafwijking onbekend zijn?

Tot nog toe hebben we in de oefeningen de standaard normale verdeling nog niet nodig gehad. We zullen ze echter wel nodig hebben als we een onbekend gemiddelde of standaardafwijking hebben.

Voorbeeld: Veronderstel dat een machine die flessen water vult ingesteld is op een gemiddelde van 1 liter en een standaardafwijking van 5 gram (=0,005 liter). Dan zal de helft van alle flessen een inhoud hebben van minder dan 1 liter. Op welk gemiddelde moet de machine worden afgesteld (met dezelfde standaardafwijking) als slechts 5% van de flessen minder dan 1 liter water mag bevatten?



We krijgen hier een grens én een percentage gegeven, maar het gemiddelde is onbekend.
De manier om dit op te lossen is om dezelfde grens te berekenen voor de standaard normale verdeling en dan de standaardiseringsformule te gebruiken om μ te vinden.

De grens waarvoor 5% van de gegevens is voor de standaard normale verdeling vind je met je rekenmachine: z_g = **invNorm**(0.05) ≈ -1,645. Invullen in de standaardiseringsformule geeft:

$$z_g = \frac{x_g - \mu}{\sigma} \Leftrightarrow z_g.\sigma = x_g - \mu \Leftrightarrow \mu = x_g - z_g.\sigma \approx 1 - (-1{,}645).0{,}005 = 1{,}008225\,.$$

→ De machine moet ingesteld worden op een gemiddelde van ongeveer 1,008225 liter.

5) Kansverdelingen

a) Stochasten

In de kansrekening is een *stochastische variabele* (kortweg *stochast*) een grootheid waarvan de waarde een reëel getal is dat afhangt van een toevallige uitkomst in een kansexperiment. Soms wordt dit ook een *toevalsveranderlijke* genoemd.

We onderscheiden twee soorten stochasten: discrete stochasten en continue stochasten.

Discrete stochasten

Een *discrete stochast* X kan slechts een eindig aantal waarden $x_1, x_2, ..., x_n$ aannemen. We noteren de kans dat x_i optreedt met p_i, of anders geformuleerd: $P(X = x_i) = p_i$. De som van al deze kansen moet uiteraard 1 zijn.

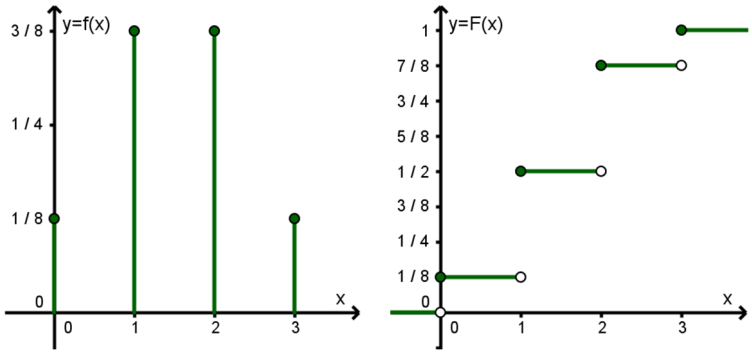
De functie $f(x) = P(X = x)$ noemen we de *kansverdelingsfunctie* (er geldt dus $f(x_i) = p_i$).

De cumulatieve verdelingsfunctie F definiëren we als $F(x) = P(X \leq x)$

Voorbeeld: We beschouwen het kansexperiment 'driemaal opgooien van een muntstuk', en we noemen X de stochast 'het aantal keer dat je munt hebt gegooid'.

Hierbij geldt voor de kansverdeling: $f(0) = f(3) = P(X = 0) = \frac{1}{8}$, $f(1) = f(2) = P(X = 1) = \frac{3}{8}$.

Hiernaast zie je zowel de kansverdelingsfunctie f als de cumulatieve verdelingsfunctie F getekend. Merk op dat f enkel gedefinieerd is voor de waarden $0, 1, 2$ en 3, en F voor alle reële waarden.



Continue stochasten

Een continue toevalsveranderlijke X kan alle waarden aannemen in een interval. Voor elke continue toevalsveranderlijke bestaat er een kansdichtheidsfunctie f zodat $P(a \leq X \leq b) = \int_a^b f(x)\,dx$. De kans dat X een waarde aanneemt tussen a en b is dus gelijk aan de oppervlakte begrepen tussen de x-as en de dichtheidsfunctie f in het interval $[a, b]$.

Rekening houdend met de definitie van het begrip kans, moet een dichtheidsfunctie f voldoen aan:

- $\forall x \in \mathbb{R} : f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x)\,dx = 1$ (de totale oppervlakte onder de dichtheidskromme moet 1 zijn)

De cumulatieve verdelingsfunctie F definiëren we als $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$.

Uit de integraalrekening weten we dan dat $D(F(x)) = f(x)$.

Merk op dat alle normale verdelingen voorbeelden zijn van continue stochasten.

b) Verwachtingswaarde van een stochast

Bij discrete stochasten

De verwachtingswaarde van een discrete stochast die de waarden x_i kan aannemen met kans p_i (en met $1 \le i \le n$) wordt gegeven door: $E(x) = \mu_X = \sum_{i=1}^{n} x_i p_i$.

Voor het voorbeeld uit de vorige paragraaf (aantal keer munt bij het opgooien van 3 munten) is dit:

$$E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2} = 1{,}5 \, .$$

Dit resultaat is heel logisch, want je verwacht de helft van de keren dat je gooit munt. Merk op dat de verwachtingswaarde $1{,}5$ zelf geen mogelijke uitkomst van het kansexperiment is. Dat hoeft ook niet. De verwachtingswaarde is de gemiddelde waarde van de stochast als je het experiment oneindig veel zou herhalen (wat uiteraard in praktijk nooit mogelijk is).

Bij continue stochasten

De verwachtingswaarde bij een continue stochast met kansdichtheidsfunctie f definiëren we als:

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

c) De standaardafwijking van een stochast

Bij discrete stochasten

De variantie van een toevalsveranderlijke is de verwachte gemiddelde kwadratische afwijking van de stochast ten opzichte van zijn verwachtingswaarde.

In formulevorm wordt dit: $Var(X) = \sigma_X^2 = \sum_{i=1}^{n} (x_i - \mu_X)^2 \, p_i$

De standaardafwijking van een stochast is de vierkantswortel uit zijn variantie: $\sigma_X = \sqrt{Var(X)}$.

Stelling: Een alternatieve manier om de variantie te berekenen is $Var(X) = E(X^2) - (E(X))^2$.

Bewijs: $\sigma_X^2 = \sum_{i=1}^{n} (x_i - \mu_X)^2 \cdot p_i = \sum_{i=1}^{n} x_i^{\,2} \cdot p_i - 2\sum_{i=1}^{n} x_i \cdot \mu_X \cdot p_i + \sum_{i=1}^{n} \mu_X^2 \cdot p_i$

$$= E(X^2) - 2\mu_X \cdot \underbrace{\sum_{i=1}^{n} x_i \cdot p_i}_{=\mu_X} + \mu_X^2 \cdot \underbrace{\sum_{i=1}^{n} p_i}_{=1} = E(X^2) - \mu_X^2 \quad \square$$

Dit rekent iets makkelijke in oefeningen.

