Handleiding Statistiek

Beschrijvende Statistieken / *Descriptives*

Toetsende Statistieken / *Inferential Statistics*

Voorspellingen / *Predictions*

Inzicht

Deel 1

Benjamin Telkamp

Voor Smalltown Boy en mijn Dochter Keya, die ik het meeste Inzicht gun en aan wie ik *altijd* denk als ik schrijf.

Colofon
tekst: Benjamin Telkamp
opmaak: Maarten van Maanen

Voorwoord

Inzicht = rust

Soms loop je tegen een probleem dat je niet zelf kan - of wil - oplossen. Sommige problemen zijn simpelweg te groot voor het nietige bestaan van een mens. Een statistiekvak kan zo'n probleem zijn. Inmiddels hebben duizenden studenten mijn statistieklessen gevolgd, in het begin vaak alleen omdat ze hun tentamen wilde halen. Maar ze kregen veel meer terug dan alleen een mooi cijfer. Het belangrijkste doel dat ik mensen wil meegeven, is Inzicht. Inzicht om zelf problemen te kunnen oplossen, dus een kritische manier van denken die je ontdoet van onnodige gedachtenkronkels (onzin) of ander onnodig emotioneel gezwam. Ik beoog een Inzicht dat je vooral ook bij het gewone leven helpt. Snel de essentie van een verhaal, probleemstelling of artikel doorhebben, helpt je bij het nadenken over een situatie. Statistisch Inzicht geeft je een nieuwe manier om de wereld - in en om je heen - handig te organiseren. Als je zaken georganiseerd hebt, betekent dat vaak ook dat je rust ervaart. *Onrust* is af en toe goed, omdat het je misschien kan aansporen om zaken te veranderen. Maar als je die zaken weer op een rijtje hebt en weer overzicht op je eigen leventje hebt, dan geniet ik vooral van de rust en vrijheid die me dat weer geeft. Zaken - in en om je heen - begrijpen kan heel bevredigend werken.

Inzicht = aandacht

Om een dieper inzicht in een verschijnsel te krijgen, zul je echt aandacht moeten geven. Om elkaar echt te begrijpen, zul je moeten doorvragen en echt proberen te luisteren. Ook hier is die vermoeiende eigenschap nodig. 'Read the manual' is mijn gedachte als mensen niet weten waar de aan- en uit knop zit van hun telefoon (bij mij staat ie nog steeds aan gelukkig). Een vrije val of zomaar ergens inspringen kan dan leuk zijn, maar het soort vallen bij tentamens - hoe leerzaam falen soms zelfs kan zijn - laat je beter achterwege. Ik heb mijn *gebrabbel* uit mijn lessen zo duidelijk en bondig mogelijk op schrift gezet in de hoop dat jullie er net zoveel aan hebben als dat het normaal gezien, uit mijn mond komt. In mijn klassikale lessen kan ik zien wie het wel en niet snapt en met mijn aandacht daarvoor corrigeren. De handleiding ligt nu voor je neus en ik kan - helaas - dat contact niet aangaan. *Mission Failed*? Niet geheel, want er zijn meer soorten aandacht, en hopelijk zul je daar één of meer van terugvinden als je dit leest. Mocht je het *voelen* dan is het ook *echt* zo.

Inzicht = fun

Wat mij betreft, is het juist de *oplossing* die bevredigt en moeilijk-doenerij niet. Iedereen die een sudoku heeft ingevuld en kloppend heeft afgemaakt, voelt iets van blijheid of trots. Mensen geven mij weleens het excuus; '... lukt niet, want ik denk te moeilijk' (alsof ze van een hogere cognitieve orde zijn en daar stiekem een beetje trots op zijn). Methodologisch redeneren bevordert het vinden van tastbare oplossingen. Aan de hand van Inzicht vind je dus sneller je oplossingen. (Overheids) instellingen hebben er nog wel eens een handje van niet eerst na te denken, neem project USIS, met Inzicht hadden ze niet zulke grove fouten gemaakt en had dat natuurlijk miljoenen gescheeld. En voor werknemer en student: uren minder frustratie. Gewoon eerst even écht nadenken en niet met de natte vinger dus, een goeie start is het halve werk en het werk moet af, want ik wil vooral plezier in mijn leventje.

Handleiding der handleidingen en het absoluut minimale dat je móet weten.

Je zal snel ontdekken dat bij het vak statistiek, het niet het rekenwerk is dat het meest ingewikkeld is, maar vooral het taaltje dat we hier spreken. Mijn ervaring is dat het juist de taal is die vaak niet wordt begrepen en tot problemen leidt. Als je zegt (of afspreekt) dat '2 maal 4' 8 is dan is dat eigenlijk makkelijker te begrijpen dan een uitspraak als '80 procent van de variatie op lengte wordt verklaard door leeftijd' of 'teamwork is goed voor het algemeen welzijn'. Zulke stellingen roepen meer vragen op dan dat we iets werkbaars hebben. En zo is het ook binnen de statistiek: Voor de berekeningen hebben we slechts een 'aantal' afspraken nodig, maar voor de taal die we gebruiken, zijn dat er velen malen meer. Natuurlijk maken we veel gebruik van synoniemen en het is dus ook zaak voor je tentamen dat je definities in verschillende vormen of situaties kan herkennnen. Een definitie is vaak kort en bondig en daarom vaak juist onbegrijpelijk. Mocht ik definities gebruiken dan maak ik veel gebruik van haakjes met daartussen extra tekst die de definities verduidelijken. Lees definities een keer met de woorden tussen haakjes en een keer zonder, zodat je weet dat je beide vormen snapt. Belangrijke definities zijn vaak apart uitgelicht in deze handleiding, zodat je er - eventueel - sneller doorheen kunt gaan. Verder heb ik definities proberen toe te lichten met voorbeelden van (verkeerd) gebruik, en toepassingen. Door begrippen in een andere bewoording te herhalen (wat ik ook altijd in mijn lessen doe), hoop ik dat je in het gewone leven, eerder een situatie of een stuk tekst - zal herkennen of begrijpen en je tentamen er uiteindelijk fluitend, in trapt. Alle rekenopgaven heb ik tot in de puntjes - met tussen oplossingen - en vaak op verschillende manieren uitgewerkt en vind je op aparte bladzijden na ieder onderwerp. Je hoeft dus altijd maar een paar bladzijden verder te bladeren voor een verlossend antwoord of uitwerking. Natuurlijk kan je op verschillende manieren iets intypen op je rekenmachine en ik laat dus ook verschillende manieren zien. Vaak hebben mensen hun eigen manier, maar ik denk als je mijn manieren snapt, je ook weer de gewone rekenregels beheerst en vooral de theorie beter snapt. Er zit meer theorie in die formules verstopt dan je vaak denkt. Mijn advies is dus: lees gewoon alles, van begin tot eind, dan weet je ook waar de aan en uit knop zit.

De essentie van waar we mee bezig zijn, in het leven en in de wetenschap.

Tijdens mijn studie kwam ik er snel achter dat de *huidige* wetenschap niet álles verklaren kan. Sterker nog: Wetenschap kwam me vaak ietwat 'religieus' over. Heel veel (vreemde) begrippen hebben een *bestaansrecht* omdat ze *bruikbaar* zijn, maar vaak weet niemand *precies* wat er met zo'n woord of begrip wordt bedoeld, of waar een begrip naar verwijst. Denk aan (psychologische) termen zoals 'Ego', 'de wil' of 'het bewustzijn' of aan normale termen als 'asbak', 'koe' of 'tafel' behoorlijk vaag eigenlijk. Wat is nou een typische tafel? Voor velen dient de straat ook als een asbak. We hanteren dan ook vaak afspraken over hoe en wat we bedoelen. Een wetenschapper kan niet alles weten en daarom heeft een wetenschapper - noodzakelijkerwijs - zo zijn 'aannames', ook wel een verzameling stellingen (ideeën) waarvan men eigenlijk niet weet of ze wel echt waar zijn, maar voorlopig wel even voor lief aanneemt (er voorlopig dus even in gelooft). Heel veel stellingen, zoals 'tijd gaat vooruit' mogen dan misschien met onze alledaagse ervaringen overeenkomen, maar dat wil nog niet zeggen dat die stelling ook waar is. Zo dacht men vroeger ook dat de zon om de aarde heen draaide. In de wetenschap is slechts een persoonlijke of subjectieve *ervaring* niet voldoende om een stelling te bewijzen. Descartes begon met de aanname: 'Ik denk, dus ik ben' (hij moest iets zeggen over het al dan niet bestaan van zijn persoon en had daar een bewijs voor nodig). Wat mij betreft een zinloze en rare aanname. ik wil graag een nieuwe - zinvolle - aanname of uitgangspunt hiervoor in de plaats. Een uitgangspunt dat meteen ook de essentie van statistiek (of wetenschap en het leven) raakt en die dus super handig is als je even - door de cijfers - de juíste getallen niet meer ziet.

Om antwoord hierop te vinden moeten we ons eerst de vraag stellen wat onderzoekers en pasgeboren baby's met elkaar gemeen hebben. Beiden tasten hun omgeving af zodat ze in staat zijn om die omgeving of werkelijkheid uiteindelijk te beheersen, of te manipuleren. *So far, so good*. Mijn volgende en belangrijkste vraag is: Wat hebben de omgeving (werkelijkheid) van de baby en de omgeving van de onderzoeker gemeen? Dus ook al zitten de twee in een totaal andere omgeving, wat is er toch gelijk of welke ervaring zullen beiden in ieder geval delen.
Dit is een diepe, maar heeft volgens mij maar één antwoord. Zowel baby als wetenschapper zullen 'verschillen' in hun omgeving ervaren. Een baby zal voelen dat het niet overal even warm is, of zal zien dat het niet altijd even licht is en zal niet altijd even hongerig zijn. Zijn vader en moeder zullen niet hetzelfde stemgeluid hebben en de baard van pappa voelt toch echt wel spannender aan dan die zachte wangetjes van mama. De onderzoeker ervaart ook verschil. Hij zal observeren dat niet iedereen in zijn steekproef hetzelfde is. Niet iedereen zal even oud, blij, lang, slim, creatief, gemotiveerd of wat dan ook zijn. Een van de kerntaken van wetenschap (en statistiek) is juist het beschrijven, voorspellen en verklaren van die *verschillen* in onze omgeving. Zo heeft een baby na korte tijd ook wel door dat een hogere stem samengaat met een zacht wangetje en een bromgeluid met een baard. We ervaren allemaal verschil en we gebruiken die verschillen om te voorspellen. Ik denk zelfs dat je kunt stellen dat verschil doet leven en dus daarom:

Ik ervaar verschil, dus ik ben

Je komt een hoop moeilijk-doenerij tegen in de wetenschap (terwijl wetenschappers het begrip *parsimony* - ook wel spaarzaamheid of simpelheid - hoog *zouden* hebben moeten zitten). Zo zegt men gek genoeg weleens in de statistiek;

- geslacht heeft een effect op lengte
- er is samenhang tussen geslacht en lengte
- geslacht en lengte zijn gecorreleerd
- lengte is geassoccieerd aan geslacht

Vier hele correcte uitspraken, maar ze betekenen allemaal hetzelfde. Wat met deze uitspraken wordt bedoeld is - enkel en alleen - dat mannen over het algemeen een *andere* lengte hebben dan vrouwen, dus dat mannen gemiddeld gezien verschillen qua lengte van vrouwen. Er is dus een (systematisch) *verschil* tussen mannen en vrouwen qua lengte. Wij weten zelfs (uit *ervaring* of onderzoek) dat mannen meestal langer zijn dan vrouwen (dit blijkt niet uit bovenstaande uitspraken, die zeggen alleen maar dat er verschil is, maar niet welke kant dat verschil op gaat). *Anyway*, je zal dus een hoop moeilijke begrippen tegenkomen, maar mocht je even in de war raken, besef dan dat bij statistiek altijd alles om *verschil* gaat. Verschillen op het een, geeft vaak een verschil op het ander.

Aapjes

Om verschillende technieken of analyses uit te leggen gebruik ik in de les altijd 'mijn aapjes'. Mijn aapjes houden de boel tastbaar en daar wordt mens, kind en dier blij van. Dus ook voor jullie voer ik mijn aapjes door, zodat jullie er ook van kunnen genieten. Na een paar oefeningen zal je zien dat je de aapjes wel kunt dromen en tijdens tentamens komen ze goed van pas. Als je het even niet meer weet, kun je je hoogst waarschijnlijk wel weer aan hun op trekken. Je hebt dan één 'kant en klaar voorbeeld' in je hoofd zitten (met antwoorden) die meteen op de meeste zaken toepasbaar is. Je zult alleen nog even de getallen moeten veranderen. Het zijn mijn 9 aapjes die ik ooit in een *bruin* verleden heb gevangen (ik doe zo min mogelijk aan ethiek in deze handleiding) omdat ik me verbaasde over de verschillen die ik bij hun zag. Het was een heerlijke verwondering, want deze lieve aapjes varieerde op lengte, leeftijd en aapsoort. Hoe mooi kan variatie zijn! Maar goed, later dus veel meer over de apen.

Onderzoek en wat algemene definities voordat we aan de slag gaan met het echte rekenwerk, het *absoluut minimale* dus.

Wetenschappers houden zich vooral bezig met het doen van onderzoek, maar wat is onderzoek in het algemeen en wat doen onderzoekers nou eigenlijk feitelijk?

Onderzoek

Definitie: Het ontdekken, beschrijven en verklaren van (observeerbare of meetbare) verschijnselen, patronen of relaties (zoals gedrag en mentale processen) in de werkelijkheid. Met de werkelijkheid bedoel ík alles wat maar 'waar te nemen' valt (observeren) in onze omgeving, dus een of meer mensen voldoen, maar ook apen, hersenen, nieren, een stad of iets anders in ons Heelal. De werkelijkheid is dus een heel ruim begrip hier.

Onderzoeksobject

Definitie: Het onderzoeksobject is die of datgene van wie of wat je informatie verzamelt voor een onderzoek. Het onderzoeksobject is dus het ding, zaak of dus object dat wordt onderzocht binnen een onderzoek en aan wie (of wat) dus de observeerbare verschijnselen (informatie) toebehoren. Binnen een onderzoek kan het één object zijn, maar meestal zijn het er meerdere.

Voorbeelden: meestal mensen (of dus slechts één mens), proefpersonen of proefdieren, maar soms ook een bepaalde dag of dagen in het jaar, landen of andere objecten zoals een school, ziekenhuis, gevangenis of één of meer steden.

Gebruik: Een onderzoek kan zich richten op Nederlandse adolescenten (onderzoeksobjecten) en hun vrijetijdsbestedingen (observeerbare verschijnselen).

Een variabele

Definitie: 'iets dat varieert', 'iets of verschijnsel dat een bepaalde grootte of waarde kan aannemen en dus verschillend qua waarde of grootte kan zijn', een (bepaald soort) grootheid, een (bepaald soort) dimensie.

Voorbeelden: geslacht, lengte, leeftijd, *soort* depressie, maar ook *mate* van depressie, opleidingsniveau, economische status, nationaliteit, kans op slagen of kans op ziek worden, temperatuur, bloeddruk, aapsoort.

Gebruik: Eén persoon kan op één moment niet verschillende lengtes hebben, het is de variabele 'lengte' die bij *verschillende* personen *verschillende* waarde zal aannemen.

Waarde of categorie (zelfde)

Definitie: een getal of naam dat kan worden toegekend aan een eigenschap van een zaak, ding of object.

Voorbeelden: een lengte van 172 cm, een IQ van 130 of een bipolaire depressiestoornis.

Let op bij gebruik: Waarden én Categorieën zijn dus niet hetzelfde als variabelen. Een variabele of dimensie kan dus wel een bepaalde waarde of categorie aannemen. Het is de variabele die (op een bepaald moment) een (bepaalde) waarde draagt (of aanneemt).

Gebruik:
- De variabele 'geslacht' kan de twee waarden (of categorieën) 'man' of 'vrouw' aannemen (alsjeblieft geen ethiek hier, maar natuurlijk erkennen we tegenwoordig meerdere soorten geslachtsvormen)
- De variabele 'lengte' neemt bij pasgeboren babies (objecten) meestal een waarde aan ergens tussen 30 en 60 cm.
- De meeste mensen die een universitaire opleiding (waarde op de variabele opleidingsniveau) hebben afgerond, scoren hoger op cognitieve dimensies (verschillende variabelen) dan mensen

met een lager opleidingsniveau (waarde).
- Vandaag (object) is de temperatuur (variabele) 30 graden Celsius (waarde).
- Een onderzoek richt zich op de relatie tussen studiekeuze (variabele) en het soort bijbaantjes (variabele) bij Nederlandse adolescenten (onderzoeksobjecten).
- Mijn onderzoek richt zich op 'het verband tussen leeftijd en lengte bij apen'. De 'onderzoeksvraag' is hier eigenlijk of de verschillen op leeftijd systematisch samen gaan met verschillen op lengte bij apen. Of makkelijker gezegd: 'Of ze dus groeien naarmate ze ouder worden'. De wijsneus, ik dus, zou meteen vragen, maar waarom 'zonodig' *omhoog* groeien en niet *omlaag*? Later gaan we hier moeilijk over doen, wees gerust.