Hernemen we nogmaals het voorbeeld uit de vorige paragraaf (aantal keer munt bij het opgooien van 3 munten) dan geeft dit:

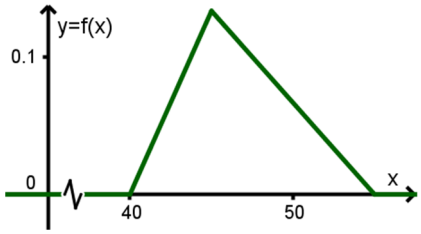
$Var(X) = \sum_{i=1}^{n} x_i^{\,2} \cdot p_i \; - \; \mu_X^2 = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 4 \cdot \frac{3}{8} + 9 \cdot \frac{1}{8} - \frac{9}{4} = \frac{3}{4}$, en dus ook $\sigma_X = \frac{\sqrt{3}}{2}$.

Bij continue stochasten

Hierbij geldt: $Var(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} (x - \mu_X)^2 \cdot f(x) dx = \int_{-\infty}^{+\infty} x^2 \cdot f(x) dx - \mu_X^2$.

We bekijken een voorbeeld van een continue stochast: Noem X de effectieve tijdsduur (in minuten) van een les wiskunde op het Oscar Romerocollege. We definiëren de kansdichtheidsfunctie als volgt:

$$f(x)=\begin{cases}\frac{2}{75}x-\frac{16}{15} & ,x\in[40,45]\\ -\frac{1}{75}x+\frac{11}{15} & ,x\in[45,55]\\ 0 & ,x\notin[40,55]\end{cases}$$

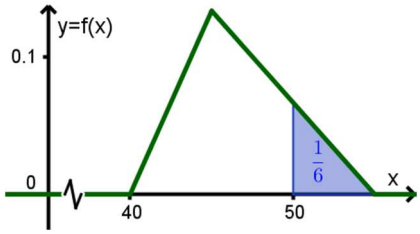


Dat dit een kansdichtheidsfunctie is kunnen we eenvoudig narekenen, want:

$$\int_{-\infty}^{+\infty} f(x)\,dx=\int_{40}^{45}\left(\frac{2}{75}x-\frac{16}{15}\right)dx+\int_{45}^{55}\left(-\frac{1}{75}x+\frac{11}{15}\right)dx=1$$

De kans dat een lesuur langer dan 50 minuten duurt wordt berekend als volgt:

$$P(X>50)=\int_{50}^{+\infty} f(x)\,dx=\int_{50}^{55}\left(-\frac{1}{75}x+\frac{11}{15}\right)dx=\frac{1}{6}$$



De verwachtingswaarde voor deze stochast is:

$$E(X)=\mu_X=\int_{-\infty}^{+\infty} x\cdot f(x)\,dx=\int_{40}^{45} x\cdot\left(\frac{2}{75}x-\frac{16}{15}\right)dx+\int_{45}^{55} x\cdot\left(-\frac{1}{75}x+\frac{11}{15}\right)dx=\frac{140}{3}\approx 46,67$$

De variantie en de standaardafwijking zijn:

$$Var(X)=\sigma_X^2=\int_{-\infty}^{+\infty} x^2\cdot f(x)\,dx-\mu_X^2$$

$$=\int_{40}^{45} x^2\cdot\left(\frac{2}{75}x-\frac{16}{15}\right)dx+\int_{45}^{55} x^2\cdot\left(-\frac{1}{75}x+\frac{11}{15}\right)dx-\left(\frac{140}{3}\right)^2=\frac{175}{18}\Rightarrow\sigma_X=\sqrt{\frac{175}{18}}\approx 3,118$$

Een lesuur wiskunde duurt op het Oscar Romerocollege dus gemiddeld 46 minuten en 40 seconden met een standaardafwijking van ongeveer 3 minuten en 7 seconden.

d) Rekenregels

Stelling: Als X een discrete stochast is, en $a\in\mathbb{R}$ een constante, dan geldt:

★ $E(X+a)=E(X)+a$ en $Var(X+a)=Var(X)$

★ $E(a.X)=a.E(X)$ en $Var(a.X)=a^2.Var(X)$

Bewijs: ★ $E(X+a)=\sum_{i=1}^{n}(x_i+a)\,p_i=\sum_{i=1}^{n}x_ip_i+\sum_{i=1}^{n}a\,p_i=E(X)+a\cdot\underbrace{\sum_{i=1}^{n}p_i}_{=1}=E(X)+a$

$$Var(X+a)=\sum_{i=1}^{n}\left(x_i+a-\left(E(X)+a\right)\right)^2p_i=\sum_{i=1}^{n}\left(x_i-E(X)\right)^2p_i=Var(X)$$

★ $E(a.X)=\sum_{i=1}^{n}(a.x_i)\,p_i=a.\sum_{i=1}^{n}x_ip_i=a.E(X)$

$$Var(a.X)=\sum_{i=1}^{n}\left(a.x_i-a.E(X)\right)^2 p_i=a^2.\sum_{i=1}^{n}\left(x_i-E(X)\right)^2 p_i=a^2.Var(X) \quad \square$$

Deze stellingen gelden ook voor continue stochasten (bewijs als oefening met integraalrekening).

Van de volgende stellingen valt het bewijs buiten het bestek van deze cursus.

Stelling: Als X en Y onafhankelijke stochasten zijn, dan geldt:

★ $E(X+Y)=E(X)+E(Y)$ en $Var(X+Y)=Var(X)+Var(Y)$

★ $E(X\cdot Y)=E(X)\cdot E(Y)$

Stelling: Als X en Y onafhankelijke, normaal verdeelde stochasten zijn, dan is ook de lineaire combinatie $S=a\cdot X+b\cdot Y$ normaal verdeeld met $\mu_S=a\cdot\mu_X+b\cdot\mu_Y$ en $\sigma_S=\sqrt{a^2\cdot s_X^2+b^2\cdot s_Y^2}$.

Voorbeeld: Kevin loopt zijn 400m gemiddeld op 50s met een standaardafwijking van 2s. Jonathan loopt zijn 400m op gemiddeld 52s met een standaardafwijking van 3s. Wat is de kans dat Jonathan een wedstrijd sneller loopt dan Kevin als je ervan uitgaat dat beide stochasten normaal verdeeld zijn.

Voor beide stochasten geldt dus: $K\sim N(50,2)$ en $J\sim N(52,3)$. We moeten de kans berekenen $P(J<K)=P(J-K<0)$. Wegens de vorige stelling is ook $J-K$ normaal verdeeld met gemiddelde $\mu=52-50=2$ en $\sigma=\sqrt{2^2+3^2}=\sqrt{13}$. De kans kunnen we dus berekenen met onze rekenmachine en is gelijk aan $P(J-K<0)=0,2895 \quad \left(=\text{normalcdf}\left(-1E99,0,2,\sqrt{13}\right)\right)$.

e) De $\sqrt{n}$ - wetten

Uit de rekenregels volgt onmiddellijk dat als $X_1,X_2,...,X_n$ allemaal onafhankelijke stochasten zijn met hetzelfde gemiddelde μ en dezelfde standaardafwijking σ, dat dan geldt:

- Voor de stochast $S=X_1+X_2+...+X_n$: $\mu_S=n.\mu$ en $\sigma_S=\sqrt{n}\cdot\sigma$
- Voor de stochast $\overline{X}=\dfrac{X_1+X_2+...+X_n}{n}$: $\mu_{\overline{X}}=\mu$ en $\sigma_{\overline{X}}=\dfrac{\sigma}{\sqrt{n}}$

Voorbeeld: Fanny haalt op haar toetsen van wiskunde gemiddeld 6 op 10 met een standaardafwijking van 1 (deze punten zijn normaal verdeeld).

a) Voor een rapportperiode worden 3 toetsen samengeteld. Wat is de kans dat Fanny daar meer dan 20 op 30 haalt?
De stochasten T_i die de punten per toets voorstellen zijn normaal verdeeld $T_i\sim N(6,1)$. De som S van drie toetsen is ook normaal verdeeld, met $\mu_S=3.6=18$ en $\sigma_S=\sqrt{3}.1=\sqrt{3}$.
De kans dat Fanny meer dan 20 haalt is dus:
$P(S>20)=0,1241 \quad \left(=\text{normalcdf}\left(20,1E99,18,\sqrt{3}\right)\right)$

b) Op het einde van het jaar heeft Fanny 16 toetsen gemaakt. Bereken de kans dat Fanny haar gemiddelde onder de helft zit.
Voor het gemiddelde geldt $\overline{X}\sim N\left(6,\dfrac{1}{4}\right)$, want $\mu_{\overline{X}}=\mu_{T_i}=6$ en $\sigma_{\overline{X}}=\dfrac{\sigma_{T_i}}{\sqrt{16}}=\dfrac{1}{4}$.
De gevraagde kans is $P(\overline{X}<5)=0,000032 \quad \left(=\text{normalcdf}\left(-1E99,5,6,1/4\right)\right)$.

6) De binomiale verdeling

Met de normale verdeling zagen we reeds een gedetailleerd voorbeeld van een familie continue stochasten. De binomiale verdeling is een voorbeeld van een familie discrete stochasten.

a) Bernoulli experimenten

Een Bernoulli-experiment is een toevalsexperiment met twee mogelijke uitkomsten, meestal aangeduid als "succes" en "mislukking". Een Bernoulli-experiment wordt beschreven door een stochast die de waarden 1 (succes) en 0 (mislukking) kan aannemen.

Is B een stochast met $P(B=1)=p$ en $P(B=0)=q=1-p$ (p is de kans op succes dus is $q=1-p$ de kans op mislukking), dan worden de verwachtingswaarde en de standaardafwijking gegeven door:

- $\mu = 0.q + 1.p = p$
- $\sigma = \sqrt{(0-p)^2.q + (1-p)^2.p} = \sqrt{p^2q + q^2p} = \sqrt{pq(p+q)} = \sqrt{pq}$

b) Binomiale verdeling

De binomiale verdeling is de kansverdeling van het aantal successen X in een reeks van n onafhankelijke Bernoulli-experimenten met kans op succes gelijk aan p. We noteren de verdeling als $X \sim B(n,p)$.

In een reeks van n Bernoulli-experimenten kunnen 0 tot en met n successen voorkomen. Het aantal successen is dus inderdaad een toevalsveranderlijke X. De kans op precies k successen, $P(X=k)$, kan gemakkelijk berekend worden als we bedenken dat elke reeks uitkomsten met k successen (en dus $n-k$ mislukkingen) dezelfde kans $p^k q^{n-k}$ heeft. Omdat er C_n^k verschillende reeksen zijn met precies k successen, wordt de kansverdelingsfunctie gegeven door: $P(X=k) = C_n^k p^k (1-p)^{n-k}$.

Voorbeeld 1: Je gooit 10 keer met een dobbelsteen. Wat is de kans dat je twee keer een 6 gooide?

De stochast X die aangeeft hoeveel keer er 6 werd gegooid is binomiaal verdeeld: $X \sim B(10,1/6)$.

Dus $P(X=2) = C_{10}^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^8 \approx 0,2907$.

Je kan dit in je rekenmachine makkelijk berekenen met de functie $\text{binompdf}(10,1/6,2)$.

Voorbeeld 2: Wat is in hetzelfde experiment de kans dat je hoogstens 4 keer een 6 gooide.

$$\begin{aligned} P(X \le 4) &= P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) \\ &= C_{10}^0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^{10} + C_{10}^1 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^9 + C_{10}^2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^8 + C_{10}^3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^7 + C_{10}^4 \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^6 \\ &\approx 0,984538 \end{aligned}$$

Het is heel wat werk om dit in te geven in je rekenmachine. Gelukkig bestaat er hier ook een snellere manier voor, want ook de cumulatieve verdelingsfunctie voor de binomiale verdeling zit in je rekenmachine. Je kan dus in plaats van het voorgaande simpelweg $\text{binomcdf}(10,1/6,4)$ invoeren.

De karakteristieken

Een binomiale verdeling $Bin \sim B(n,p)$ is eenvoudigweg de som van n onafhankelijke Bernoulli experimenten met kans op succes gelijk aan p. Uit het voorgaande volgt dus onmiddellijk dat:

- $\mu_{Bin} = n.\mu_{Ber} = n.p$
- $\sigma_{Bin} = \sqrt{n}.\sigma_{Ber} = \sqrt{n}.\sqrt{pq} = \sqrt{npq}$

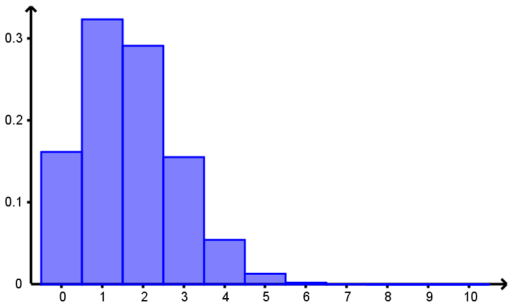
Voorbeeld: Wat is de verwachtingswaarde en de standaardafwijking voor het aantal keer dat je 6 gooit als je 10 worpen doet met een gewone dobbelsteen?

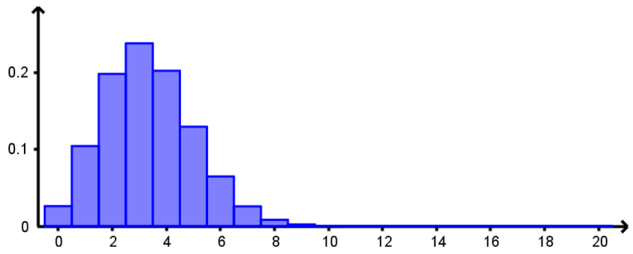
$$\mu = 10 \cdot \frac{1}{6} = \frac{5}{3} \text{ en } \sigma = \sqrt{10 \cdot \frac{1}{6} \cdot \frac{5}{6}} = \frac{5\sqrt{2}}{6}$$

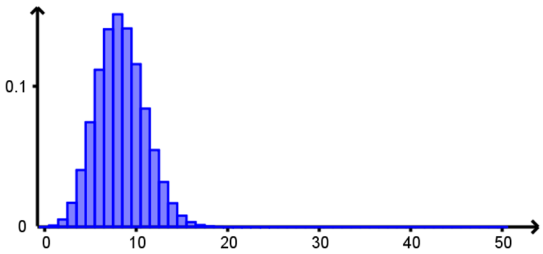
c) De normale benadering

Tekenen we de grafiek van de binomiale verdelingsfunctie uit het vorige voorbeeld ($n=10$) dan krijgen we de grafiek rechts.

Als we het aantal keer dat het experiment wordt uitgevoerd (n) opdrijven, dan krijgen we grafieken die meer en meer gaan lijken op een normale verdeling. We illustreren dit met $n=20$ en $n=50$:







We stellen dus vast dat als n voldoende groot dat dan de binomiale verdeling $B(n,p)$ kan benaderd worden door de normale verdeling $N\left(np, \sqrt{npq}\right)$.

Men heeft afgesproken dat dit mag van zodra drie voorwaarden gelden: • $n \geq 20$ • $np \geq 5$ • $nq \geq 5$

De continuïteitscorrectie

Er is bij deze benadering wel een belangrijk iets op te merken: een discrete stochast benaderen met een continue stochast vraagt om een aanpassing van je grenzen. We illustreren met een voorbeeld:

Beschouw de stochast X als het aantal keer dat je 6 gooit als je 100 keer met een dobbelsteen werpt.