Observatie

definitie: Meting (bepaling aan de hand van een meetinstrument) van een bepaalde waarde op een variabele, toebehorend aan een onderzoeksobject.

voorbeeld:
- We kunnen observeren (meten of bepalen) of een bepaald persoon (onderzoeksobject) een man (één van de twee waarden die de variabele geslacht kan aannemen) dan wel een vrouw (de andere waarde of categorie) is.
- Aan de hand van een IQ-test (meetinstrument) observeren of meten we hoe hoog een bepaald persoon scoort.

Apen, even warm worden

Voordat we inhoudelijk naar theorie en analyses gaan kijken, wil ik dus eerst even mijn apen introduceren. Op basis van de gegevens (data) gaan we vast wat rekenen, zodat we wat rekenregels tegenkomen die later van pas zullen komen. Ook tijdens deze berekeningen behandel ik al belangrijke theorie, maar ik zal nog niet *alle* begrippen die ik hier gebruik uitleggen, soms doe ik dat pas verderop in deze handleiding. Maar ook hier genoeg theorie voor je eerste tentamen statistiek. Het is voor nu ook nog even niet nodig dat je alle begrippen meteen snapt, alles op z'n tijd en ik wil nu graag puur even rekenen en wat extra aandacht aan de rekenregels geven.

De dataset

In onderstaande tabel vind je de scores (datapunten) die ik verzameld heb tijdens mijn onderzoek naar aapjes. Mijn 9 aapjes vormen een steekproef en heb ik eerlijk geselecteerd uit de hele populatie (ergens in land waar je heel lekker kunt eten, Indonesië ofzo). Je ziet 3 variabelen (de drie kolommen) in deze tabel. De eerste variabele 'i' - spreken we even af - noemen we respondent of *case*-nummer en is alleen maar om mijn andere - of echte - waarnemingen of observaties te organiseren (ik heb dus alle aapjes genummerd van 1 tot en met 9). In de tweede kolom vind je 9 scores qua 'lengte in cm'. Uit luiheid gebruiken we eigenlijk altijd een letter voor een variabele. Hier staat Y_i voor de score qua lengte in cm voor het 'i-de' aapje. Je mag dus zeggen dat de score qua lengte voor bijvoorbeeld het tweede aapje ook wel te schijven is als $Y_2 = 130$. In de derde kolom vind je de variabele X_i die in *dit geval* voor de 'leeftijd in jaren' staat (voor aapje nummer i welliswaar). Dus $X_9 = 2.0$ betekent alleen maar dat aapje nummer 9 een leeftijd heeft van 2.0 jaar.

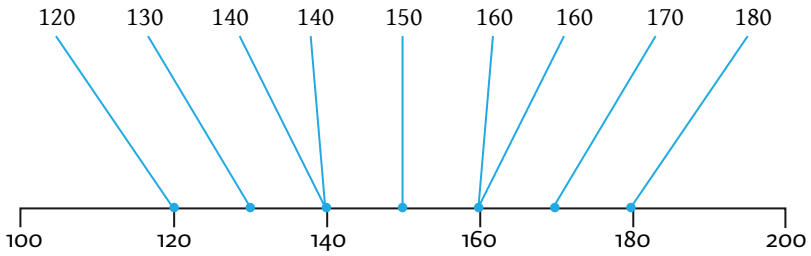
TABEL 0A

i	Y_i	X_i
1	120	1.0
2	130	1.0
3	140	1.0
4	140	1.5
5	150	1.5
6	160	1.5
7	160	2.0
8	170	2.0
9	180	2.0

$n = 9$

We gaan onze eerste statistieken uitrekenen: het *gemiddelde* en *standaarddeviatie* voor de variabele Y (lengte in cm), in symbolen ook wel $\bar{Y}$ en S_y. Verder heb ik ook de scores voor lengte even grafisch weergeven aan de hand van een getallenlijn. Je ziet maar zeven punten op de getallenlijn in plaats van negen, maar dat komt dus omdat we twee keer een dubbele waarneming hebben, bij 140 en 160 cm.

figuur 0.A



Het gemiddelde is ook wel de verwachte waarde (voor een variabele). Het gemiddelde is een *statistiek* (een beschrijvend getalletje) die het centrum van een verzameling datapunten aangeeft waardoor we ook wel de *ligging* (of positie) van onze scores weten. Als je aan een getallenlijn denkt (zie figuur 0.1) dan is het gemiddelde ook wel een soort plaatsbepaling (van het centrum van de datapunten). Naast die plaatsbepaling (beschrijving van positie) heeft het gemiddelde nog een andere belangrijke rol. Je zou kunnen zeggen dat het gemiddelde ook wel de beste gok is als je een voorspelling wilt doen. Dus als één van onze negen aapjes binnen zou komen wandelen, wat zou dan je verwachting of voorspelling zijn qua lengte? Vandaar dus ook de naam 'verwachte waarde' voor het gemiddelde. Omdat het gemiddelde je beste gok is (bij gebrek aan andere informatie), kun je zeggen dat het gemiddelde je meest basale voorspelling, theorie of (voorspel)model is. Termen als 'intercept model' of het 'nul-model' zijn ook veel gebruikte termen voor het gemiddelde. Als je een model maakt, probeer je alleen maar de realiteit of de werkelijkheid te benaderen of weer te geven. Een model stelt meestal een representatie (soort kopietje) van iets anders (werkelijkheid) voor. Sommige modellen zijn complex zoals een regressiemodel met tien voorspellers om het algemeen welzijn van een persoon te voorspellen, een maatpak (helemaal op maat gemaakt en representeert de vorm van een lichaam), een landkaart met alle wandelwegen van Nederland of zelfs een fotomodel (die modelleert om schoonheid te representeren) en sommige modellen (of theorieën) zijn heel simpel zoals het gemiddelde zelf, een spijkerbroek, het liefst een Levi's 28/32 (28 staat hier voor de breedte en 32 voor de lengte en met slechts twee 'parameters' (beschrijvers) weet de winkelier dus al genoeg en pakt dan zo de juiste spijkerbroek (model) uit de kast), een speelgoedautootje, of een plattegrond van je schoolgebouw. Een model is dus een ruim begrip, maar de modellen die we in de statistiek bouwen zijn eigenlijk altijd bedoeld om verschijnselen (in de werkelijkheid) te beschrijven of te voorspellen.

De berekening van het gemiddelde voor een variabele (*mean, average, expected value*).

In woorden zou de berekening voor het gemiddelde van variabele Y zijn: *eerst* alle scores optellen en *dan pas* delen door het aantal. Je kan het ook moeilijk(er) zeggen: om het gemiddelde te vinden, deel je de sommatie van alle scores door het aantal waarnemingen in je steekproef.

in formule vorm:

$$\bar{Y} = E(Y) = \frac{\sum_{i=1}^{i=n} Y_i}{n}$$

We komen dus nu voor het eerst het 'sommatieteken' tegen $\sum_{i=1}^{i=n}$. Officieel heet dit teken ook wel 'sigma', maar die term ga ik niet gebruiken (omdat de standaardafwijking ook die naam draagt) en gebruik ik dus gewoon de het woord 'sommatieteken'. In deze formule zie je een onderschift "i=1" en een bovenschrift "i=n" bij het sommatieteken. Heel vaak laten ze onder en bovenschrift

weg, dat doe ík nog even niet, ik wil graag dat je beseft waar ze voor staan. 'i' staat bij ons nu voor respondentnummer en de letter 'n' staat voor de totale steekproefgrootte, dus 9 bij ons, of ook wel de hoogst mogelijke waarde voor i. Later krijgen de i-tjes een andere betekenis (bijvoorbeeld groepsnummer) en hebben we ook j-tjes en k-tjes nodig om de boel te organiseren en laat ik ze nu dus staan, zodat je eraan kunt wennen. De formule staat nu in een breukvorm. Het bovenste gedeelte van de breuk (de teller, *numerator*) is dus $\sum_{i=1}^{i=n} Y_i$ en het sommatieteken hierin vertelt ons dus dat we *eerst* datgene dat achter het sommatieteken staat (hier alleen Y_i), voor elke waarde van i (beginnend bij $i=1$ en eindigend bij $i=n$, dus $i=9$ bij ons) moeten invullen (vervangen) en daarna pas deze negen waarden moeten optellen

$$\sum_{i=1}^{i=n} Y_i = 120 + 130 + 140 + 140 + 150 + 160 + 160 + 170 + 180 = 1350$$

De som of sommatie van alle scores is dus (heeft een waarde van) 1350. Stel dat ik alleen de middelste drie lengtes van mijn aapjes zou willen optellen dan zouden dus alleen het onder en bovenschrif veranderen:

$$\sum_{i=4}^{i=6} Y_i = 140 + 150 + 160 = 450$$

Maar we waren er nog niet want in het onderste gedeelte van de breuk (noemer, *denominator*) stond ook nog een 'n'. Die staat vaak voor de totale steekproefgrootte, maar soms kom je een kleine letter n én een grote letter N tegen binnen één onderzoek. In dat geval bedoelen ze met de kleine n de groepsgroottes van de subgroepen binnen je steekproef (aantal mannen en vrouwen) en de grote N voor totale steekproefgrootte (aantal mensen). *Anyway*, we moeten de sommatie van de scores (1350) dus nog delen door 9 en we hebben de gemiddelde waarde gevonden. Natuurlijk reken je vaak met tussen-antwoorden en type je dus niet altijd de hele berekening in één keer in. Om het formule gevoel toch een beetje op te krikken, schrijf ik het toch even op zoals je het allemaal in één keer zou moeten intypen:

$$\bar{Y} = E(Y) = \frac{\sum_{i=1}^{i=n} Y_i}{n} = [120+130+140+140+150+160+160+170+180] / 9 = 150$$

Ik gebruik altijd blokhaken om aan te geven dat ik een sommatieteken uitwerk. Met je rekenmachine type je (natuurlijk) gewone haakjes in plaats van blokhaken. De intype-manier wordt dus als volgt:

$$\bar{Y} = E(Y) = \frac{\sum_{i=1}^{i=n} Y_i}{n} = (120+130+140+140+150+160+160+170+180) / 9 = 150$$

Het gemiddelde voor lengte (voor deze steekproef) is dus 150 en is dus ook de verwachting of voorspelling als je wilt voorspellen wat de waarde van een aapje zal zijn, als er een willekeurig aapje binnen komt lopen.

Soms kom je de formule voor het gemiddelde in een andere vorm tegen. En omdat we later met veel moeilijkere formules moeten werken, wil ik dat je beide vormen even goed snapt en kunt toepassen. De moeilijke versie ziet er als volgt uit.

$$\bar{Y} = E(Y) = \frac{1}{n} \cdot \sum_{i=1}^{i=n} Y_i$$

Mischien ken je de regel 'delen door een getal is hetzelfde als vermenigvuldigen met het omgekeerde'. Neem bijvoorbeeld 8/2= $\frac{8}{2}$=4. We delen dus hier het getal 8 door het getal 2. Wat de regel eigenlijk zegt, is dat je 8 ook kunt vermenigvuldigen met het omgekeerde van het getal 2. Het 'omgekeerde' van het getal 2 is $\frac{1}{2}$ en het omgekeerde van bijvoorbeeld 100 is ook wel 1 gedeeld door 100 of dus $\frac{1}{100}$ (één honderdste). Dus je had ook $8 \cdot \frac{1}{2}$ =4 kunnen doen of draai het om $\frac{1}{2} \cdot 8$=4. Je ziet dus dat ik het 'keer-teken' met een puntje doe, maar als ik het uitschrijf voor

de rekenmachine zal ik een "*" gebruiken voor het maalteken. De ouderwetse keer 'x' kunnen we nu niet meer gebruiken omdat de meeste variabelen uit luiheid of compactheid de naam 'X' krijgen en we willen zo min mogelijk verwarring.

TABEL 0B

getal/waarde	omgekeerde in breukvorm	omgekeerde in 2 decimalen
2	$\frac{1}{2}$	0.50
4	$\frac{1}{4}$	0.25
100	$\frac{1}{100}$	0.01
3	$\frac{1}{\frac{1}{3}} = 3$	3.00
$\frac{2}{3}$	$\frac{1}{\frac{2}{3}} = \frac{3}{2}$	1.50
n	$\frac{1}{n}$	dit kun je pas uitrekenen als je de waarde van n weet
n-1	$\frac{1}{n-1}$	dit kun je pas uitrekenen als je de waarde van n weet
Benjamin	$\frac{1}{Benjamin}$	dit kun je pas uitrekenen als je de waarde van Benjamin weet

Dus het liefst had je de formule als volgt ingevuld:

$$\bar{Y} = E(Y) = \frac{1}{n} \cdot \sum_{i=1}^{i=n} Y_i = \frac{1}{9} \cdot [120+130+140+140+150+160+160+170+180] = 150$$

of qua intypen op je rekenmachine:

$$\bar{Y} = E(Y) = \frac{1}{n} \cdot \sum_{i=1}^{i=n} Y_i = 1/9*(120+130+140+140+150+160+160+170+180) = 150$$

Merk op dat ik 1/9 niet tussen haakjes heb gezet, ik weet dat jullie gek zijn op haakjes, maar ik doe het alleen als het *nodig* is. Denk dus ook na als ik géén haakjes gebruik en type alsjeblieft de formules letterlijk in zoals ik ze uitschrijf (en zie dan dat het blijkbaar zo mag, tenzij je een rekenmachine uit de tijd van *Kniertje* hebt, maar dan zal je een nieuwe moeten halen; een zogenaamd wetenschappelijk rekenmachientje of het liefst een Grafische Rekenmachine, Texas TI (nog wat), want daar heb ik allemaal geweldige trucjes voor die ik in appels en peren zal uitleggen) Het leren lezen van woorden is één ding, maar het lezen van formules is van een hele andere orde. Soms zul je dus gewoon - symbool voor symbool - een formule moeten *uit*spellen tijdens het overnemen van een formule in je *ruitjes*-schrift.

Standaardafwijking, standaarddeviatie, *standarddeviation*

De volgende - en misschien wel de meest belangrijke - statisitiek die we gaan berekenen, is de standaardafwijking, ook deze beschrijft een karakterestiek (eigenschap) van een verzameling scores voor een variabele, bij ons de variabele lengte (Y_i) dus. De standaardafwijking is een spreidingsmaat en vertelt je in hoeverre de scores (van een variabele) juist bij elkaar of juist uit elkaar liggen. Als je naar de getallenlijn kijkt, gaat het dus nu om de concentratie van datapuntjes. Als de punten dicht bij elkaar liggen, is er weinig spreiding en heeft de standaardafwijking een relatief lagere waarde dan als de punten juist verder uit elkaar liggen.

Definitie: De standaardafwijking is de gemiddelde afwijking (afstand of verschil) van een observatie (of score) naar het gemiddelde.

In gewone woorden zou je ook wel kunnen zeggen dat de standaardafwijking, de grootte van de gemiddelde gokfout (afstand, verschil, afwijking) is, als je de scores van jouw variabele probeert terug te voorspellen (gokken) met het gemiddelde als beste gok. Als ik standaardafwijking zeg, denk ik vaak gewoon: de gemiddelde gokfout (als je het spelletje zou spelen, dus zou gokken of voorspellen wat de lengte van een aapje is voordat hij binnenkomt en het gemiddelde als beste gok gebruikt, ik noem dit 'het spelletje').