Voor de stochast X geldt dat $X \sim B(100, 1/6)$.

Deze stochast kan benaderd worden met de stochast $X' \sim N\left(\frac{100}{6}, \frac{5\sqrt{5}}{3}\right)$, want de parameters zijn:

$$\mu = np = 100 \cdot \frac{1}{6} = \frac{50}{3} \text{ en } \sigma = \sqrt{npq} = \sqrt{100 \cdot \frac{1}{6} \cdot \frac{5}{6}} = \frac{5\sqrt{5}}{3}$$

a) Bereken de kans dat je van de 100 keer dat je gooit in totaal 10 keer een 6 gooide.

Exact is dit $P(X=10) = C_{100}^{10} \left(\frac{1}{6}\right)^{10} \left(\frac{5}{6}\right)^{90} \approx 0{,}02140$

Benaderd: $P(X=10) \approx P(9{,}5 \le X' < 10{,}5) \approx 0{,}02175$

(met de GRM: $\text{binompdf}(100,1/6,10)$ en $\text{normalcdf}(9.5,10.5,50/3,5\sqrt{5}/3)$)

Je neemt deze grenzen omdat je de waarden in het interval [9,5 ; 10,5[afrondt tot 10.

b) Bereken de kans dat je van de 100 keer minstens 20 keer een 6 gooide.

Exact is dit $P(X \ge 20) = 1 - P(X \le 19) = 0{,}2197$

Benaderd is dit $P(X \ge 20) \approx P(X' \ge 19{,}5) \approx 0{,}2235$

(met de GRM: $1\text{-binomcdf}(100,1/6,19)$ en $\text{normalcdf}(19.5,1E99,50/3,5\sqrt{5}/3)$)

De centrale limietstelling

De benadering die hierboven gebruikt wordt is een speciaal geval van de centrale limietstelling. Dit is misschien wel de belangrijkste stelling in heel de statistiek. Ze zegt: 'De som van een aantal onafhankelijke toevalsveranderlijken zal altijd naar een normaal verdeelde stochast neigen als het aantal maar groot genoeg is'. Het bewijs van deze stelling valt (ver) buiten het bestek van deze cursus.

7) Schatten met puntschatters en betrouwbaarheidsintervallen

Het doel van statistiek is altijd om op basis van een goede steekproef een uitspraak te kunnen doen over de populatie. De meest eenvoudige manier om dit te doen is met behulp van puntschatters.

a) Puntschatters

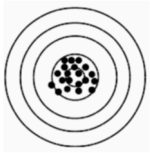
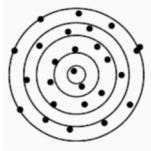
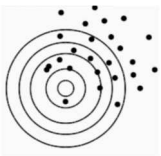
Als we een parameter van een populatie (voorbeelden zijn een proportie p, een gemiddelde m of een standaardafwijking σ) willen schatten kunnen we gebruik maken van een waarde die we berekenen aan de hand van een steekproef. Deze waarde is een stochast, want hij is afhankelijk van het toeval (de steekproef). We noemen zulke stochasten ook wel eens *steekproefgrootheden*.

We spreken af dat we steekproefgrootheden die parameters schatten (*puntschatters* genoemd) noteren met dezelfde letters als de parameter die ze willen schatten, maar dan met een hoedje op. Zo schat $\hat{p}$ de parameter p, schat $\hat{\mu}$ de parameter μ enzovoort.

Goede puntschatters voldoen aan twee eigenschappen:

- Hij moet *zuiver* zijn: dit wil zeggen dat de verwachtingswaarde van de puntschatter gelijk is aan de parameter die hij schat.
- Hij moet efficiënt zijn: dit wil zeggen dat de standaardafwijking van de puntschatter zo klein mogelijk is.

Stel dat je met een dartspijltje mikt naar de roos van een dartsbord. Dan kan je zeggen dat het resultaat van je worp een schatter is van de roos op het dartsbord. De begrippen zuiver en efficiënt kan je dan heel eenvoudig als volgt grafisch voorstellen:

	Zuiver	Niet zuiver
Efficiënt		
Niet efficiënt		

Er kan bewezen worden dat het gemiddelde van een steekproef $\overline{X}$ een zuivere en efficiënte schatter is voor het gemiddelde μ van een populatie, en dat S, de stochast die de standaardafwijking berekent bij een steekproef, een zuivere en efficiënte schatter is voor de standaardafwijking σ van de populatie.

b) Betrouwbaarheidsintervallen

Een puntschatting is interessant maar het is zeer beperkt. Het is veel zinvoller om een interval te geven waarbinnen de te schatten parameter met een zekere *betrouwbaarheid* ligt.

Om dit probleem op te lossen bekijken we eerst het begrip kritieke z-waarden:

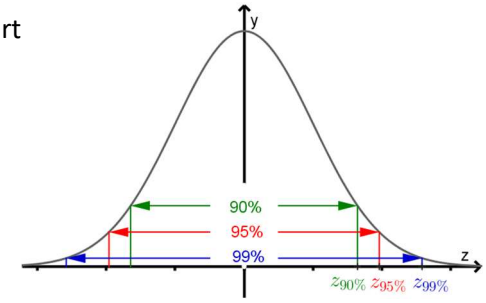
Kritieke z-waarden

De *kritieke* z*-waarde* z_α^* die bij een bepaald percentage $\alpha\,\%$ hoort wordt gegeven door de z-score waarvoor geldt:

$$P\left(-z_\alpha^* < Z < z_\alpha^*\right) = \alpha\,\%\,.$$

Veelgebruikte waarden zijn $z_{90}^* = 1{,}65$, $z_{95}^* = 1{,}96$ en $z_{99}^* = 2{,}58$.

Dit is eenvoudig geïllustreerd op de figuur hiernaast.



Waarschijnlijkheidsintervallen voor steekproefproporties

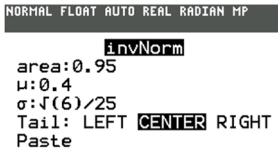
We bekijken eerst de eenvoudige richting: Als de proportie p van een eigenschap gekend is in een populatie, dan stellen we een symmetrisch interval op rond p zodat in 95% van de steekproeven met grootte n die je neemt de proportie $\hat{p}$ van die eigenschap in de steekproef in dat interval ligt.

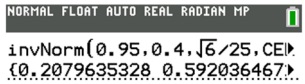
Aan de hand van de kritieke z-waarden kan je dan de kritieke x-waarden berekenen met behulp van de formule $X = \mu + \sigma \cdot Z$.

Voorbeeld: Van een bepaalde munt is geweten dat hij vervalst is. Hij werpt in 40% van de gevallen munt. Stel het interval op voor de steekproefproportie $\hat{p}$, symmetrisch om de gekende proportie $p = 0,4$, waarbinnen de steekproefproportie met 95% waarschijnlijkheid zal liggen, als de steekproefgrootte $n = 25$ is.

Wegens de centrale limietstelling is de steekproefproportie ongeveer normaal verdeeld met parameters $\mu = p = 0,4$ en $\sigma = \frac{1}{n}\sqrt{npq} = \sqrt{\frac{0,4.0,6}{25}} = \frac{\sqrt{6}}{25}$.

De kritieke z-waarde voor 95% is $z^*_{95} \approx 1,96$, dus de kritieke x-waarden zijn (kan ook met GRM):





$$x_1 \approx 0,4 + 1,96 \cdot \sqrt{6}/25 \approx 0,592$$

$$x_2 \approx 0,4 - 1,96 \cdot \sqrt{6}/25 \approx 0,208$$

We kunnen dus zeggen dat er voor elke steekproef een kans is van 95% dat de steekproefproportie $\hat{p}$ (hoeveel % van de 25 keer je munt gooit met de valse munt) in het interval $[0,208\,;\ 0,592]$ zal liggen.

De formule die we hier hebben opgesteld valt heel eenvoudig te veralgemenen:

Komt een kenmerk in een populatie voor met proportie p dan wordt het $\alpha\,\%$-waarschijnlijkheidsinterval voor $\hat{p}$ gegeven door $\left[\, p - z^*_\alpha \cdot \sqrt{\frac{pq}{n}}\ ;\ p + z^*_\alpha \cdot \sqrt{\frac{pq}{n}} \,\right]$.

Betrouwbaarheidsintervallen voor proporties

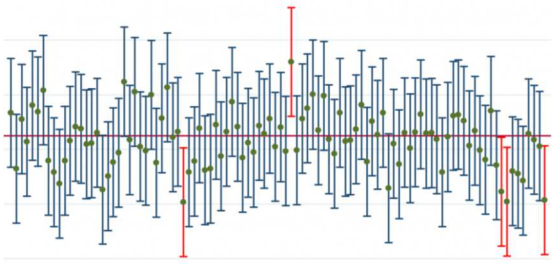
Draaien we deze redenering om, dan kunnen we stellen dat in $\alpha\,\%$ van de steekproeven met grootte n de populatieproportie p in het interval $\left[\, \hat{p} - z^*_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ;\ \hat{p} + z^*_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \,\right]$ zal liggen.

We noemen dit het $\alpha\,\%$-betrouwbaarheidsinterval dat we met die steekproef bekomen.

De waarde $m = z^*_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ noemen we de foutenmarge van het betrouwbaarheidsinterval.

!Belangrijk!: Dit interval drukt geen kans uit. De werkelijke proportie p zal ofwel in het interval liggen, ofwel niet. Vandaar dat we hier spreken van betrouwbaarheid in plaats van waarschijnlijkheid.

De betekenis van een betrouwbaarheidsinterval kunnen we ook eenvoudig grafisch illustreren.
Op de figuur hiernaast staan voor 100 verschillende steekproeven telkens het bijhorende betrouwbaarheids-interval getekend. Je ziet dat in 95% van de gevallen de populatieparameter die je wil schatten (de horizontale bordeaux lijn) in het interval ligt.



Voorbeeld: Om te zien hoeveel mensen er akkoord gaan met een nieuw wetsvoorstel doet men een peiling bij 120 mensen. Daarvan gaat 30% akkoord met het wetsvoorstel. Wat is het 90%-betrouwbaarheidsinterval voor mensen die akkoord gaan in de populatie op basis van deze steekproef?

$$\left[0{,}3-1{,}65\cdot\sqrt{\frac{0{,}3\cdot 0{,}7}{120}};\ 0{,}3+1{,}65\cdot\sqrt{\frac{0{,}3\cdot 0{,}7}{120}}\right]\approx\left[0{,}231;\ 0{,}369\right]$$

Je GRM kan dit ook sneller berekenen met de functie **1-PropZInt** (STAT – TESTS – A:1-PropZInt).

Merk op dat je hier bij x moet ingeven hoeveel mensen er positief stemden in de steekproef (dus een absoluut aantal en geen relatief).

Betrouwbaarheidsintervallen voor het populatiegemiddelde μ

Volledig analoog aan de intervallen voor proporties kunnen we ook intervallen opstellen voor het steekproefgemiddelde $\hat{\mu}=\overline{X}$ en het populatiegemiddelde μ.

We zagen in de vorige hoofdstukken reeds dat als X een stochast is met gemiddelde μ en standaardafwijking σ, dat dan $\overline{X}$ bij benadering normaal verdeeld is als n voldoende groot is (centrale limietstelling) met parameters $\mu_{\overline{X}}=\mu$ en $\sigma_{\overline{X}}=\dfrac{\sigma}{\sqrt{n}}$ (de $\sqrt{n}$-wet). Met andere woorden:

$$P\left(\overline{X}\in\left[\mu-z_\alpha^*\cdot\frac{\sigma}{\sqrt{n}};\mu+z_\alpha^*\cdot\frac{\sigma}{\sqrt{n}}\right]\right)=\alpha\,\%\,.$$

De redenering omdraaien impliceert dat we nu ook een $\alpha\,\%$-betrouwbaarheidsinterval kunnen opstellen voor μ als we een steekproef nemen van grootte n met gemiddelde $\overline{x}$, namelijk:

$$\left[\overline{x}-z_\alpha^*\cdot\frac{\sigma}{\sqrt{n}};\overline{x}+z_\alpha^*\cdot\frac{\sigma}{\sqrt{n}}\right].$$

!Belangrijk!: deze formule geldt enkel als de populatiestandaardafwijking σ gekend is, en als de steekproefverdeling $\overline{X}$ (bij benadering) normaal verdeeld is. Anders moet er een andere formule gebruikt worden die gebruik maakt van de student t-verdeling met behulp van de steekproefstandaardafwijking s (in plaats van de standaard normale verdeling – zie later).

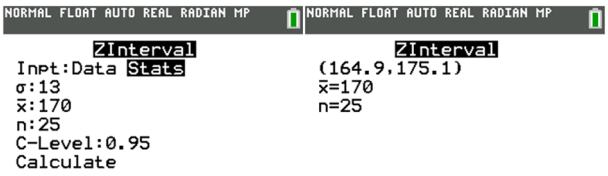
De waarde $m=z_\alpha^*\cdot\dfrac{\sigma}{\sqrt{n}}$ noemen we de foutenmarge van het betrouwbaarheidsinterval.

Voorbeeld: De lengte van een laatstejaarsstudent op het ORC is normaal verdeeld met een standaardafwijking van 13 cm. Een steekproef van 25 personen bekomt een gemiddelde lengte van 170 cm.

a) Stel een 95%-betrouwbaarheidsinterval op voor de gemiddelde lengte van een laatstejaarsstudent op het Oscar Romerocollege op basis van deze steekproef.

Het interval is $\left[170-1{,}96\cdot\dfrac{13}{\sqrt{25}};170+1{,}96\cdot\dfrac{13}{\sqrt{25}}\right]$, of eenvoudiger $\left[164{,}904;175{,}096\right]$.

Ook dit kan uiteraard sneller met de GRM:



b) Hoe groot zouden we onze steekproef moeten maken als een 99%-betrouwbaarheidsinterval maximaal 2 cm breed mag zijn.