Hier gaan we even dieper (en makkelijker) over nadenken door het spelletje te spelen. Stel je voor dat *onze* 9 aapjes op de gang staan en dat er willekeurig (je weet niet welke) één aapje binnen komt wandelen. Je hebt de data inmiddels gezien, je weet dus ook welke 'waarden' binnen zouden kunnen komen wandelen. Inmiddels weet je ook dat het gemiddelde je beste gok is, je weet misschien nog niet waarom, maar dat wordt nu hopelijk duidelijk. Stel, jij zegt dus: 'Het aapje dat binnenkomt, zal wel 150 cm zijn'. Er is maar één aapje precies 150 cm lang, dus een grote kans dat je precies goed gokt (en er dus 0 cm naast zit met jouw gok of verwachting), heb je niet (die kans is slechts 1 op 9 of 1/9 of 0.11111....). Maar daar gaat het nu ook niet om (gek genoeg). Het gaat dus niet om 'zo *vaak* mogelijk *precies* goed gokken', maar juist om '*gemiddeld* gezien er zo dicht bij in de buurt komen'. En daarop moet je beste keuze - qua voorspelling - gebaseerd zijn. Je wilt - gemiddeld genomen - de kleinst mogelijke gokfout weten, als je het spelletje herhaald. En als je dus het gemiddelde kiest als beste gok, dan is je gemiddelde gokfout - dus de standaardafwijking - het kleinst.

Goed, we gaan hem - de standaardafwijking - berekenen, in dit geval dus voor de variabele Y (lengte in cm). We bouwen het langzaam op, want tijdens onze berekening komen we ook weer wat rekenregels en theorie tegen die je later weer zal moeten gebruiken, dus blijf opletten.

Individuele afwijking

Definitie: Een individuele afwijking van een observatie naar het gemiddelde is de afstand van de waarde van een waarneming naar het gemiddelde (denk dus een individuele gokfout).

Wij spreken af dat als een score boven (of rechts van, als je aan de getallen lijn denkt) de verwachting (gemiddelde) ligt, die score een positieve afwijking heeft ten op zichte van het gemiddelde. Als de score onder of links van de verwachting ligt, noemen we het dus een negatieve afwijking. Om de afstand dus juist te berekenen, neem je altijd de waarde van de observatie en daar trek je de verwachting van af. Onthoud vast: een verschil (afstand tussen) is altijd 'specifiek min algemeen' en een observatie is natuurlijk veel specifieker dan de verwachting (het gemiddelde). Maar dus altijd, in deze volgorde.

$$\text{Residual} = \text{Observed} - \text{Expected}$$

Een residu is ook wel iets dat je overhoudt (in dit geval verschil of afstand) na een bepaalde behandeling (een aftrekking), je hebt meerdere soorten residuen, maar dat is nu nog even niet aan de orde. Voor nu zou je een (individuele) gokfout dus ook wel een residu kunnen noemen.

Laten we vast alle residuen (ten op zichte van het gemiddelde uit rekenen).

$$\text{Residual} = \text{Observed} - \text{Expected} = Y_i - \bar{Y}$$

De afwijking van aapje nummer 1 is $Y_1 - \bar{Y} = 120 - 150 = -30$ en is dus een negatieve afwijking. Voor respondentnummer 1 kan je dus zeggen dat de gokfout een waarde heeft van -30 en bijvoorbeeld voor aapje nummer 8 dus 20. Omdat we straks gaan optellen en we telkens gelijksoortige handelingen gaan doen, zet ik de resultaten vast in kolomen in de tabel hieronder. Als je naar de de individuele afwijkingen kijkt dan zie je dat de kleinste afstand 0 is (voor aapje nummer 5, want die ligt precies op het gemiddelde) en de grootste afstand 30 of - 30 is. De

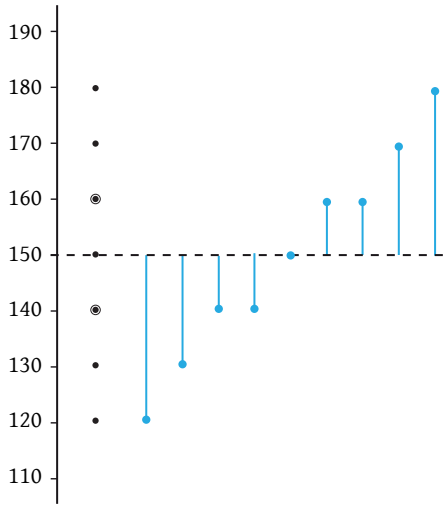
standaardafwijking is de gemiddelde afwijking (van een observatie naar het gemiddelde). Dus je zou misschien zeggen dat als je alle individuele afwijkingen bij elkaar optelt en vervolgens deelt door het aantal, dan heb je de standaard afwijking gevonden. Maar helaas, zo werkt het dus niet (maar ik zou het wel zo voelen als ik jou was) We komen nog twee of drie problemen tegen waarvoor we nog moeten corrigeren (een oplossing voor moeten vinden).

TABEL 0C

i	Y_i	$Y_1 - \bar{Y}$	$(Y_1 - \bar{Y})^2$
1	120	-30	900
2	130	-20	400
3	140	-10	100
4	140	-10	100
5	150	0	0
6	160	10	100
7	160	10	100
8	170	20	400
9	180	30	900
	$\sum_{i=1}^{i=n} Y_i = 1350$	$\sum_{i=1}^{i=n} (Y_1 - \bar{Y}) = 0$	$\sum_{i=1}^{i=n} (Y_1 - \bar{Y})^2 = 3000$

Het *foute* gevoel moet dus zijn: Alle gokfouten optellen en daarna delen door het aantal, want dan weet je de gemiddelde lengte van die gokfouten dus de standaardafwijking (zie het dus als negen blauwe streepjes waar je de gemiddelde lengte van berekent, bij mij in de les zijn deze streepjes *altijd* blauw).

figuur 0B



Probleem 1: Alle gokfouten optellen geeft nul, daarom gaan we eerst de gokfouten kwadrateren en daarna pas optellen. We kwadrateren hier om alle negatieve waarden positief te maken, zodat we ze wel kunnen optellen.

In formule-vorm zou de optelling of sommatie van de individuele afwijkingen er als volgt uitzien:

$$\sum_{i=1}^{i=n} (Y_i - \bar{Y})$$

Wat dus wil zeggen dat je *eerst* het hele gedeelte na het sommatieteken voor ieder aapje moet invullen en uitrekenen en daarna pas die uitkomsten per aapje moet optellen. (dus ook wel gewoon de optelling van de getallen in de derde kolom, zie tabel).

Uitgeschreven gaat ie als volgt:

$$\sum_{i=1}^{i=n}(Y_i-\bar{Y}) = [(120–150)+(130–150)+(140–150)+(140–150)+(150–150)+(160–150)+ (160–150)+(170–150) + (180–150)]$$

of als je dit intypet:

$$\sum_{i=1}^{i=n}(Y_i-\bar{Y}) = (120–150)+(130–150)+(140–150)+(140–150)+(150–150)+(160–150)+ (160–150)+(170–150)+(180–150)$$

of met tussen-antwoorden (de uitkomsten van de individuele afwijkingen):

$$\sum_{i=1}^{i=n}(Y_i-\bar{Y}) = –30 + –20 + –10 + –10 + 0 + 10 + 10 + 20 + 30$$

Ook hier even aandacht voor de minnetjes (en plussen). Eigenlijk kennen we twee soorten min-tekens. De gewone min op je rekenmachine is de min van 'aftrekken' (ik noem hem als nodig de 'aftrek-min' dus als je twee getallen van elkaar wil aftrekken). Die andere min op je rekenmachien is ook wel de min om aan te duiden dat een getal een negatieve waarde heeft (een minnetje tussen twee haakjes op je rekenmachien). Ik noem hem 'de min van negatief'.
En in de bovenstaande formule is de eerste min na het '='- teken (die dus voor de 30 staat) een min van negatief getal, de tweede (voor 20) ook. Eigenlijk zie je hierboven dus ook een optelling of sommatie van positieve en negatieve getallen. De meeste mensen zouden zeggen dat als je 8-6=2 doet, dat dat een aftrekking (aftreksom) is, maar ik zou het liever willen zien als de optelling van een positief getal (8) en een negatief getal (-6), dus 8+-6=2. Als je namelijk 8 euro in je portemonnee hebt én (plus) je hebt ook nog een schuld van 6 euro, zou je dus 2 euro overhouden. Misschien vind je dat ik moeilijk doe, maar later zul je me dankbaar zijn.... *Anyway*, bijna iedereen weet dat '+ - ' gewoon min wordt. Dus uiteindelijk ziet de som van alle individuele afwijkingen er als volgt uit.

$$\sum_{i=1}^{i=n}(Y_i-\bar{Y}) = –30 –20 –10 –10 + 0 + 10 + 10 + 20 + 30 = 0$$

En heeft dus de waarde nul, hier voelt het een beetje alsof de gokfouten 'verdwenen' zijn, maar bij optelling van gewone residuen (of gokfouten) krijg je dus *altijd* nul. Je kan ook zeggen dat de gokfouten elkaar opheffen omdat de uitkomst van de som dus nul is. Eigenlijk moeten we dus van alle negatieve waarden af, zodat we de gokfouten wel kunnen optellen. Wij gaan de gokfouten straks kwadrateren om van de minnetjes af te komen, maar eerst nog een uitwijding over *absolute waarden*.
We zeggen ook wel dat de waarden '30' en '-30' *absoluut* gezien even groot zijn, omdat ze (op de getallenlijn) dezelfde afstand tot nul hebben. Je moet *even ver* wandelen om vanuit 30 of -30 naar 0 te lopen, alleen de richting verschilt. We gebruiken absoluut-tekens om een getal om te zetten naar zijn absolute waarde (of aan te kondigen dat de absolute waarde eraan komt na het = teken). Absoluut-teken(s) doe je met twee verticale strepen om een waarde heen:

$$|–30| = 30$$

Je zegt dan: de absolute waarde van -30 is 30. Of:

$$|20| = 20$$

Het getal twintig is dus *al* gelijk aan zijn eigen absolute waarde.
Je zegt in dit laatste geval dus dat de absulote waarde van 20, dus gewoon 20 is (best flauw dus).

Dus *absoluut* gezien, is bij ons de kleinst mogelijke gokfout 0 cm en de grootste 30 cm.
Alle andere gokfouten hadden dus een absolute waarde ergens tussen de 0 en de 30.

Als je zou moeten schatten wat *ongeveer* de waarde is van de gemiddelde gokfout zou je dus ergens tussen de 0 en de 30 in moeten gaan zitten, zeg voor het moment dat die waarde bijvoorbeeld ongeveer 15 zal zijn. Hou dit gevoel, een gemiddelde gokfout van ongeveer 15 cm, even vast. Als een aapje binnen komt wandelen zeggen wij dat ie wel 150 cm zal zijn, maar het kan dus zijn dat:

- een aapje (maximaal) 30 cm boven de verwachting zit
- een aapje (maximaal) 30 cm onder de verwachting zit
- een aapje precies op de verwachting zit (er 0 cm vandaan zit)
- een aapje tussen de 0 en de 30 cm van de verwachting vandaan zit

maar gemiddeld zitten ze dus *ongeveer* 15 cm onder of boven de verwachting

De som van de gokfouten geeft dus nul en heeft geen zin, we moeten dus van die minnetjes af. Sommige formules voor de standaardafwijking gebruiken absoluut-strepen om van alle minnetjes af te komen, wij gebruiken een andere formule (of manier) en kwadrateren dus eerst de gokfouten voordat we ze optellen. In de statistiek zul je heel vaak waarden moeten kwadrateren (een getal keer zichzelf doen) (ja, tot vervelens toe) en daarna optellen, we hebben daar dus ook een naam voor: 'De Som van de Kwadraten' of kortweg de 'Kwadratensom' of '*Sum of Squares*'. Het is dus níet 'het kwadraat van de som', want dan zou je eerst optellen en dan pas kwadrateren.

een individuele gekwadrateerde afwijking = $(Y_i - \bar{Y})^2$

Voor aapje nummer 1 word het dus

$$(Y_i - \bar{Y})^2 = (120-150)^2 = (-30)^2 = -30 \cdot -30 = 900$$

Rekenregels

Ook hier weer even aandacht voor de rekenregels. Voor het gewone rekenwerk heb je een aantal handelingen (operaties) waarvan de volgorde dwingend is:

1. haakjes
2. machten (wortels en andere *machts*-wortels)
3. vermenigvuldigen (en delen)
4. optellen (en aftrekken)

1. Haakjes

Altijd eerst uitwerken - voor zover mogelijk - wat tussen haakjes staat, dus eerst een voorbeeld zonder haakjes en dan een paar met.

$2 \cdot 3 + 5 = 6 + 5 = 11$ De vermenigvuldiging moet dus eerst gebeuren en daarna de optelling

$(5+2) \cdot 3 = (5+2)3 = (7)3 = 7 \cdot 3 = 21$ Altijd eerst opschonen (herleiden of korter schrijven) wat tussen haakjes staat. Zodra je niet verder kan zoals bij '(7)', dan zijn de haakjes overbodig geworden.

$5 + (2 \cdot 3) = 5 + 6 = 11$ Hier zijn de haakjes dus overbodig omdat je sowieso eerst moet vermenigvuldigen.

2. Machten

Bij statistiek komen jullie eigenlijk vooral kwadraatjes tegen, maar soms ook hogere of lagere machten. Neem bijvoorbeeld 'twee tot de macht vier' of 'twee tot de vierde macht (verheven)' en 'je verheft twee, tot de macht (van) vier'. Je schrijft het als: 2^4, en het betekent ook wel: 2· 2· 2· 2 of ook wel' het getal 2, 3 keer met zichzelf vermenigvuldigd'. Ja, pas op, drie keer, want

als je een getal één keer met zichzelf vermenigvuldigd heb je het al tot de macht 2 gedaan (gekwadrateerd). In 2^4 is het 'grondtal' 2 en noem je 4 dus (de waarde van) de 'macht' of 'exponent'.

Wortels

Bij de 'gewone' wortel (huhuh..., maar heet ook wel 'tweedemachts'wortel), werkt het precies *omgekeerd* als bij 'een getal tot de tweede macht (verheffen)'.

Neem bijvoorbeeld 9^2 waarbij ik het getal 9 dus tot de tweede macht verhef. Je spreekt het uit als: 'negen tot de macht twee' of 'het kwadraat van negen'. 9 is hier het 'grondtal' en 2 is de macht (waarmee je 9 verheft).

handeling bij 9^2 : Hier doe je '9', één keer *zichzelf* dus: $9 \cdot 9 = 81$.

En nu bijvoordeeld juist de wortel van '9', dus $\sqrt{9}$. Je spreekt het uit als; 'de wortel van negen'. Om de wortel van negen te vinden is de vraag hierbij eigenlijk:

- *Wat* keer *wat* is 9?
- *Wat* of *welke waarde* zou één keer zichzelf, 9 zijn?
- welk getal voor X (*wat*), zou keer zichzelf als antwoord '9' geven?
- Bij de vraag wat is de waarde van $\sqrt{9}$ hoort ook wel:

$X^2 = 9$ Welk getal moet ik tot de macht twee verheffen om negen te krijgen of:

$X \cdot X = 9$ dus voor *welke waarde* van X klopt deze vergelijking?

Of: Het antwoord is natuurlijk '3', want $3^2 = 3 \cdot 3 = 9$. Trouwens, de wortel van 9 is óók '-3' omdat $-3 \cdot -3 = 9$. Dit laatste mag je *voorlopig* vergeten. Sterker nog: Er bestaan ook hogere (of lagere) machtwortels dan de tweedemachts (gewone) wortel. Voorlopig hebben we die niet nodig en daar ga ik - gelukkig voor jullie - nu ook niet op in. Al met al voor ons:

$\sqrt{9} = 3$ Dit is ook het enige antwoord dat je rekenmachientje geeft en niet de negatieve waarde dus.

Een paar wortels, dus kijk even of je rekenmachine doet wat ie moet doen.

TABEL 0D

wortel-behandeling	exact resultaat of antwoord	reden (bewijs)	resultaat afgerond in drie decimalen
$\sqrt{1}$	1	$1 \cdot 1 = 1$	1.000
$\sqrt{2}$	$\sqrt{2}$	$\sqrt{2} \cdot \sqrt{2} = 2$	1.414
$\sqrt{3}$	$\sqrt{3}$	$\sqrt{3} \cdot \sqrt{3} = 3$	1.732
$\sqrt{4}$	2	$2 \cdot 2 = 4$	2.000
$\sqrt{9}$	3	$3 \cdot 3 = 9$	3.000
$\sqrt{10}$	$\sqrt{10}$	$\sqrt{10} \cdot \sqrt{10} = 10$	3.162

3.Vermenigvuldigen en delen

Ik hou het hier vooral even bij het moeilijke 'taaltje' dat je moet snappen en moet kunnen vertalen naar een vermenigvuldiging of deling. Hoevaak zie ik mensen niet denken, na een vraagstuk: 'Uhm, moet ik nou juist delen of keer doen?'. Beide operaties lijken weer erg op elkaar, ze zijn alleen verschillend omdat ze het omgekeerde van elkaar zijn. Beetje vaag vooralsnog, maar je kan 'iets' of een hoeveelheid, groter of kleiner maken, meer of minder. Stel, je hebt heel veel dingen van hetzelfde, bijvoorbeeld heel veel briefjes van 10 euro, zeg 30 stuks. Natuurlijk ben je geïnteresseerd in het totale bedrag. Maar wat is de snelste manier? Natuurlijk niet optellen. Omdat elk briefje dezelfde waarde heeft maken we 'de waarde ook wel

dertig keer zo belangrijk of zo groot'. Je pakt je rekenmachientje (je hoeft van mij niet te kunnen hoofdrekenen, zelfs dit niet) en tikt het in. Maar eigenlijk maak je het getal 10 met een *factor* 30 groter, het getal vermenigvuldig je dus met 30 (de factor) en natuurlijk geeft dat 300. of je nou 10 keer 30 doet of dat je het omdraaid: 30 keer 10, het geeft allebei hetzelfde antwoord. Wat algemener:

$$a \cdot b = b \cdot a$$

Bij '12 gedeeld door 4 is 3', als je dus aan het delen bent (door 4) maak je een bepaalde hoeveelheid (hier 12) kleiner, ook wel *zoveel keer* (4) kleiner als *waar* (4) je die hoeveelheid door deelt. Dus als je weet dat je iets 4 keer kleiner moet maken, dan moet je dus door 4 delen (of vermenigvuldigen met $\frac{1}{4}$).

4. Optellen en aftrekken

Denk er vooral aan, als je optelt, dat je een positieve waarde toevoegt aan (optelt bij) een willekeurige andere waarde, dat je dus ook wel naar rechts moet wandelen (als je aan de getallenlijn denkt). Als je een negatieve waarde toegevoegt aan een willekeurig andere waarde (dus aftrekt), schuif je op naar links. En nog een hersenkraker: als je een negatieve waarde van een (andere) willekeurige waarde *aftrekt*, dan schuif je dus naar rechts op (min min wordt plus). Toch nog even wat voorbeelden voor als je toch nog in de war raakt:

$$3+5 = 8$$
$$3+ -5 = 3-5 = -2$$
$$3- -5 = 3+5 = 2$$
$$-3+5 = 2$$
$$-3+ -5 = -3-5 = -8$$
$$-3-5 = -8$$
$$-3- -5 = -3+5 = 2$$

We lopen de berekening $(120\text{-}150)^2$ aan de hand van de operaties nog een keer door. Lekker moeilijk doen over makkelijke dingen.

- staan er haakjes in?
Ja en daarom moeten we eerst kijken wat er tussen de haakjes staat en dat *zover mogelijk* oplossen. Tussen de haakjes staan *geen* haakjes, machten of vermenigvuldigingen, er staat alleen maar een aftrekking, dus die kan je meteen doen. Tussen de haakjes staat '120-150', ook wel een (aftrek) som van twee termen (120 en -150). We noemen deze twee termen 'gelijksoortig'. In een sommetje zoals '4+2a' zijn de twee termen (4 en 2a) niet gelijksoortig en kan je het sommetje dus ook niet verder uitwerken.

$$(120\text{-}150)^2 = (-30)^2$$

We kunnen nu dus zeggen dat wat er tussen haakjes staat echt één waarde is (geworden).

- staan er machten in $(-30)^2$?
Ja, de tweede macht, het kwadraat (een kwadraat is een macht) komen we tegen en het kwadraat slaat hier op alles wat tussen haakjes staat, '-30' dus. Een kwadraat betekent ook wel dat je de waarde (-30) die je 'kwadrateert', keer zichzelf moet doen.

$$(-30)^2 = -30 \cdot -30$$

Alledrie de minnen hier zijn van 'negatief'. Intypen als -30*-30 en je uitkomst is dan 900, min keer min is altijd plus. Vaak typen mensen het toch fout in en typen ze letterlijk -30^2 in en dat

geeft toch écht een ander antwoord:

$$-30^2 = -30 \cdot 30 = -900$$

Je ziet hier dat het kwadraatje dus - blijkbaar - alleen maar op die 30 slaat en dus geen betrekking op het minnetje heeft. Dus als je weet dat je een negatieve waarde moet kwadrateren bijvoorbeeld '-6', dan zijn er maar twee correcte manieren:

$(-6)^2 = 36$ of $6^2 = 36$ en bij de laatste laat je min-teken dus gewoon weg.

Blauwe streepjes en blauwe vierkantjes

Wat gebeurt er eigenlijk als je '6 cm' kwadrateert? Als je de oppervlakte van een vierkant wil berekenen hoef je alleen maar de breedte maal de lengte te doen. En aangezien een vierkant vier gelijke zijdes heeft, kun je dus ook de lengte van één zijde kwadrateren!

$6^2 = 6 \cdot 6 = 36$ Maar officieel zou je ook de meeteenheid in je bereking moeten zitten.

$$(6\ \text{cm})^2 = (6 \cdot \text{cm})^2 = 6\ \text{cm} \cdot 6\ \text{cm} = 6 \cdot 6 \cdot \text{cm} \cdot \text{cm} = 36\ \text{cm}^2$$

Het resultaat is dus 36 centimeter kwadraat of ook wel 36 vierkante centimeter.

Verder met de echte kwadratensom, de som van de gekwadrateerde afwijkingen.

$$\sum_{i=1}^{i=n} (Y_i - \bar{Y})^2 = (120-150)^2 + (130-150)^2 + (140-150)^2 + (140-150)^2 + (150-150)^2 + (160-150)^2 + (160-150)^2 + (170-150)^2 + (180-150)^2$$

In de vierde kolom van de tabel vind je de gekwadrateerde residuen, natuurlijk allemaal positief (daar ging het juist om), maar dus wel een stuk groter geworden. De kleinste gekwadrateerde gokfout is 0 en de grootste heeft een waarde van 900. Je mag dus ook denken in termen van blauwe gekwadrateerde streepjes (de gekwadrateerde afstand van een observatie naar het gemiddelde). Blauwe vierkantjes dus (waarvan de lengte van hun zijdes - blauwe streepjes - gelijk is aan de wortel van hun oppervlakte)! Als je een grove *schatting* zou moeten geven van de waarde van 'de gemiddelde gekwadrateerde afstand van een observatie naar het gemiddelde', wat zou je dan kunnen zeggen? Als het kleinste kwadraat 0 is en het grootste 900, zou ik er tussenin gaan zitten, zeg 450. Dus de gemiddelde gekwadrateerde gokfout heeft ongeveer een waarde van 450. Qua berekening zou je zeggen als je de gemiddelde gekwadrateerde afwijking wil bereken, moet je ze eerst optellen en daarna delen door het *aantal*, bij ons 9 dus. Maar ook hier gaan we wat anders doen:

Probleem 2: We delen de kwadratensom niet door 'n', maar door 'n-1', ook wel het aantal vrijheidsgraden of *degrees of freedom* genoemd.

Het draait allemaal om gokken, maar neem even het volgende voorbeeld. Stel, ik heb drie portemonnees met daarin wat geld en ik vertel je dat er gemiddeld - per portemonnee - 10 euro in zit. Dat betekent dus dat als ik een portemonnee open, jij zou zeggen dat er precies 10 euro in zit omdat dat blijkbaar de verwachte waarde is. Als ik nu de eerste open en er blijkt 7 euro in te zitten, heb jij dus een gokfout gemaakt van $X_i - \bar{X} = X_1 - \bar{X} = 7\text{-}10 = -3$, ik gebruik hier even x-en (voor de lol). Als ik ook de tweede portemonne open maak, zeg jij natuurlijk weer 10 euro en hier zat bijvoorbeeld 11 euro in. Nu heb je dus informatie over de eerste twee portemonnees, wat zou je nu - gegeven deze informatie - voor de derde portemonnee zeggen? Omdat je weet dat het gemiddelde 10 is ($\bar{X} = 10$, gegeven) en er 3 observaties zijn ($n=3$, gegeven) weet je ook (zou je moeten kunnen beredeneren):

- dat de optelling van de drie scores 30 zou moeten zijn of (som van X_i)
- dat de optelling van alle residuen nul zou moeten zijn
- dat in de derde portemonnee dus 12 euro moet zitten

Want als je kijkt naar de formule voor het gemiddelde en daar alles invult dat er gegeven is kan je de waarde van X_3 dus uitrekenen.

$$\bar{X} = \frac{\sum_{i=1}^{i=n} X_i}{n} = \frac{X_1 + X_2 + X_3}{n} = \frac{\text{som van } X_i}{n}$$

invullen wat je weet geeft:

$$10 = \frac{\text{som van } X_i}{3} \qquad \text{of als aan de residuen denkt:}$$

$$\sum_{i=1}^{i=n} (X_i - \bar{X}) = (X_1 - \bar{X}) + (X_2 - \bar{X}) + (X_3 - \bar{X}) = 0$$

Om deze vergelijking op te lossen, kun je gebruik maken van het volgende: De vergelijking staat ook wel in van de vorm $3 = \frac{6}{2}$ (ik kies hier dus even makkelijke getallen). We zijn op zoek naar de waarde van de *som van* X_i zodanig dat de vergelijking klopt. Wat moet je met 2 en 3 doen om 6 te krijgen? Met elkaar vermenigvuldigen, want $6 = 2 \cdot 3$. Omdat onze vergelijking dezelfde vorm heeft kunnen we dus hetzelfde doen:

$$\textit{som van } X_i = 3 \cdot 10 = 30$$

Als je dus al weet dat de som 30 moet zijn (omdat het gemiddelde ook al bekend was) én je weet dat X_1=7 en X_2=11, dan moeten X_3 wel een waarde zijn van 12, want 7+11+12=30. Maar als je dus het gemiddelde weet van een verzameling (set) getallen dan weet je dus ook wat de optelling van die getallen moet zijn en je kunt dus altijd de laatste waarneming (bij ons net X_3) dus zelf uitrekenen als de rest (n-1) van de waarnemingen, gegeven zijn.

Laten we de andere vergelijking ook maar meteen uitwerken

$$\sum_{i=1}^{i=n} (X_i - \bar{X}) = (X_1 - \bar{X}) + (X_2 - \bar{X}) + (X_3 - \bar{X}) = 0$$

$$(7 - 10) + (11 - 10) + (X_3 - 10) = 0$$

Omdat in deze vergelijking er geen machten of vermenigvuldigingen gebruikt worden, staan de haakjes er hier voor Jan Joker, ze kunnen dus weg.

$$7 - 10 + 11 - 10 + X_3 - 10 = 0$$

Bij elkaar nemen wat gelijksoortig is;

$$7 + 11 - 10 - 10 - 10 + X_3 = 0$$

$$18 - 30 + X_3 = 0$$

$$-12 + X_3 = 0$$

En sorry, heel even zoals in de brugklas, de 'balancemethode':
Omdat dit een *vergelijking* is en dat is gewoon een stelling. Die stelling luidt als volgt: 'Het linker deel is gelijk aan het rechter deel.' Of: Het deel links van het '='-teken - dus '$-12+X_3$'- is gelijk (van waarde) aan het rechter deel, dus '0'. De vraag is hier dus eigenlijk: 'Voor welke waarde

van X_3 klopt deze stelling?'. Om deze vraag op te lossen, kun je aan *beide* kanten een *éénzelfde* hoeveelheid erbij gooien (dus optellen). Ik ga aan beide kanten er 12 bij knallen;

$$-12+X_3+12 = 0 + 12$$

en weer de gelijksoortige termen bij elkaar rapen en de boel opruimen.

$$-12+12+X_3 = 12$$

$$0+X_3 = 12$$

$$X_3 = 12$$

En wat roep je dan als antwoord? Waarschijnlijk roep je nu iets als 'IKS-drie is 12' en daarop zeg ik (dolgelukkig): 'FOUT!'... en geef ik je stralend het goede antwoord: ' In de derde portemonnee zit 12 *EURO*!', want de *harde* - en dus tastbare - realiteit gaat niet over X-en of wiskundig geleuter, maar gewoon over appels en peren, dus laten we *die* vooral benoemen. Zou even mooi zijn: Sta je bij de bakker en je vraagt hem om brood, maar je krijgt een briefje met een brood-*recept* in je hand gedrukt. Nu ik toch uitweid, mocht je zover als hier gekomen zijn en min of meer begrepen hebben, zou ik me geen zorgen maken over de rest - met aandacht - wordt het een makkie.

Almost wrapping things up,

Samengenomen hebben we nu dus ontdekt dat als je een bekende set (verzameling) getallen probeert 'terug' te voorspellen, dat je het laatste getal of waarde dus niet hoeft te gokken, maar gewoon kan uitrekenen. Dus als je onze aapjes één voor één laat binnenlopen op willekeurige volgorde, moet je dus de eerste 8 (n-1) aapjes gokken, maar als je tussendoor netjes je acht observaties opschrijft (onthoud) kun je dus het laatste aapje, de negende, netjes uitrekenen (zijn lengte dan). Het 'laatste' aapje zit dus eigenlijk altijd 'vast' qua waarde, maar de eerste acht hebben dus alle 'vrijheid'. En daarom dus 'vrijheidsgraden'. We zeggen ook wel: 'Een set van 9 waarnemingen heeft acht vrijheidsgraden (en één waarneming zit dus vast (gegeven een bepaald gemiddelde)). Simpel gezegd: als je negen getallen hebt, hoef je er maar acht te gokken, omdat je het laatste getal dus kunt uitrekenen. Of nog korter; 'Some things are redundant to say...(*duh*)' . Maar goed, een set van n getallen heeft dus n-1 vrijheidsgraden en het vertelt ons in termen van gokfouten, dat je bij de aapjes dus maar acht gokfouten hebt en niet negen! De kwadratensom wordt dus gedeeld door het aantal vrijheidsgraden (8), omdat we maar acht gokfouten hebben (die laatste kon je uitrekenen). En tenslotte was toch de *gemiddelde gekwadrateerde gokfout* het doel? Ja, dus punt. Afgezien dat ik nog steeds een 'dergrees of freedom'-party wil geven, moet je er tijdens berekeningen wel heel vaak rekening mee houden. Die vrijheidsgraadjes komen bovendien in een grote verschijdenheid voor. Dus genieten.

Terug naar de uitwerking

$$\sum_{i=1}^{i=n} (Y_i - \bar{Y})^2 = (120-150)^2 + (130-150)^2 + (140-150)^2 + (140-150)^2 + (150-150)^2 + (160-150)^2 + (160-150)^2 + (170-150)^2 + (180-150)^2$$

$$\sum_{i=1}^{i=n} (Y_i - \bar{Y})^2 = (-30)^2 + (-20)^2 + (-10)^2 + (-10)^2 + (-0)^2 + (10)^2 + (10)^2 + (20)^2 + (30)^2$$

$$\sum_{i=1}^{i=n} (Y_i - \bar{Y})^2 = 900 + 400 + 100 + 100 + 0 + 100 + 100 + 400 + 900$$

$$\sum_{i=1}^{i=n} (Y_i - \bar{Y})^2 = 3000$$

Nu de kwadratensom delen door het aantal vrijheidsgraden, *n*-1.

$$\frac{\sum_{i=1}^{i=n}(Y_i-\bar{Y})^2}{n-1}=\frac{3000}{8}=375$$

of (liever):

$$\frac{1}{n-1}\cdot\sum_{i=1}^{i=n}(Y_i-\bar{Y})^2=\frac{1}{9-1}\cdot 3000=\frac{1}{8}\cdot 3000=1/8*3000=375$$

Nu vast moeilijk doen is, goed voor later, maar we hebben dus een antwoord: de gemiddelde kwadrateerde gokfout heeft een waarde van 375 (officiëel: 375 cm², met éénheden, maar vergeet dit echt alsjeblieft). Weet je nog, wij hadden 450 *geschat*, prima dus, niet helemaal dezelfde waarde, maar wij waren wel heel grof.

Maar wie wilde er nou een *gekadrateerde* afwijking? Niemand, mag ik hopen, dus we moeten nog de wortel nemen om het laatste probleem op te lossen, bijna ademhalen dus. Wij hadden de gokfouten gekwadrateerd (vierkant gemaakt) en daarvan de gemiddelde waarde berekend. We moeten dus nog de wortel trekken om eindelijk klaar te zijn.