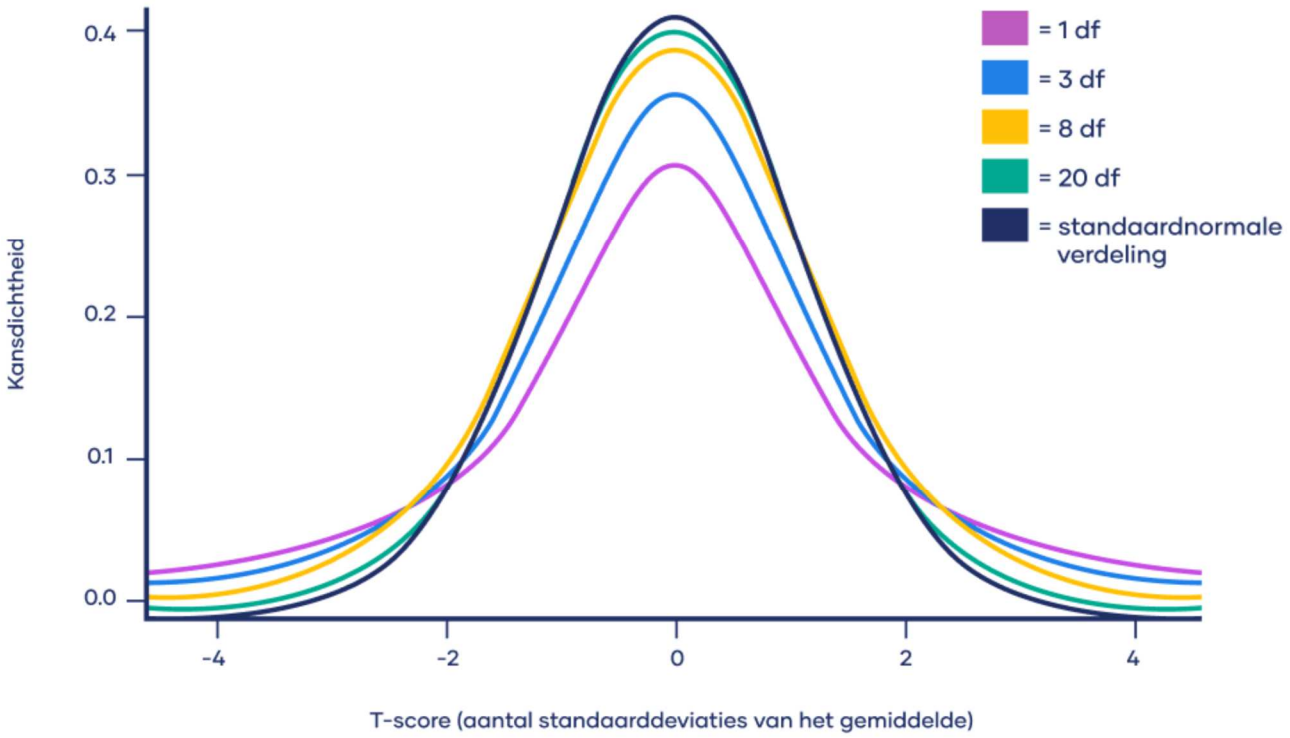
Dan mag de foutenmarge dus maar 1 zijn, dus $1 > 2{,}58 \cdot \frac{13}{\sqrt{n}} \Leftrightarrow n > (2{,}58.13)^2 \approx 1125$. Dit is uiteraard onmogelijk.

De (student) t-verdeling

De t-verdeling (In het Engels: t-distribution of Student's t-distribution) wordt gebruikt als de data bij benadering normaal verdeeld is (en dus een klokvorm volgt), maar waarbij de populatievariantie onbekend is. Ze wordt ook gebruikt bij te kleine steekproeven waar de centrale limietstelling nog niet geldt (meestal geldt de afspraak $n < 30$ en dan moet de data ook wel unimodaal en symmetrisch zijn). De variantie in een t-verdeling wordt dan geschat op basis van het aantal vrijheidsgraden van de dataset. Het aantal vrijheidsgraden bij een steekproef is de steekpoefgrootte verminderd met 1: $df = n - 1$ (hierbij staat df voor degrees of freedom).

Deze verdeling is een meer conservatieve vorm van de standaardnormale verdeling. Dit betekent dat de t-verdeling een lagere kansdichtheid geeft voor het centrum en een hogere kansdichtheid voor de staarten dan de standaard normale verdeling.

Naarmate het aantal vrijheidsgraden toeneemt, zal de t-verdeling steeds dichter bij de standaardnormale verdeling (z-verdeling) komen te liggen, totdat ze nagenoeg hetzelfde zijn. Er geldt dus: $\lim_{n \to +\infty} T_n = Z$.



Vanaf 30 vrijheidsgraden komt de t-verdeling ongeveer overeen met de z-verdeling. Daarom gebruik je voor voldoende grote steekproeven nog steeds de z-verdeling in plaats van de t-verdeling.

De kansdichtheidsfunctie van de t-verdeling (met $df = \nu$) is: $f(x) = \frac{(\nu-1)!!}{\sqrt{\pi \cdot \nu} \cdot (\nu-2)!!} \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$.

Hierbij is de dubbelfaculteit (!!) gedefinieerd als: $n!! = \begin{cases} n \cdot (n-2) \cdot (n-4) \cdot \ldots \cdot 5 \cdot 3 & \text{,n oneven} \\ n \cdot (n-2) \cdot (n-4) \cdot \ldots \cdot 4 \cdot 2 & \text{,n even} \end{cases}$.

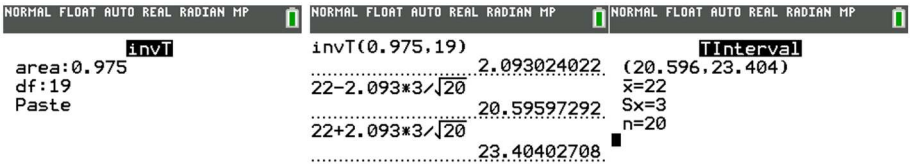
De formules die we zagen voor betrouwbaarheidsintervallen bij de z-verdeling blijven gewoon gelden bij de t-verdeling, alleen gebruiken we dus nu s in plaats van σ als standaardafwijking:

$$\left[\overline{x} - t^*_{n-1,\alpha} \cdot \frac{s}{\sqrt{n}} ; \overline{x} + t^*_{n-1,\alpha} \cdot \frac{s}{\sqrt{n}}\right]$$

Voorbeeld: We zijn geïnteresseerd in hoeveel zakgeld zesdejaars op ORC per week krijgen van hun ouders. Een enkelvoudige aselecte steekproef bij 20 zesdejaars van het ORC geeft een gemiddelde van €22 en een steekproefstandaardafwijking van €3. Stel een 95% betrouwbaarheidsinterval op.

We vinden de kritieke t-waarde met onze rekenmachine: $t^*_{19,95} \approx 2,093$, zodat:

$$\left[22 - 2,093 \cdot \frac{3}{\sqrt{20}} ; 22 + 2,093 \cdot \frac{3}{\sqrt{20}}\right] \approx [20,596 ;\ 23,404]$$



Ook dit kan uiteraard sneller met de GRM. Merk op dat je om de kritieke t-waarde te bereken bij **invT** de linkerstaart moet ingeven (95% symmetrisch rond het gemiddelde wil zeggen 97,5% tot de waarde).

8) Toetsen van hypothesen

a) Concept

Met het toetsen van hypothesen tracht je op een statistisch verantwoorde manier aan de hand van een steekproef een uitspraak (*de nulhypothese*) over de populatie te aanvaarden of net te weerleggen (je aanvaardt dan een *alternatieve hypothese*). Statistisch verantwoord wil zeggen dat we met een bepaalde kans deze nulhypothese aanvaarden of net niet. De kans dat we de nulhypothese niet aanvaarden terwijl ze toch juist zou zijn noemen we het significantieniveau α van de toets.

We bekijken 4 inleidende voorbeelden om dit concept te verduidelijken.

b) Uitgewerkte voorbeelden

Een test met behulp van de binomiale verdeling

Voorbeeld: Een leerling haalt 7/10 op een multiple choice toets terwijl hij beweert op alles gegokt te hebben. Bij elke vraag waren er 3 antwoordmogelijkheden. Test zijn bewering op het 5% significantieniveau.