Trek de wortel van de gemiddelde kwadraten;

$$\sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{i=n}(Y_i-\bar{Y})^2}=\sqrt{\frac{1}{9-1}\cdot 3000}=\sqrt{\frac{1}{8}\cdot 3000}=\sqrt{1/8*3000}=\sqrt{375}\approx 19.36$$

De (gewone) gemiddelde gokfout - de standaardafwijking - heeft dus een waarde van 19.36 en ik kies hier even voor een afronding op twee decimalen. Geeft me meteen een reden om over afrondingen te praten. 19.36 is *slechts* een afronding, eigenlijk komen nog heel veel cijfers achter het cijfertje '6', sterker nog; het zou zo maar kunnen dat het echte *aantal* cijfers achter de komma (bij ons een punt) oneindig groot is. *Exact* gezien heeft de standaardafwijking van y een waarde van $\sqrt{375}$, dit zou dus een exact antwoord zijn en in onze eind-antwoorden geven wij altijd een *benadering* in meestal 2 maar soms ook 3 decimalen (2 of 3 cijfers achter de komma), we moeten dus de lange getallen afronden, hoe gaat dat ook al weer? Hier even een paar voorbeelden.

Afrondingen

TABEL 0E

exacte waarden	afronding in vijf decimalen	afronding in drie decimalen	afronding in twee decimalen
$\sqrt{(375)}$	19.36492	19.365	19.36
7.1234567890	7.12346	7.123	7.12
9.989898989	9.98990	9.990	9.99
99.99999999999	10.00000	10.000	10.00
π (het getal of constante pi)	3.14159	3.142	3.14
e (het getal of constante 'e', komen we later nog wel een keer tegen)	2.71828	2.718	2.72

De regel hierbij is dat als je bijvoorbeeld op drie decimalen moet afronden, kijk je altijd *alleen* één cijfer verder (dus het vierde cijfer na de komma (punt) in het exacte getal) en als dat cijfer een waarde heeft van 4 of lager, dan blijft het derde cijfer gelijk. Maar als het vierde cijfer 5 of hoger is, dan wordt het derde cijfer één punt hoger. In die speciale gevallen (in de tabel het vierde exacte getal) waar het derde cijfer een 9 is (en het vierde cijfer 5 of hoger), zal het derde cijfer dus eigenlijk 10 moeten worden, maar dat gaat niet zomaar en zal het tweede cijfer ook mee moeten veranderen (ook ééntje hoger), maar als het tweede cijfer ook een 9 is.... Over het algemeen zal je niet zakken op een verkeerde afronding tijdens je tentamens, dus maak je voorlopig niet te veel zorgen hierover, al gaande weg wordt het makkelijker. Sommige getallen zijn zo bijzonder dat we ze een naam of symbool hebben gegeven, zoals bij pi, omdat dit eigenlijk een te *lang* getal is (oneindig veel cijfers achter de komma, waarschijnlijk) en we het niet altijd willen

afronden schijven we het dus als een symbool (de Griekse letter π). En hetzelfde geldt dus ook voor het getal e, maar later hier meer over.

Finally. Het belangrijkste, de interpretatie.

Dus de gemiddelde lengte van een blauw streepje is dus 19.36 cm, of netter; de gemiddelde afwijking van een observatie naar het gemiddelde heeft dus een waarde van 19.36 cm, de standaardafwijking. Nu weten we dus eindelijk wat de waarde van de gemiddelde gokfout is, of qua gevoel nog beter; we weten nu *wat* we moeten gokken (het gemiddelde van 150 cm) en *hoe goed* we kunnen gokken (de standaardafwijking van 19.36 cm). Als we dus zeggen of voorspellen dat een aapje 150 cm zal zijn, zitten hun lengtes gemiddeld 19.36 cm van onze verwachting vandaan (erboven of eronder). Als laatste nog even de juiste namen en symbolen bij de formules.

De gemiddelde gekwadrateerde afwijking voor de Y-scores wordt de *variantie* van *Y* genoemd een heeft als symbool: S_y^2

$$\text{Variantie van } Y = S_y^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{i=n} (Y_i - \bar{Y})^2$$

$$S_y^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{i=n} (Y_i - \bar{Y})^2 = \frac{1}{8} \cdot 3000 = 1/8 * 3000 = 375$$

$$S_y^2 = 375$$

Om de waarde te vinden van de standaardafwijking van Y, neem je dus de wortel van de variantie:

$$\text{standaardafwijking van } Y = S_y = \sqrt{S_y^2} = \sqrt{375} \approx 19.36$$

Minimale Oefening

Het gemiddelde, variantie en standaardafwijking voor de variabele X_i berekenen we natuurlijk op dezelfde manier, maar vervangen we de y-tjes door de x-jes in de formules.

$$\bar{X} = \frac{1}{n-1} \cdot \sum_{i=1}^{i=n} X_i$$

$$\bar{X} = \frac{1}{9} \cdot [1+1+1+1.5+1.5+1.5+2+2+2]$$

$$\bar{X} = \frac{1}{9} \cdot (1+1+1+1.5+1.5+1.5+2+2+2) = \frac{1}{9} \cdot (13.5) = 1/9 * 13.5 = 1.5$$

of even voor de 'kicksaus' ook op een andere manier:

$$\bar{X} = \frac{1}{9} \cdot (3 \cdot 1 + 3 \cdot 1.5 + 3 \cdot 2)$$

Omdat bij ons de waarde 1, 1.5 en 2 allemaal drie keer voorkomen, heb ik die waarden met 3 vermenigvuldigd. Tussen de haakjes staat nu de optelling van *drie* termen: 3 keer 1, 3 keer 1.5 en 3 keer 2, zijn alledrie als normale getalletjes te schrijven en zijn dus gelijksoortig. Omdat hier de waarde 1, 1.5 en 2, alledrie met 3 worden vermenigvuldigd, mag je die drie ook buiten haakjes halen:

$$\bar{X} = \frac{1}{9} \cdot 3 \cdot (1 + 1.5 + 2)$$ de drie termen die nu tussen de haakjes staan zijn nog steeds gelijksoortig, dus opschonen:

$$\bar{X} = \frac{1}{9} \cdot 3 \cdot (4.5)$$ de haakjes rond 4.5 staan er nu weer voor joker en je kan ze dus weghalen;

$$\bar{X} = \frac{1}{9} \cdot 3 \cdot 4.5 = 1/9 * 13.5 = 1.5$$

En nu de variantie:

$$S_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{i=n} (X_i - \bar{X})^2$$

$$S_x^2 = \frac{1}{9-1} \cdot [(1-1.5)^2 + (1-1.5)^2 + (1-1.5)^2 + (1.5-1.5)^2 + (1.5-1.5)^2 + (1.5-1.5)^2 + (2-1.5)^2 + (2-1.5)^2 + (2-1.5)^2]$$

Ik ga meteen wat rekenregels toepassen om wat handiger te kunnen rekenen (je zou maar 1000 observaties hebben...)

$S_x^2 = \frac{1}{8} \cdot [3 \cdot (1-1.5)^2 + 3 \cdot (1.5-1.5)^2 + 3 \cdot (2-1.5)^2]$ en ik vervang de blokhaken:

$S_x^2 = \frac{1}{8} \cdot (3 \cdot (1-1.5)^2 + 3 \cdot (1.5-1.5)^2 + 3 \cdot (2-1.5)^2)$

dus weer tussen de blokhaken drie gelijksoortige termen met allemaal een 3 erin, dus die 3 kunnen we buitenhaakjes halen.

$S_x^2 = \frac{1}{8} \cdot 3 \cdot ((1-1.5)^2 + (1.5-1.5)^2 + (2-1.5)^2)$ wat tussen haakjes staat schoon ik op:

$S_x^2 = \frac{1}{8} \cdot 3 \cdot ((-0.5)^2 + (0)^2 + (0.5)^2)$ dan de kwadraatjes wegwerken

$S_x^2 = \frac{1}{8} \cdot 3 \cdot (0.25 + 0 + 0.25)$ dan weer opschonen:

$S_x^2 = \frac{1}{8} \cdot 3 \cdot (0.5)$ haakjes nu voor jan joker:

$S_x^2 = \frac{1}{8} \cdot 3 \cdot 0.5 = 1/8 * 3 * 0.5 = 0.1875$ of

$S_x^2 = \frac{1}{8} \cdot 3 \cdot 0.5 = \frac{1}{8} \cdot \frac{3}{1} \cdot 0.5 = \frac{3}{8} \cdot 0.5 = 3/8 * 0.5 = 0.1875$

3 is te schrijven als $\frac{3}{1}$ (drie eende of drie eerste, een breuk dus) en je kunt breuken met elkaar vermenigvuldigen door teller maal teller en noemer maal noemer te doen, dus respectievelijk, 1 keer 3 en 8 keer 1.

nu de standaardafwijking:

$S_x = \sqrt{S_x^2} = \sqrt{.1875}$ dit zou je exacte antwoord voor de standaardafwijking zijn.

$S_x = \sqrt{S_x^2} = \sqrt{.1875} = 0.4330$ en je antwoord afgerond in vier decimalen.

Zo, nu hebben we toch echt wel die aapjes behoorlijk uitgemolken en zijn we de nodige rekenregels tegenkomen.

Meetniveaus van variabelen

Een variabele kan dus verschillende waarden aannemen. Verschillende waarden of categorieën (dit is hetzelfde) kennen we toe om aan te tonen dat dingen nou eenmaal verschillen. Sommige studenten zijn nou eenmaal universitaire studenten en anderen zijn HBO studenten. Het meetniveau van een variabele vertelt ons wat we met die mogelijke waarden (of categorieën) kunnen doen, dus hoe je die waarden kunt gebruiken. Dus het meetniveau van een variabele zegt iets over de aard van die waarden of hoe die verschillende waarden (van een variabele) van elkaar verschillen. Met de twee mogelijke waarden *man* en *vrouw* van de variabele geslacht kun je beduidend minder doen (statistische toepassing, berekening) dan met de waarden die (altijd) in getallen worden uitgedrukt, zoals waarden bij de variabele lengte in cm. Je zou kunnen zeggen dat de waarden 'man' of 'vrouw' minder kwaliteit(en) hebben dan de waarden 172 cm of 151 cm omdat de laatste twee waarden meer informatie in zich dragen over *hoe* die twee waarden van elkaar verschillen. Het meetniveau van een variabele heeft zes verschillende niveaus die aangeven hoe de waarden of categorieën in complexiteit van elkaar verschillen. De mogelijke meetniveaus (op volgorde van complexiteit) zijn *nominaal, ordinaal, interval, ratio* en *absoluut*.
Eigenlijk kan je dus zeggen dat het 'meetniveau' (van een variabele zoals geslacht of lengte in cm) een variabele is die iets zegt over de mogelijke waarde van een andere variabele (zoals geslacht of lengte in cm).
Het beoordelen van het juiste meetniveau van een variabele of schaal – of praktische toekenning aan - is uitermate belangrijk, de onderzoeksvraag en meetniveaus van de variabelen in je onderzoek bepalen geheel welke statistische analyse voor jou vraagstuk of onderzoek van toepassing is.

Nominaal

Wanneer een variabele een nominaal meetniveau heeft, zeggen we dat de mogelijke waarden die die variabele aan kan nemen slechts een *benaming* is van of voor het gemetene (het object) en dus *slechts* een verschil (tussen de mogelijke waarden of categorieën) aangeeft.
Dit is het meest basale (dus van lage complexiteit) meetniveau dus met de minste kwaliteit of mogelijkheden. Dus als twee objecten verschillen van naam (waarde), weten we *alleen* dat die twee objecten *niet hetzelfde* zijn (qua meting dus). Een nominaal meetniveau draagt alleen maar de kwaliteit 'verschil' in zich (dus best flauw eigenlijk). Je kiest een andere naam (voor een object) om aan te duiden dat ie anders is dan anders genoemde objecten.

Voorbeelden.
Voorbeelden van nominale variabelen: Naam, Geslacht, Nationaliteit, diersoort, provincie, land, aapsoort, of al dan niet slagen.
Voorbeelden van nominale waarden: Benjamin, man, Belgische, aap, Zuid-Holland, België, bonobo, gezakt.

Statistische toepassing of mogelijke berekeningen.
- Eigenlijk kun je alleen maar (de verschillende) categorieën tellen en eventueel daarna rapporteren hoe vaak (dus de frequentie van) een bepaalde categorie (procenteel of relatief) voorkomt (in een steekproef).

- Tellen en rapporteren (output) met spss: Analyse-descriptives-frequencies, variabele toevoegen en frequency table aangevinkt laten, dan ok of paste.

- Vaak coderen we categorieën als getallen, we kunnen bijvoorbeeld afspreken dat de waarde 'vrouw' wordt gecodeerd als '0' en 'man' als de waarde '1'. Deze keuze is dus geheel willekeurig (arbitrair) en doet er dus eigenlijk niet toe omdat de betekenis van die getallen dus puur nominaal bedoeld is. Een andere keuze tot codering zou bijvoorbeeld '78' en '132' kunnen zijn, maar waarom zou je moeilijke getallen kiezen als het ook simpel kan? Dus kies je (her) coderingen zo handig mogelijk.

Ordinaal

Bij variabelen van een ordinaal meetniveau hebben de waarden een *extra* kwaliteit bovenop de nominale kwaliteit (verschil). Deze extra kwaliteit is een natuurlijke *ordening* van de categorieën die dus dwingend (of logisch) is. De twee kwaliteiten bij een ordinaal meetniveau zijn dus 'verschil' en 'een dwingende *volg-* of *rangorde'*. Ordinale waarden staan dus altijd op dezelfde volgorde qua grootte of hoeveelheid, maar het blijft bij ordinaal wel onduidelijk hoeveel verschil, (qua grote, hoeveelheid of afstand) *tussen* de verschillende waarden ligt. Aan de benamingen (waarden) 'goed', 'beter' en 'best' hoor je wel dat er *verschil* (eerste kwaliteit) en *rangorde* (tweede kwaliteit) in zit, maar je weet nooit *hoeveel* (derde kwaliteit) meer de categorie 'best' is dan 'beter'.

Voorbeelden.
Ordinale variabelen: opleidingsniveau, een vijf (of zeven) puntschaal met als eerste antwoordmogelijkheid 'helemaal mee oneens' tot en met 'helemaal mee eens' als laatste antwoordmogelijkheid (vaak gecodeerd als 1 t/m 5, ook wel een *likert*-schaal genoemd), leeftijdscategorie (voor mensen), rangen in het leger, Beoordelingssysteem in Nederland met (afgeronde) cijfers van 1 t/m 10.
Ordinale waarden: HAVO, een beetje mee oneens (2), jonge mensen, maarschalk, een 8.

Toepassing.
- Havo is een hoger opleidingsniveau dan VMBO (MAVO), maar het is onduidelijk hoeveel moeilijker, hoeveel meer werkende hersencellen, of motivatie - of wat dan ook - je zou moeten hebben om HAVO te kunnen halen dan VMBO. We weten al helemaal niet of die 'afstand' *tussen* twee naast elkaar liggende waarden zoals VMBO en HAVO vergelijkbaar is met de afstand met het volgende *sprongetje* van HAVO naar VWO. We weten slechts dat het één hoger of meer is dan het andere.

Statistische toepassing, gebruik of mogelijke berekeningen.
- strikt genomen zou je bij ordinaal, net zoals bij nominaal, alleen maar kunnen tellen en bijvoorbeeld aan de hand van een frequentieverdeling kunnen rapporteren (een tabel of grafiek die aangeeft hoevaak de mogelijke categorieën voorkomen in een steekproef, in aantallen of procentueel natuurlijk). Het zou raar zijn als je de genoemde Opleidingscategorieën als 1, 2 en 3 codeert en dan vervolgens rapporteert dat het gemiddelde opleidingsniveau 2.46 was in je steekproef, het zegt misschien wel wat (namelijk dat er dus relatief veel VWO-ers in je steekproef zaten) maar netjes is het niet.

- Vaak wordt bij (strikt) ordinale schalen gesmokkeld en wordt er (onterecht) een derde kwaliteit aan ordinale schalen verleend. Mensen gaan dan de getallen (coderingen) eigenlijk te serieus nemen. Dit gebeurd vaker als er wat meer opties (waarden) zijn bij een variabele. Zoals bij *likert*-scales (bijv een vijf (of zeven) puntschaal met als eerste antwoordmogelijkheid 'helemaal mee oneens' tot en met 'helemaal mee eens' als laatste antwoordmogelijkheid (gecodeerd als 1 t/m 5), vaak gebruikt bij opiniepijlingen of vragenlijst bijvoorbeeld voor een depressie(stoornis). Vaak wordt er dan gerapporteerd dat men gemiddeld 3.8 antwoordde (scoorde) op een 5 puntschaal op de vraag 'In hoeverre bent u het met kernenergie eens?' Oplopend van oneens tot en met eens is, strikt genomen, een ordinale schaal, maar het wordt behandeld *alsof* de schaal een intervalniveau behelst. In zo'n geval wordt dus onterecht de derde kwaliteit - hoeveelheid, afstand, of grootte *tussen* waarden - aan de waarden toegekend. In de praktijk zal dit gebruik niet veel schade aanrichten, maar voorzichtigheid in interpretatie van de scores is dus geboden.