We berekenen de kans dat deze leerling door puur te gokken 7/10 haalt (of nog beter), onder de aanname dat hij inderdaad op alles gegokt heeft. De nulhypothese H_0 is dus de aanname dat hij alles zou gegokt hebben, en dat dus de stochast T die zijn punten aangeeft binomiaal verdeeld zou zijn. Anders gezegd geldt dan $H_0 : T \sim B(10,1/3)$. Is deze kans kleiner dan het significantieniveau dan verwerpen we de nulhypothese en aanvaarden we de alternatieve hypothese (H_a : hij heeft niet op alles puur gegokt maar had wel degelijk voorkennis, of dus $H_a : T \not\sim B(10,1/3)$).

De kans dat hij 7 of meer haalt is $P(T \geq 7) = 1 - P(T \leq 6) = 0,0197$ $(= 1 - \text{binomcdf}(10,1/3,6))$

Er is dus slechts 2% kans dat deze leerling 7 of meer zou halen als hij op alles had gegokt. We verwerpen dus deze hypothese op het 5% significantieniveau. (merk op dat we ze wel hadden aanvaard op het 1% significantieniveau).

De overschrijdingskans die we berekenden ($p = 0,0197$) noemen we de p *-waarde* van deze toets.

Een test in verband met proporties

(Bij het toetsen van hypothesen noteren we proporties met π in de populatie en $\hat{\pi}$ in de steekproef om verwarring met de p -waarde te vermijden).

Voorbeeld: Er is geweten dat 6% van de pasgeborenen in Vlaanderen een bepaalde afwijking heeft. Een gynaecoloog twijfelt hier echter aan en denkt dat dit bij hem in de praktijk niet zo is. Hij doet een steekproef bij 400 geboortes en komt uit dat bij hem slechts 17 van de baby's deze afwijking vertonen. Test de bewering '6% van de pasgeborenen in Vlaanderen heeft de afwijking' op het 10% significantieniveau.

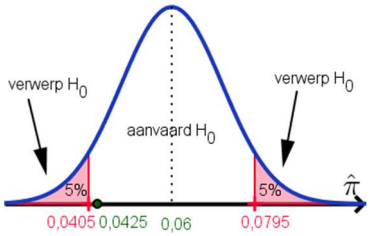
Wegens de centrale limietstelling is zijn de steekproefproporties $\hat{\pi}$ ongeveer normaal verdeeld met parameters $\mu = \pi = 0,06$ en $\sigma = \frac{1}{n}\sqrt{n\pi(1-\pi)} = \sqrt{\frac{0,06.0,94}{400}} \approx 0,01187$.

De nulhypothese is $H_0 : \pi = 0,06$, en de alternatieve hypothese is $H_a : \pi \neq 0,06$

De kans dat de proportie dus zo ver afwijkt van de gegeven waarde of nog extremer is, is gelijk aan:

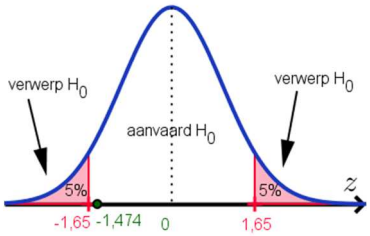
$p = 2.P(\hat{\pi} < 0,0425) \approx 0,1405$. We aanvaarden dus (nipt) de nulhypothese.

Alternatieve methode ①: Je kan de kritieke $\hat{\pi}$-waarden berekenen die het aanvaardingsgebied voor de nulhypothese afbakenen. Dit is makkelijk in te zien als je de verdeling van $\hat{\pi}$ grafisch voorstelt.



De kritieke grenzen zijn $\text{invnorm}(0.05, 0.06, 0.01187) = 0,0405$ en $\text{invnorm}(0.95, 0.06, 0.01187) = 0,0795$. Onze steekproefproportie ($\hat{\pi} = 0,0425$) ligt dus in het aanvaardingsgebied.

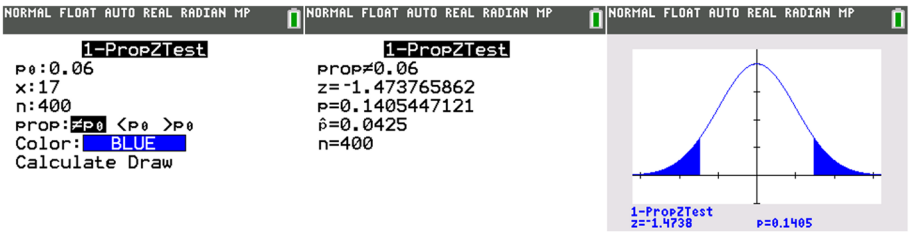
Alternatieve methode ②: Je kan de z-score van de geobserveerde steekproefproportie berekenen. Dit is $z = \dfrac{0,0425 - 0,06}{0,01187} \approx -1,474$.



De kritieke z-waarden zijn gekend voor 90%-intervallen, namelijk 1,65 en -1,65. Onze geobserveerde waarde ligt hierbinnen dus aanvaarden we de nulhypothese.

In het grootste deel van de literatuur wordt de voorkeur gegeven aan het tweede alternatief. De berekende z-score wordt dan de toetsingsgrootheid genoemd. Je kan deze bij deze soort toetsen bereken met de formule: $z = \dfrac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}}$.

Uiteraard kan de GRM dit heel snel met behulp van STAT – TESTS – **5:A-PropZTest**:



Belangrijk om op te merken is dat we hier tweezijdig hebben getoetst. Dat wil zeggen dat de alternatieve hypothese geen voorkeur uitdrukt voor groter of kleiner (> of <) maar enkel een verschil met de nulhypothese aanduidt (≠). Bij tweezijdige toetsen is het verwerpingsgebied in twee gedeeld (vandaar dat je bij het berekenen van de p-waarde ook maal twee moet doen). Had de gynaecoloog hier het vermoeden gehad dat er bij hem minder afwijkingen zouden zijn, dan was de alternatieve hypothese $H_a : \pi < 0,06$ geweest en dan hadden we de nulhypothese wel verworpen (want dan lag het verwerpingsgebied volledig links en dan lag onze geobserveerde steekproefproportie erin. De p-waarde had dan 0,0703 geweest).

Een pientere leerling kan hier ook opmerken dat we helemaal de centrale limietstelling niet nodig hebben, maar dat we exact kunnen werken met de binomiale verdeling.

De p-waarde (dus de kans dat we deze, of nog een extremere, steekproef zouden observeren) is:

$$p = 2.P\big(B(400\,;0,06) \le 17\big) \approx 0,1611.$$

De reden waarom er in de literatuur vaak benaderend wordt gewerkt is omdat die methode veel universeler is (wegens de centrale limietstelling) en het vaak makkelijker is om continu te werken dan om discreet te werken (dit geldt zeker als je deze materie in zijn historische context ziet).

Twee testen in verband met gemiddelden

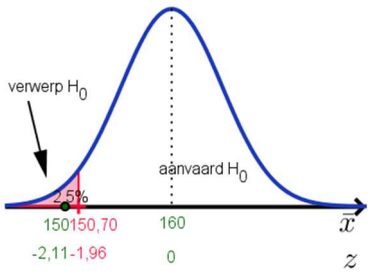
Voorbeeld 1: Kurt loopt al heel zijn leven marathons. De tijd (in minuten) waarin hij deze loopt is normaal verdeeld met een gemiddelde van $\mu = 160$ en een standaardafwijking van $\sigma = 15$. Een mental coach zegt dat hij hem sneller kan doen lopen. De volgende tien marathons (gesteund door de mental coach) loopt hij in gemiddeld 150 minuten. Test of de coach gelijk heeft met een significantie van $\alpha = 2,5\%$.

De nulhypothese is $H_0 : \mu = 160$, de alternatieve hypothese is $H_a : \mu < 160$ (eenzijdige toets).

We mogen hier de normale verdeling gebruiken want er is gegeven dat zijn tijden normaal verdeeld zijn, en dus is ook het steekproefgemiddelde normaal verdeeld.

Er geldt: $\bar{X} \sim N\left(160, \frac{15}{\sqrt{10}}\right)$. De p-waarde voor deze toets is de kans dat het gemiddelde 150 is of nog sneller: $p = P\left(\bar{X} < 150\right) = 0{,}0175 < 0{,}025 = \alpha$. We verwerpen H_0. De coach heeft gelijk!

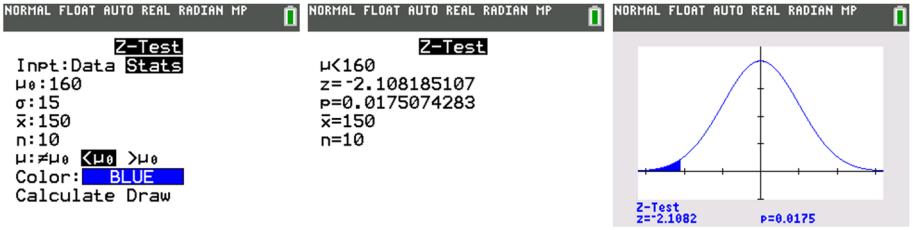
Alternatieve methode ①: De kritieke $\bar{x}$-waarde die het aanvaardings-gebied afbakent is $\text{invnorm}\left(0.025, 160, 15/\sqrt{10}\right) = 150{,}70$. Ons geobserveerd gemiddelde is nog sneller, het ligt in het verwerpingsgebied.



Alternatieve methode ②: De z-score van het geobserveerde steekproefgemiddelde is $\frac{150-160}{15/\sqrt{10}} \approx -2{,}11$. Dit is kleiner dan de kritieke z-waarde (-1,96) en ligt dus in het verwerpingsgebied.

Ook hier wordt meestal de voorkeur gegeven aan tweede alternatief. De toetsingsgrootheid kan deze bij deze soort toetsen berekend worden met de formule: $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$.

Uiteraard kan de GRM dit heel snel met behulp van STAT – TESTS – **1: Z-Test**:



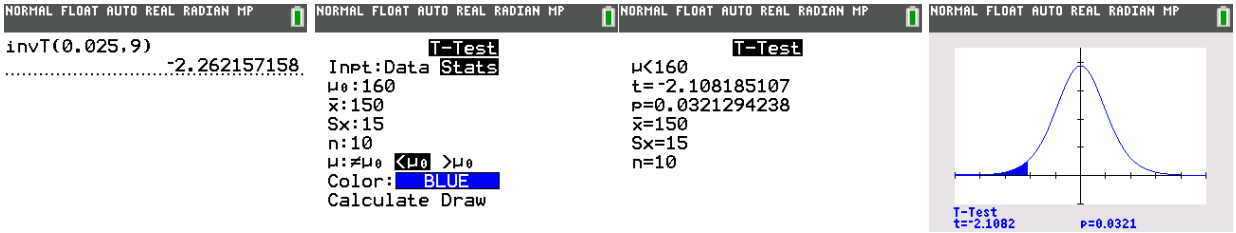
Voorbeeld 2: Dave loopt ook al heel zijn leven marathons. Zijn tijden liggen symmetrisch rond het gemiddelde (in minuten): $\mu = 160$. Een mental coach zegt dat hij hem sneller kan doen lopen. De volgende tien marathons (gesteund door de mental coach) loopt hij in gemiddeld 150 minuten, met een standaardafwijking van 15 minuten. Test of de coach gelijk heeft met een significantie van $\alpha = 2{,}5\%$.

De nulhypothese is $H_0 : \mu = 160$, de alternatieve hypothese is $H_a : \mu < 160$ (eenzijdige toets).

We moeten hier de t-verdeling gebruiken want we kennen enkel de stadaardafwijking van de steekproef en niet van de populatie (en het aantal steekproeven is niet groot genoeg om de centrale limietstelling te gebruiken ($n = 10 < 30$).

De t-score van het geobserveerde steekproefgemiddelde is $\frac{150-160}{15/\sqrt{10}} \approx -2{,}11$. Dit is groter dan de kritieke t-waarde (-2,26) en ligt dus in het aanvaardingsgebied. De p-waarde vinden we met de GRM: $p \approx 0{,}032 > 0{,}025 = \alpha$. Dus in dit geval geloven we de mental coach niet.

Uiteraard kan de GRM dit heel snel met behulp van STAT – TESTS – **2: t-Test**:



c) Soorten fouten

Bij het opstellen van de toets definiëren we een significantieniveau α. Dit is de kans op een fout van de eerste soort: het verwerpen van de nulhypothese terwijl ze juist is. Uiteraard bestaat ook het omgekeerde: het aanvaarden van de nulhypothese terwijl ze fout is. Dit heet dan een fout van de tweede soort. We noteren de kans hierop met β. Het complement $1-\beta$ wordt ook wel eens de *power* van de toets genoemd (dit is de kans dat je de nulhypothese verwerpt als ze fout is).

Voorbeeld: Boer Bavo beweert dat de aardappelplanten op zijn boerderij elk gemiddeld 1 kg aardappelen opleveren met een standaardafwijking van 0,2 kg (normaal verdeeld). Zijn vrouw Jeanine denkt dat dit minder is en doet een steekproef van 20 plantjes en vindt een gemiddelde van 0,85 kg.

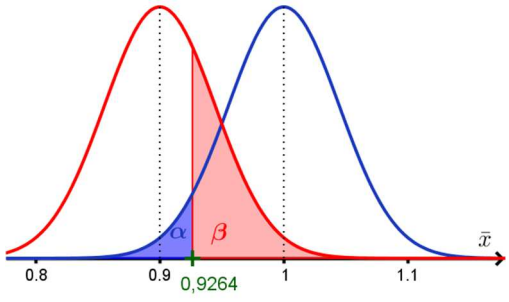
a) Voer een test uit op het 5% significantieniveau.

De kritieke $\bar{x}$-waarde die het aanvaardingsgebied afbakent is $0{,}9264$, want $\bar{X}$ is normaal verdeeld met parameters $\mu = 1$ en $\sigma = \dfrac{0{,}2}{\sqrt{20}}$, en $\text{invnorm}\left(0.05, 1, 0.2/\sqrt{20}\right) \approx 0{,}9264$.

We verwerpen dus de nulhypothese en aanvaarden het alternatief, namelijk dat $\mu < 1$.

b) Stel dat Jeanine gelijk heeft en dat het gemiddelde inderdaad minder is, namelijk 0,9 kg is, wat is dan de kans dat we toch de nulhypothese aanvaardden met deze toets?

We gaan dan uit van een andere kansverdeling voor $\bar{X} \sim N\left(0{,}9\,;\ \dfrac{0{,}2}{\sqrt{20}}\right)$ (getekend in rood).



De kans dat we de nulhypothese hadden aanvaard is aangeduid in rood, en bedraagt:

$$\beta = \text{normalcdf}\left(0.9264, 1E99, 0.9, 0.2/\sqrt{20}\right) \approx 0{,}27749$$

Dit is een vrij grote kans. Wat belangrijk is om te beseffen (en wat je eenvoudig kan zien op de figuur) is dat je nooit beide kansen α en β klein kan houden. Hoe kleiner je α maakt, hoe groter β wordt, en omgekeerd.