- Ook het Nederlands cijfersysteem wordt vaak onterecht overgekwalificeerd. Strikt genomen kan je alleen maar zeggen dat iemand met een '9', *hoger* scoort dan iemand met een '8.5'. Dit noem je dus een ordinale uitspraak, omdat je alleen de kwaliteit 'volgorde' toekent (mag toekennen) aan het verschil tussen een 8.5 en een 9 en de afstand (verschil van '.5 cijfer') tussen

de twee getallen dus negeert. En laten we wel wezen: '.5' kan *onmogelijk* een 'halve fout' meer of minder betekenen en al helemaal niet omdat de categorie 'fout' sléchts nominaal van karakter is en dus niet numeriek (getalsmatig) uit te drukken valt en dus ook zeker niet te halveren! Maar zoals geen ander dan de student zelf, weet de student vaak beter hoe hij zijn jaargemiddelde moet berekenen dan zijn eigen docent.

- Er wordt - eigenlijk dus onterecht - met ordinale variabelen of schalen vaak *gerekend* alsof ze van intervalniveau zouden zijn. In dit geval wordt dus een hoger meetniveau (interval in dit geval) toegekend of aangenomen voor een variabele dan - strikt gezien - terecht is. Zeker binnen de sociale studies komt dit vaak voor. Maar voor ons wetenschappers is deze aanname dat een variabele zich 'interval' gedraagt een belangrijke én noodzakelijke aanname om überhaupt verder te kunnen rekenen en dus ook verdere conclusies te kunnen trekken (en zo bijv. kunnen publiceren aan de hand van een artikel).

Interval

In geval van een intervalvariabele (of -schaal) is er ook sprake van de derde kwaliteit waarop waarden zich van elkaar kunnen onderscheiden, namelijk 'een betekenisvol *verschil* in hoeveelheid, afstand of grootte *tussen* waarden'. Dit verschil kan en word nu uitgedrukt in numerieke waarden (echte getallen zoals 130) én eenheden (bijv IQ-punten). We weten nu als extra dus ook *hoeveel* verschil (in eenheden) er *tussen* twee gemeten waarden is (of ligt) en dus niet alleen of een waarde *meer* of *minder* is dan een andere. Deze de kwaliteit betreft *alleen* de grootte van het *verschil tussen* twee waarden, de kwaliteit kan je niet op één enkele waarden plakken, dan is die betekenisloos, natuurlijk lost de vierde kwaliteit dit probleem op: 'een betekenisvolle nul-waarde'. Dan zou je dus zelfs met de eerste van de twee hoogste meetniveaus te maken hebben.

Voorbeelden.
Intervalvariabelen: Temperatuur gemeten in graden Celcius (°C), afstanden *tussen* posities (van punten) op een lijn (of assenstelsel) uitgedrukt in meters, intelligentie gemeten in *IQ-punten* van de Nederlandse Nationale IQ-Test (maar dan moet je je dus niet afvragen waar die 'IQ-punten' precies voor staan (dan behalve een vaste afstand, die optelbaar, deelbaar enz. is), anders wordt het gewoon weer een ordinale variabele), tijdstip van de dag in uren, minuten en seconden.

Intervalwaarden: 27 °C, punt B ligt 0.21 meter van punt D vandaan, 130 IQ-punten, 15:59:59

Statistische toepassing of mogelijke berekeningen.
Hier wordt het leuk, bijna alles mag qua berekeningen, je kunt gemiddelden, of andere statistieken zoals standaardafwijkingen voor een variabele berekenen. Soms wil je scores of verschillende variabelen optellen om een nieuwe of totaal score te berekenen, denk aan een proefwerkcijfer, IQ-scores of test bestaande uit meerdere vragen om de mate van depressie vast te stellen. Scores verhoudingsgewijs bekijken mag dus eigenlijk niet want als je twee scores door elkaar deelt is de betekenis van die verhouding eigenlijk nietszeggend.

Vaak hoor je een uitspraak als 'Het is twee keer zo warm als gister', dus als het bijv. vandaag 6 °C is en gister 3 °C. Maar dit is een Ratio-uitspraak en gaat dus over een kwaliteit die over het ratio-meetniveau gaat. Bij interval toekenningen of uitspraken kijk je slechts naar het *verschil tussen* twee waarden (6-3=3), maar bij Ratio zelfs naar de *deling* van de twee (6/3=2). Een deling is een verhouding, ook wel in Latijn: *ratio*. Alleen als de waarde nul dus zou aangeven dat er geen temperatuur (warmte, trilling, beweging) is, mag je verhoudingsgewijze (ratio) uitspraken doen. Het heeft dus ook geen zin om te zeggen dat iemand twee keer zo goed is als een ander als hij een negen voor zijn proefwerk haalt en die ander een 4.5. Je kan alleen zeggen dat het proefwerk-*cijfer* dus twee keer zo groot is en niet de cognitieve vaardigheid van het object of iets dergelijks.

Als het buiten qua temperatuur 0 °C is, betekent dat dan dat er geen temperatuur (of warmte) is? Natuurlijk niet en daarom zeggen we dat temperatuur in graden Celsius strikt genomen een interval variabele is omdat het dus niet ook de vierde kwaliteit bezit, een echt of absoluut nul-punt of nul-waarde, waarbij je meteen weet dat nul ook echt 'niks of leeg' betekent. Bij temperatuur in graden Celsius is waarde '0 °C' 'toevallig' geplakt op het vriespunt van water bij een bepaalde luchtdruk, het is dus een arbitraire (willekeurige) keuze geweest. We hadden ook kunnen kiezen om het nulpunt van temperatuur op de gemiddelde lichaamstemperatuur te zetten, dan weet je meteen aan het plus of minteken (+/-) van de waarde of de temperatuur van een object hoger of lager is dan die van de gemiddelde mens. Aan de andere kant, het is thuis ook wel makkelijk om vanille-ijs te maken met de gekozen temperatuur-schaal zoals het nu is. Over het algemeen verdient het dan ook de voorkeur om het nulpunt zo handig mogelijk te kiezen voor een (nieuwe) interval variabele. In de praktijk van belang vanwege bijvoorbeeld huilende moeders als ze horen dat hun kind een intelligentie van 0 of zelfs -30 IQ punten heeft), maar als wetenschapper is je keuze vooral van belang voor je verdere (statistische) analyses. Eerst even nadenken scheelt je vaak achteraf een hele hoop werk (onnodig hercoderen, en makkelijkere interpretaties van uitkomsten) als je de getallen zo handig en natuurlijk mogelijk kiest.

Praktische toepassingen.
- In de sociale wetenschappen zijn het vooral de eerste drie meetniveaus die van belang zijn bij onderzoek. Als variabelen zelfs hogere meetniveaus hebben, vinden we dat misschien prachtig, maar niet echt relevant voor nodige of verdere analyses. Bij meer biologische, technische of beta-studies, zijn vaker ook de hogere meetniveaus van variabelen van belang, denk aan biologie, medicijnen, genetica, hersenonderzoek, scheikunde of natuurkunde. Hogere meetniveaus *mag* je altijd negeren en variabelen op een lager meetniveau opvatten of behandelen. In de psychologie is het vaak genoeg om slechts te weten hoeveel punten mensen van elkaar *verschillen* en heeft het geen (praktische) zin om te weten hoe hoog iemand op zichzelf scoort (een absolute score). Mijn dochter had laatst een 5.6 voor haar wiskunde proefwerk, maar het was wel het hoogste van de klas! Bij psychologie gaat het meestal om de grootte van verschillen tussen scores en dus niet de absolute waarde (wat is een angst score van 325 punten) of om verhoudingen daarvan zoals in een uitspraak: 'Jij hebt drie keer zoveel angst als ik'. Zinvolle interval uitspraak zou kunnen zijn: 'Op deze faalangst test scoor ik 14 punten meer dan jij.' zolang je de schaal gewoon in punten laat staan en niet gaat nadenken over de betekenis van een angst-punt nadenkt (dan behalve dat één angst-punt voor een vaste denkbeeldige afstand staat op een getallenlijn).

Ratio

De vierde kwaliteit die wordt *toegevoegd* aan het interval-meetniveau is het absoluut nulpunt (op een schaal of variabele) en zo ontstaat het op - één na - hoogste meetniveau, het ratio-meetniveau. de waarde nul of '0' heeft nu wel echt betekenis en verwijst dus naar 'niks, leeg of afwezig' en is dus niet zomaar ergens opgeplakt. De andere mogelijke absolute (op zichzelf) waarden dan de waarde '0' zoals '0.45', '1', '17' of '99.99' die het object toebehoren, hebben nu (met de eenheid) ook betekenis (en dus niet meer *alleen* betekenis voor het 'verschil' *tussen* twee waarden zoals bij interval). De absolute waarden hebben dus nu ook betekenis op zichzelf en daarmee ook de *verhouding* van twee (absolute) waarden, ook wel *deling*, *ratio* of breuk. Deze kwaliteit van 'betekenis voor 0, absolute waarde en de verhouding' mag dus alleen toegekend worden aan variabelen met (minimaal) een ratio-meetniveau. Als schalen van ratio-meetniveau zijn (of nog hoger) kun je dus nog meer kwaliteiten toekennen en zijn uitspraken als 'ik ben 1.14 keer zo groot als mijn vrouw' dus betekenisvol. Eén resterend probleem is nog wel dat we nog niet meteen kunnen begrijpen wat de éénheid (waarmee gemeten is) precies betekent, hiervoor zijn nog altijd afspraken nodig. De eenheid 'meter' moet dus wel door iemand zijn gedefinieerd. Afspraken over (meet-) éénheden zijn dus altijd nodig voor interpretatie van scores als je met een ratio-meetniveau (of lager niveau) te maken hebt. De vijfde - en hoogste - kwaliteit: 'de

éénheid 1 staat voor alles (of voor perfectie)'. Bij deze kwaliteit is de waarde '1' op zichzelf - dus zonder eenheid - meteen duidelijk en zou het probleem van afspraken over eenheden moeten oplossen. Het 'getal' of score spreekt dan helemaal voor zich. Maar die kwaliteit is dus alleen weggelegd voor waarden met het hoogst mogelijk meetniveau: het absolute meetniveau.

Voorbeelden.
Van ratio-variabelen: Leeftijd in jaren, Massa in kilogram (kg), lengte in meters (m), Temperatuur in Kelvin (K), concentratie van aantal witte bloedcellen (in miljarden) per liter bloed.
Van ratio-waarden: 0.5 jaar, 56 kg, 1.72 m, 273.15 K, 21.3 miljard witte bloedcellen per liter bloed.

Gebruik en toepassingen.
- De eenheid 'meter' voor de ratio-variabele 'lengte' is oorspronkelijk in 1793 gedefinieerd als de afstand van de evenaar tot de noordpool gedeeld door 10 miljoen of ook wel de omtrek (uitgedrukt in een afstand) van de aarde gedeeld door 40 miljoen', dus als je die omtrek in 40 miljoen stukjes verdeelt dan heb je een echte meter te pakken. Tegen woordig gebruiken ze andere (éénduidigere) manieren om af te spreken wat een meter precies is (aan de hand van lichtsnelheid bijvoorbeeld). Gelukkig ligt er ook ergens in een museum in Parijs een *lat* waarmee we kunnen laten zien hoe wij de eenheid 'meter' hebben afgesproken (geoperationaliseerd), maar besef wel dat die lat altijd bij dezelfde (constante) temperatuur en druk moet liggen anders verandert toch echt de (absolute) lengte van die lat.

- Als ik zou zeggen dat ik 1.139 keer zolang ben als mijn vrouw, ken ik dus ook een ratio-kwaliteit toe aan de variabele lengte in cm. Het verhoudingsgetal '1.139' wordt hier ook wel een 'vergrotingsfactor' of gewoon factor genoemd omdat je mijn vrouw haar lengte (151 cm) alleen maar met 1.139 hoeft te vermenigvuldigen om mijn lengte te vinden.

- Temperatuur gemeten in Kelvin (K) word ook wel een 'thermodynamische' temperatuur genoemd omdat temperatuur ook wel te maken heeft met trillingen of bewegingen (vandaar dynamisch) van kleine deeltjes. Als een object een temperatuur heeft van 0 Kelvin kan je dus ook wel zeggen dat alle deeltjes gestopt zijn met bewegen en is er dus een *afwezigheid* van beweging en kan je dus echt zeggen dat de waarde 0 hier dus ook verwijst naar niks (geen beweging van deeltjes). Nul Kelvin komt trouwens overeen met ongeveer '- 273.15 graden Celsius' en is dus de koudst mogelijke temperatuur (geen beweging, warmte of temperatuur) in dit heelal, waarschijnlijk zijn er praktisch geen objecten met precies die temperatuur van 0 Kelvin, maar theoretisch (of hypothetisch) weten we wel wat we bedoelen met 0 Kelvin (geen temperatuur, beweging of warmte). Zo ook met lengte van een mens in centimeters: *niemand* heeft een lengte van nul centimeter, maar we weten wel wat we met '0 cm' bedoelen, namelijk geen (of een afwezige) lengte.

Absoluut

Het laatste - en hoogst mogelijke - meetniveau krijg kent dus nog een kwaliteit extra: 'Het getal of waarde 1, betekent meteen alles of perfect'. Er zijn niet veel variabelen die hieraan voldoen en veel boeken benoemen dit meetniveau niet eens. Waar het bij deze laatste kwaliteit om gaat, is dat de eenheid voor ons getallen gebruik (dus gewoon het getal of waarde '1') meteen één heel ding (alles aanwezig of perfectie) aangeeft. Behoorlijk vaag dus. Er bestaat dan ook wel wat onenigheid over de precieze betekenis van dit laatste meetniveau. Een kans is een getal of waarde dat je uitdrukt in een getal dat bij '0' begint als laagste mogelijke waarde en als hoogste mogelijke waarde een '1' krijgt. Alle getallen ertussen zijn natuurlijk ook mogelijk bij kansen. Is een kans dan ook een variabele? Ja, omdat een kans ook iets is wat (per object) kan variëren. De ene persoon heeft nou eenmaal meer kans om te slagen voor een bepaald vak dan een ander en daarmee varieert dus de mogelijke kans op slagen. We drukken dus kansen in getallen uit, maar besef dus dat je geen eenheid nodig hebt om het getal duidelijk te maken, de waarde staat dus

helemaal op zichzelf omdat het alle mogelijke informatie of kwaliteiten aanwezig zijn.

Voorbeelden.
absolute-variabelen: kans, proportie en correlatie
absolute-waarden: een kans van 0.8, een proportie van .5, of een correlatie van .3

Toepassing en gebruik.
Een kans-waarde van '0' voor het slagen voor een tentamen betekent dat iemand het tentamen *sowieso* niet zal halen. Hoe hoger de waarde wordt, des te meer zal een persoon geneigd zijn om het tentamen te halen. Bij een waarde van '.5' betekent het dat er net zoveel kans is op slagen als op zakken (bij herhaling zal iemand het tentamen even vaak halen als dat hij er voor zakt) van bijvoorbeeld '.8' dat je naar verwachting 4 van de 5 (zelfde) tentamens zou halen. En ten slotte: de waarde 1 bij een kans betekent dat het verschijnsel waar de kans voor bedoeld is sowieso zal optreden, alles ervan is dus aanwezig. Iemand zal bij een waarde van 1 het tentamen zonder enige twijfel halen en is de uitkomst (het halen van je tentamen) dus geheel bepaald of gedetermineerd.

Een correlatie is een getal tussen de -1 en 1 die aangeeft in hoeverre twee variabelen samenhangen, zoals lengte en gewicht van mensen samenhangen, hoe kleiner iemand is, des te lager zal zijn gewicht zijn (we spreken hier over een positief verband omdat de twee variabelen bij een object dus *vaak* dezelfde kant op wijzen (in dit geval laag en laag, of klein en licht, allebei dus negatief afwijken van het gemiddelde).

Het negeren van hogere meetniveaus

Als twee mensen dus twee verschillende waarden hebben voor bijvoorbeeld de kans op slagen, zeg de waarden 0.7 en 0.9, *kan* je die informatie puur nominaal opvatten door alleen maar te zeggen dat er *verschil* is tussen de twee personen omdat '0.7' en '0.9' nou eenmaal anders klinken. Als alleen van belang is dat de ene persoon *meer* kans heeft op slagen, kun je dus met een ordinale uitspraak of kwaliteit volstaan: 'De ene persoon is meer geneigd te slagen dan de andere'. Ook al heeft een kansvariabele dus strikt genomen zelfs een absoluut meetniveau, je kan een variabele dus altijd op een lager niveau gebruiken of opvatten door informatie of kwaliteiten te negeren. Let dus op het verschil in het meetniveau qua uitspraak of toepassing enerzijds en het strikte meetniveau van variabelen anderzijds. Binnen de sociale wetenschappen is vaak de toekenning van interval kwaliteiten verreweg voldoende voor de meeste statistische analyses.

Gebruik en toepassing.
- Als ik zeg dat het verschil qua lengte in cm tussen mij en mijn vrouw twee keer zo groot is als het verschil qua lengte tussen mij en mijn dochter, ken ik dus een interval-kwaliteit toe aan onze scores. Omdat ik slechts de afstanden *tussen* waarden benoem en daarvan de relatieve groottes vergelijk en dit mag alleen bij interval variabelen (of van hoger meetniveau) terwijl strikt genomen de variabele lengte in cm van ratio-meetniveau is. Maar soms hanteer of benoem je een lager meetniveau, dus soms moet je het verschil weten.

Andere soorten indelingen (dan het meetniveau) voor soorten variabelen

Continu versus discreet

We hebben nu het meetniveau gehad, maar er zijn nog meer indelingen waar we op moeten letten. Neem een dobbelsteen, noem hem voor het gemak even de variabele X. Je kan wel de waarde $X = 3$ of $X = 6$ gooien, maar het is onmogelijk om bijvoorbeeld $X = 3.5$ te gooien. Omdat deze variabele X nogal *beperkt* is qua aantal mogelijkheden noem je deze variabele discreet (als je vreemd gaat, wil je ook graag dat je vrienden daar discreet of dus - beperkt - over zijn). Als een variabele een onbeperkt aantal waarden heeft noemen we het een continue variabele, zoals bij lengte in cm. Er zijn natuurlijk waarden die (praktisch) niet voorkomen, zoals een persoon van 289 cm lang. Maar tussen bijvoorbeeld 170 cm en 172 cm liggen *oneindig* veel andere waarden zoals bijv. 171.3567 cm. Je kunt ook wel denken dat als je de mogelijke waarden van een

dobbelsteen op de getallenlijn zet, moet je 'springen' om van 1 naar 2 te gaan. Maar als je de waarden van lengte in cm (van personen) op de getallen lijn uitzet hoef je nooit te springen om bij een andere waarde uit te komen, want elk punt op die lijn staat voor een mogelijke waarden. Een lijn of *lijnstuk*, bestaat per definitie uit oneindig veel punten (dus waarden). Sowieso zijn nominale en ordinale variabelen *altijd* discreet en bij de hogere meetniveaus (interval, ratio en absoluut) hangt het dus van het aantal én de ligging van mogelijke waarden af. Zijn het er *oneindig* veel én liggen de waarden echt tegen elkaar aan, dan noemen we het een continue variabele en zijn het een beperkt aantal opties dan noemen we hem dus discreet van karakter. Heeft een variabele slechts twee waarden (niveaus of categorieën) zoals bij geslacht, zakken of slagen, wel of niet ziek zijn of een andere groepsindeling waar je maar keuze hebt uit twee groepen, noemen we hem ook wel *dichotoom*, *binair* of op zijn echt Nederlands; 'tweewaardig'. Dichotome variabelen zijn heel handig voor statistische analyses als je die variabelen - of beter - *de twee categorieen* (her)codeerd met eentjes (bijv. ziekte aanwezig) en nulletjes (ziekte afwezig) dan kun je hele leuke analyses doen (bijvoorbeeld een regressie-analyse).

Voorbeelden.
Voorbeelden van discrete variabelen: alle nominale, ordinale en dichotome variabelen, geslacht, al dan niet slagen, aapsoort, leeftijdscategorie (jong, middel, oud), nationaliteit. soort ziekte of aandoening, een telling (van aantal mannen in een steekproef, je kan nooit een telling van '70.5' man vinden)
Voorbeelden van continue variabelen: Lengte in centimeter, leeftijd in jaren, lichaamstemperatuur in graden Celsius, correlatie, kans (op slagen), proportie (gedeelte).

Toepassing en misbruik.
Ook hier word weleens gesmokkelt en wordt gedaan alsof een telling (discreet) zich toch continue gedraagt. Een strikt continue variabele heeft meer kwaliteiten (meer mogelijkheden) dan een discrete variabele. Soms passen we 'continuïteits correcties' toe zodat we toch weer de analyses kunnen doen die we graag zouden willen doen en soms zijn we lui en laten we het voor wat het is en smokkelen we een beetje.

Het Beschrijven van Data aan de hand van Statistieken en Parameters

Als je weet wie en wat je wilt onderzoeken is het dus zaak om eerst informatie over de te onderzoeken verschijnselen te verzamelen, dus je metingen doen bij je respondenten (onderzoeksobjecten). Bij het verzamelen van de data ben je in de 'Toetsingsfase' van de empirische cyclus van de Groot (de Groot, 1961) beland. Hier gebeurt eigenlijk het meeste werk:
- Metingen verrichten van alle mogelijke verschijnselen binnen jouw onderzoek (data-verzameling),
- vervolgens voer je de data in een statistiek programma zoals 'SPSS' of 'R' (data-verwerking),
- daarna analyseer je de data (Data-Analyse of ook wel gewoon: berekeningen uitvoeren).
- Aan de hand van je data-analyse of beter - je resultaten of uitkomsten - kun je *beschrijven* wat er gebeurt binnen jouw steekproef én kijk je of jouw bevindingen te *generaliseren* vallen naar situaties buiten jouw steekproef door conclusies op basis van je analyses te trekken. Dit behoort allemaal tot de Toetsingsfase en deze handleiding gaat dus ook vooral over alle processen binnen deze fase.

Data-punten

We verzamelen dus informatie over verschijnselen (waarden op variabelen) die de onderzoeksobjecten (proefpersonen) toebehoren. Het woord 'datum' betekent ook wel 'gegeven' of 'informatie'. Eén datapunt is één enkel gegeven (over een object), bijvoorbeeld iemands lengte (172 cm). Een datapunt staat dus voor één enkele waarde (van een variabele voor een bepaald object) en als je meerdere datapunten hebt verzameld, dan heb je te maken met een data-verzameling - of beter - een *dataset*. Een dataset bestaat dus uit (heel veel) verzamelde waarden die weer onder te verdelen zijn (in één of) meerdere variabelen. Een van de eerste stappen na het verzamelen van je data(punten) is het *beschrijven* van je gegevens. De

meest precieze beschrijving van je data, zou een *letterlijke opsomming* zijn van *al* je datapunten zijn, maar dat zou wel heel vervelend en saai worden, het is helemaal niet raar als je meer dan tienduizend waarnemingen of waarden hebt verzameld. Beschrijven doen we hier aan de hand van een *samenvatting* van onze data. De dataset *samenvatten* (beschrijven) doen we aan de hand van statistieken.

Het beschrijven van steekproefdata

Een statistiek is een beschrijvende waarde (meestal uitgedrukt in een getal) die een eigenschap of karakterestiek van de dataset beschrijft. Een statistiek zegt *altijd* iets over datapunten uit een *steekproef*, dus is een beschrijver of samenvatter van steekproefdata. Er bestaan natuurlijk meerdere manieren om iets over een verzameling datapunten te zeggen: soms wil je iets zeggen over waarden die onder één enkele variabele vallen, maar vaak wil je ook meer zeggen, bijvoorbeeld *hoe* (waarden van) verschillende variabelen bijvoorbeeld samenhangen in je steekproef, ook wel het verband of correlatie tussen twee variabelen genoemd.

Het beschrijven van populatiedata

Natuurlijk is het doel van een wetenschapper om uiteindelijk uitspraken te doen over populatiegevens. Populatiedata (of gegevens) beschrijven we aan de hand van 'parameters' (met de klemtoon op 'ra'). Dus het gemiddelde van een populatie noemt men een parameter. We hebben hier alleen een *ontzettend* groot probleem. Niemand is in staat om met absolute zekerheid iets zeggen over sitiuaties waarin niet alle objecten gemeten zijn. Een populatie (van objecten) is per definitie *oneindig* groot qua aantal objecten. Omdat dus niet alle gegevens praktisch te verzamelen zijn, kun je nooit precies zeggen of beschrijven wat er in een populatie precies of exact gebeurt. Dus het ware of echte gemiddelde voor een bepaalde populatie kan dus niemand weten (behalve God of een ander *alwetend* figuur en dat is behoorlijk vaag). Voor de notatie van parameters gebruiken we altijd griekse letters (ook lekker vaag dus). Onthoud dus: vage dingen (parameters) doen we aan de hand van vage letters (griekse alfabet).

TABEL OF

soort statistiek	soort maat	bijbehorende statistiek	bijbehorende parameter	uitspraak van parameter
gemiddelde	centrummaat	$\bar{x}$	μ_x	mu
standaardafwijking	spreidingsmaat	S_x	σ_x	sigma
variantie	spreidingsmaat	S_x^2	σ_x^2	sigma kwadraad
covariantie	ruwe samenhangsmaat	S_{xy}	σ_{xy}	sigma
correlatie (coëfficient)	gestandaardiseerde samenhangsmaat	r_{xy}	ρ_{xy}	rho
regressiegewicht	constante, startgetal of *intercept*	b_0	β_0	beta
regressiegewicht	richtingscoëfficiënt / *slope*	b_1	β_1	beta

Voorbeelden.
Voorbeelden van statistieken voor een variabele (X), dus voor steekproefdata: gemiddelde $\bar{X}$, variantie S_x^2, standaardafwijking S_x.
Voorbeelden van parameters voor de relatie tussen twee variabelen (X en Y) voor populatiedata: correlatie ρ_{xy}, covariantie σ_{xy}, regressiegewicht β_1.

Statistieken (steekproef) en Parameters (populatie)

Natuurlijk kunnen we wel *schatten* wat er *ongeveer* gebeurt in een populatie op basis van een steekproef (een deelverzameling van je populatie die je wilt onderzoeken). De waarde voor het echte gemiddelde voor een variabele van een populatie (de parameter 'het gemiddelde') kan dus niemand weten, maar we gebruiken een statistiek (het gemiddelde van een steekproef) als beste schatting voor het ware gemiddelde van de populatie. Een statistiek is dus een schatter

(of puntschatting, omdat het één enkele waarde is) voor een parameter. De puntschatting (statistiek) hoeft dus niet precies hetzelfde te zijn als het echte gemiddelde (parameter), maar naarmate je steekproef groter is, zal de *waarschijnlijkheid* (of kans) dat jouw puntschatting (statistiek) meer lijkt op de ware waarde van de *parameter* toenemen. In het algemeen: Hoe groter de steekproef, des te betrouwbaarder (precieser, accurater) worden jouw schattingen voor echte populatie-waarden of verschijnselen.

Het beschrijven van data aan de hand van statistieken, *Descriptive Statistics*.

1§1 Waar ligt de data? Centrummaten

Als wetenschapper moeten we gebeurtenissen om ons heen kunnen beschrijven. We willen bijvoorbeeld weten hoe slim (in IQ-punten) of hoe lang (in centimers) de respondenten in een steekproef zijn. Dit beschrijven van de verkregen data (de verzameling opgemeten scores van je proefpersonen in je steekproef) is eigenlijk altijd stap één van de data-analyse in een onderzoek. Om één van de variabelen in een dataset te kunnen beschrijven, hebben we twee hoofdvragen nodig.

De eerste vraag gaat over *waar* de data zich bevindt. Bij deze vraag denk je aan een maat die het centrum of een soort midden van alle datapunten (de opgemeten scores van alle individuen) aangeeft. Het gemiddelde (*mean*), de mediaan (*median*) of de modus (*mode*) kunnen hiervoor gebruikt worden. Afhankelijk van de data geeft de ene maat een handigere beschrijving dan de andere, maar alledrie de maten zijn bedoeld om aan te geven rond welk punt (getal of waarde) de data (alle waarden in je steekproef) zich bevindt.

Voor de tweede vraag willen we weten hoe de data *verdeeld* is of hoe de datapunten in je steekproef ten opzichte van elkaar *verschillen* of *variëren*. Hierbij kun je denken aan de vraag in hoeverre de verschillende scores bij elkaar of juist verder van elkaar verwijderd liggen.

Dit beantwoordt dus de vraag in hoeverre de scores op elkaar lijken (homogeen zijn) of juist verschillen (heterogeen zijn). Om de mate van verschillen aan te geven voor de scores op een bepaalde variabele, gebruiken we de zogenaamde spreidingsmaten zoals de standaardafwijking (*standarddeviation*) en de variantie (*variance*). Er zijn meerdere manieren uiteraard.

Soms is het goed mis met de data en gedragen de scores zich niet zoals we graag zouden willen zien en gebruiken we andere trucjes om toch de data te kunnen beschrijven. Grafisch, in een grafiekje wordt snel duidelijk hoe de data zich gedraagt (waar de verschillende scores zich bevinden) en dus verdeeld is. Ik zou het hier kunnen gaan vertellen, het middel en het doel, maar een van de belangrijkste lessen geef ik je meteen mee: Eerst doen en dan pas gaan denken! Het lijkt een beetje op de film *Karate Kid*, frustrerend maar waar. De leerling in deze film moet allerlei – in eerste opzicht – niet gerelateerde oefeningen doen. Hij wil natuurlijk gewoon vechten. Uiteraard wordt hij dik beloond in een of ander duel waar hem dán pas duidelijk wordt waarvoor hij de zinloze oefeningen eindeloos moest herhalen (muren verven met een bepaalde beweging of zijn jas tot in de eeuwigheid op – en af – hangen). Dat geldt hier in de statistiek (vechten) ook dus en kunnen we het beste maar gewoon beginnen met de opgaven (verven). Voor het grootste deel behandel ik in de opgaven de te leren stof en formules. Je zult dus wel moeten 'doen'! Succes.

Belgedrag onder jongeren

In een fictief onderzoek werd gekeken naar het belgedrag onder jongeren. In een steekproef van 20 personen (n=20, de kleine letter n staat in de meeste gevallen voor het aantal mensen in een steekproef of een conditie) werd de respondenten gevraagd hoeveel minuten zij de voorgaande week gebruik hadden gemaakt van hun mobiele telefoon. In de tabel hieronder de verkregen data. Voorlopig gebruik ik 'X_i' ter vervanging van de variabele 'het aantal belminuten', zolang ik dus nog in het algemeen spreek over de variabele 'het aantal belminuten' en nog niet weet over specifiek welke persoon met welke waarde ik het heb. De waarde waar X_5 voor staat is 25. of anders gezegd: De score X voor persoon nummer 5 heeft de waarde 25. De kleine letter i (het subscript) staat dus voor respondent- of proefpersoonnummer. De i-tjes zijn er eigenlijk slechts ter organisatie en als je dus weet over welke persoon het gaat, gebruik je zijn nummer in plaats van de letter i.

TABEL 1A

i	X_i
1	13
2	18
3	25
4	58
5	25
6	31
7	39
8	42
9	17
10	35
11	46
12	22
13	18
14	20
15	26
16	14
17	33
18	19
19	20
20	21

Aan de hand van een aantal statistieken en grafieken en dergelijke gaan we de data beschrijven. Er is dus een verschil tussen een 'statistiek' en 'de data'. De data is de verzameling van de scores (datapunten) in onze steekproef en een statistiek heeft een overstijgende functie. Een statistiek is een *beschrijvend getalletje* en zegt dus iets over - of beschrijft - de data (als geheel).
Voordat we beginnen is het handig om de data (de verschillende scores) op volgorde (rangorde) te zetten, zodat we bijvoorbeeld makkelijk de mediaan kunnen berekenen. Neem de data over en orden de scores van lage naar hoge waarden. Je kunt de verschillende scores opnieuw nummeren en de oorspronkelijke waarden voor i negeren, zodat nu de persoon met de laagste score, de waarde i=1 krijgt en de hoogste score i=20. De verschillende waarden voor i representeren dan meteen de verschillende rangnummers.

Opgaven

1.1a Als eerste statistiek bekijken we de mediaan. De mediaan is een centrummaat en hoort dus bij de vraag waar de data zich bevindt en is de waarde van de variabele voor de middelste observatie, persoon of rangnummer. Op welke plek ligt de mediaan?

1.1b Welke waarde heeft de mediaan? (er is dus een verschil tussen de vraag '*Waar* ligt de mediaan?' en '*Wat* is de (waarde van de) mediaan?')

1.1c Welke waarden hebben Q_1 en Q_3? Welke X-scores vallen in het eerste quartiel?

1.1d Geef de *Five-Number Summary* voor de score X.

1.1e Teken een boxplot zonder uitbijters (ofwel de zogenaamde *outliers*) betrek dus alle scores bij de boxplot. De twee staarten, aan de onder en bovenkant van de boxplot lopen dus tot en met het minimum en het maximum van de scores.

1.1f Om wel rekening te houden met eventuele uitbijters gaan we nu een *modified boxplot* fabriceren. Hiervoor moet je eerst de waarde van de IQR (*interquartile range*) weten. Bereken deze waarde. Een score wordt als uitbijter gezien als deze boven de waarde '$Q_3+1.5\cdot IQR$' valt en wordt dan slechts als puntje aangegeven. De bovenkant van de boxplot (staart) eindigt dan bij, of op de laatste score die nog wel echt mee mag doen en dus nog net op of onder de waarde $Q_3+1.5\cdot IQR$ valt. Voor uitbijters aan de linker of onderkant van de verdeling doe je hetzelfde maar kijk je vanaf Q_1 maar dan wel in de tegengestelde richting als bij de andere staart. Dus $Q_1-1.5\cdot IQR$ is dan de ondergrens van scores die nog mee zouden mogen doen voor de staart aan de onderkant van de boxplot.

1.1g Soms is er sprake van scheefheid van de verdeling van de scores in de data (*skewness*). Is hier sprake van een verdeling die links of rechtsscheef is? Zou het gemiddelde ($\bar{X}$) links of rechts van de mediaan liggen? Je hoeft het gemiddelde nog niet uit te rekenen, geef slechts een schatting.

1.1h Wat is hier de waarde van de modus? Welk probleem ontstaat hier? In hoeverre kunnen we hier zeggen dat de modus een goede beschrijver voor (deze) data is?

1.1i Waar zou de mediaan liggen als er een oneven aantal observaties zouden zijn, bijvoorbeeld bij 21 scores waarbij de waarde van de nieuwe score 83 is? Teken ook voor deze 21 gevens een gewone boxplot en eentje die aangepast is waarin de *outliers* zichtbaar zijn.

1.2a Een andere manier om grafisch de verdeeldheid van de data te laten zien is een *stem-and-leaf plot.* De zogenaamde stam met blaadjes met in de stam de tientallen en de blaadjes zijn de waarden binnen die tientallen die in de data voorkomen, als een score drie keer voorkomt, krijg je ook drie dezelfde blaadjes te zien. Teken een *stem-and-leaf plot*. Wat kunnen we over de vorm zeggen? Links of rechtsscheef Komt dit overeen met de eerder getekende boxplot?

1.3a Bereken het gemiddelde voor de X-scores ($\bar{X}$ of $E(X)$), ook wel de verwachting (of *expected value*) voor de variabele X. Zeker bij scores die zich wenselijk gedragen (hier kom ik later op terug) is het gemiddelde de meest gebruikte statistiek om het centrum van de data te beschrijven.

1.3b Loop de voorgaande vragen nog een keer door en bekijk de verschillen tussen de mediaan, de modus en het gemiddelde. Hoe hadden de boxplot en de stem-and-leaf plot eruit kunnen (of moeten) zien zodanig dat het gemiddelde, de mediaan en de modus ongeveer gelijk waren geweest?

Uitwerking

TABEL 1B

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X_i	13	14	17	18	18	19	20	20	21	22	25	25	26	31	33	35	39	42	46	58

1.1a De mediaan ligt tussen het vakje 10 en 11, dus eigenlijk bij rangnummer 10.5. Je kunt dus zeggen wanneer er een even aantal observaties is dat de mediaan altijd de data in twee gelijke stukken hakt (twee stukken van 50 procent) én de mediaan dus tussen de twee middelste observaties komt te liggen. Wanneer er een oneven aantal observatie zijn wordt de data ook in twee gelijke stukken gehakt, maar ligt de mediaan op de (enige) middelste persoon (of rangnummer).

1.1b $Med = \frac{22 + 25}{2} = 23.5$

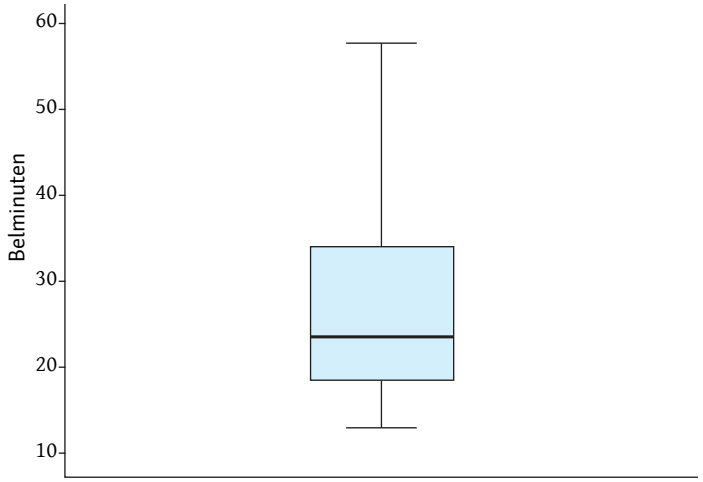
1.1c $Q_1 = 18.5$ deze waarde is eigenlijk niks anders dan de mediaan van de linker helft van je data, dus behorende bij de eerste tot en met de tiende waarneming.
$Q_3 = 34$ deze waarde is eigenlijk niks anders dan de mediaan van de rechter helft van je data, dus behorende bij de tiende tot en met de twintigste waarneming.

1.1d *The Five-Number Summary* = $\overline{min\ Q_1\ med\ Q_3\ max}$ = $\overline{13\ \ 18.5\ \ 23.5\ \ 34\ \ 58}$

De scores die in het eerste quartiel liggen zijn: 13, 14, 17, 18 en nog een keer 18.

1.1e

figuur 1A



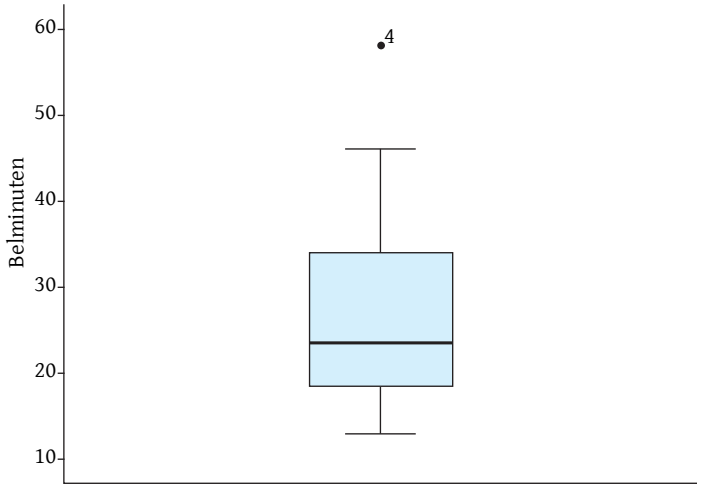
1.1f *IQR* = 34−18,5 = 15.5

1.5·*IQR* = 23.25

34+23.25 = 57.25 (de bovengrens van de boxplot, hoger dan deze waarde zou de *staart* dus nooit kunnen komen, maar de *outliers* dus wel!)

18.5−23.25 =−4.75 (de ondergrens van de boxplot, belminuten kunnen niet negatief zijn, dus *no worries*)

figuur 1B



1.1g Rechtsscheef. Omdat er een aantal scores vrij hoog zijn ten opzichte van de rest zal het gemiddelde iets omhoog worden getrokken en dus rechts (of boven) van de mediaan liggen.

1.1h 18, 20 en 25 zijn de modussen, want die komen alle drie het meest (2 keer) voor. Vreemd, maar dit kan gewoon. Blijkbaar zijn er dus meerdere scores in de mode!

1.1i Als we 21 observaties hadden geteld, was de mediaan toebedeeld aan rangnummer 11, dan zouden er dus tien scores onder en boven die waarde vallen. Stel dat de waarde van die 21ste score bijvoorbeeld 83 zou zijn, erg hoog dus, dan zou dat niet zoveel uit maken voor de waarde van de mediaan. Die wordt dan de score 25 die hoort bij de elfde persoon (de score van de middelste persoon ingeval van 21 scores). Je zou kunnen stellen dat door toevoeging van een score de mediaan slechts een half plekje naar rechts schuift (in ons voorbeeld dus van 10.5 naar rangnummer 11) onafhankelijk van de waarde van die toegevoegde score. Omdat de (hoge) waarde van die toegevoegde score dus eigenlijk niet van belang is voor de mediaan, kunnen we ook stellen dat de mediaan robuust (opgewassen tegen) is voor outliers. De mediaan is dus vrij onveranderlijk of stabiel als het gaat om de invloed van outliers in de data op deze statistiek (mediaan).

minimum = 13

$Q_1 = \frac{18+19}{2} = 18.5$ Dus eigenlijk de mediaan van de linker helft, deze loopt ook van 1 t/m 10 en niet van 1 t/m 11!

med = 25 De elfde observatie deelt nu de data in twee gelijke stukken

$Q_3 = \frac{35+39}{2} = 37$ Dus de mediaan van de rechterhelft, deze loopt van 12 t/m 21 en de twee middelste rangnummers daarvan zijn nu 16 en 17!

maximum = 83

$IQR = Q_3 - Q_1 = 37 - 18.5 = 18.5$ De IQR heb je dus nodig om eventueel een *modified boxplot* te maken.

$1.5 \cdot IQR = 1.5 \cdot 18.5 = 27.25$

$Q_3 + 1.5 \cdot IQR = 37 + 27.25 = 64.25$ De bovenkant of de rechterstaart van de modified boxplot, loopt dus tot de waarde in de data die nog net onder of even groot is als 64.25 ofwel t/m de waarde 58 dus.

De ondergrens van deze boxplot is gewoon weer 13 omdat de uiterste grens weer negatief zal zijn en dus sowieso niet zal voorkomen in de data.

1.2a

1	3 4 7 8 8 9
2	0 0 1 2 5 5 6
3	1 3 5 9
4	2 6
5	8

De verdeling is rechts scheef, als je de *stem- and leaf-plot* een kwartslag tegen de klok indraait, zit de staart rechts.

1.3a

$$\bar{X} = E(X) = \frac{\sum_{i=1}^{i=n} X_i}{n} = \frac{1}{n} \cdot \sum_{i=1}^{i=n} X_i$$

Het sommatieteken $\sum$ met *i*=1 eronder en *i*=*n* erboven wil zeggen dat je datgeen wat achter het sommatieteken staat, in ons geval dus alleen X_i, eerst voor alle waarden van i (dus *i*=1 t/m *i*=*n*=20) moet invullen en vervolgens bij elkaar moet optellen. Als ik dus alleen de sommatie voor alle X-waarden zou willen hebben voor i=1 t/m i=20 krijg je dus het volgende:

$$\sum_{i=1}^{i=20} X_i = [13+14+17+18+18+\ldots\ \ldots+46+58] = 542$$

Om een sommatieteken uit te werken gebruik ik vaak rechte haakjes. Alles links van het eerste is-gelijk-aan-teken noem ik ook vaak wel dat 'ding', maar dat ding heeft dus een waarde van 542 (als je hem netjes en dus correct hebt uitgewerkt).
Let op: vaak staat er niks onder of boven het sommatieteken, ze bedoelen dan dat je gewoon elke waarde moet nemen

$$\bar{X} = E(X) = \frac{[13+14+17+18+18+\ldots\ \ldots+46+58]}{20} = \frac{542}{20} = 27.1$$

of: $\bar{X} = E(X) = \frac{1}{20} \cdot [13+14+17+18+18+\ldots\ \ldots+46+58] = \frac{1}{20} \cdot 542 = 27.1$

Zeker voor later is het handig als je beseft dat delen door 20 hetzelfde is als vermenigvuldigen met het omgekeerde van 20. Het omgekeerde van een getal is hetzelfde 1 gedeeld door dat getal, in dit geval dus $\frac{1}{20}$.

1.3b Het gemiddelde, de mediaan en de modus verschillen van elkaar. Als de boxplot of de *stem-and-leaf plot* perfect symmetrisch waren geweest, waren het gemiddelde en de mediaan aan elkaar gelijk geweest. De modus is een beetje een apart verhaal, maar ook hier geldt grofweg

hetzelfde. Bij symmetrie in je data (en een normaal-verdeling, kom ik later op terug) zou je wel verwachten dat de modus hetzelfde zou moeten zijn als het gemiddelde en de mediaan.

1§2 Hoe is de data verdeeld? Spreidingsmaten

In paragraaf 1.1 hebben we gekeken naar de vraag 'Waar ligt de data?' Nu gaan we kijken in hoeverre de data verdeeld is. Als je weet waar (het gros van) de data zich bevindt kun je een voorspelling doen voor iemand uit je steekproef – of zelfs buiten je steekproef, bijvoorbeeld iemand uit de rest van de populatie. De data uit het belgedrag onderzoek heeft als laagste waarde een score van 13 en als hoogste een score van 58. Als je dan toch een voorspelling of gok moet doen voor een willekeurig persoon uit je steekproef, is het wel handig als je in ieder geval een waarde tussen deze twee uiterste doet.
Het gemiddelde is vaak de beste gok mogelijk. Niet iedereen – of vaak zelfs niemand - binnen je steekproef heeft precies dezelfde waarde als het gemiddelde, maar toch is het handig om het gemiddelde als beste gok of voorspelling te gebruiken. Het gemiddelde gebruiken als beste gok wil dus ook niet zeggen dat je dan altijd precies goed gokt, maar wel dat je er gemiddeld gezien (als je dus vaker dan 1 keer zou gokken) er het dichtst bij zit met je gok. Anders gezegd: als je het gemiddelde gebruikt als beste voorspelling voor een willekeurig persoon uit je steekproef, zal je gemiddelde gokfout het kleinst zijn.
Wanneer de scores van de verschillende proefpersonen in een steekproef dicht bij elkaar liggen, dan zal het duidelijk zijn dat de gemiddelde gokfout ook kleiner zal zijn dan als de proefpersonen onderling juist grotere verschillen vertonen op hun scores. De gemiddelde gokfout heet ook wel de standaardafwijking of standaarddeviatie (de officiële naampjes) en is dan ook een heel handige en veel gebruikte spreidingsmaat. Zoals je later niet zal ontgaan: de standaardafwijking is *heilig* binnen de statistiek.

We gaan verder op de data uit het belgedrag onderzoek. Voor de volgende vragen is het het handigst als je SPSS gebruikt om je handmatige (GR) berekeningen te controleren. Open het bestand 'belminuten.sav' met SPSS. De volgende commando's zijn voorlopig voldoende.

Analyse > *Descritpive Statistics* > *Frequencies...* of *Descriptives...*

Om vervolgens specifieke statistieken op te vragen, zul je nog wel de nodige *statistics* moeten aanvinken.

Opgaven

1.4a Bereken het gemiddelde voor de X-scores, ook wel de verwachte waarde (of *expected value*) voor de variabele X.

1.4b Van waar tot waar bevinden zich de mogelijke scores? Hoe groot of lang is de lengte van het interval waarop de geobserveerde scores zich bevinden anders gezegd hoe groot is de spreidingsbreedte (*range*) van onze scores.

1.4c Als je het gemiddelde gebruikt als beste gok voor een willekeurige score, wat is dan de kleinste en wat is dan de grootste gokfout (absoluut gezien)? Geef een schatting voor de gemiddelde gokfout. Wat ik onder een gokfout voor een willekeurige persoon in je steekproef versta, is de afstand tussen de geobserveerde score van die persoon tot het gemiddelde. In het geval van belminuten, onze variabele, is het dus het verschil (afstand, afwijking) tussen de score van een persoon en het gemiddelde op het aantal bel minuten, ook wel zijn individuele afwijking tot het gemiddelde genaamd. Als een persoon rechts (boven) van het gemiddelde zit qua score dan zien wij dit als een positieve afwijking en negatief als hij zich links van het gemiddelde bevindt. In welke volgorde moet je dan telkens de twee waarden van elkaar aftrekken om niet onterecht een positieve (of negatieve) afwijking te vinden?

Handmatige berekening

We gaan over tot de echte berekening van de standaardafwijking. Handmatig is het een hoop werk. Aangezien we meestal met statistiek programma's zoals SPSS werken, hoeven we dit zelf dus eigenlijk nooit te doen. Waarom gaan we dit dan doen? Omdat het de nodige rekenvaardigheid en inzichten zal verschaffen die we nodig hebben om later verdere stappen

te kunnen nemen. De standaardafwijking is dus de gemiddelde gokfout als je het gemiddelde zou gebruiken als beste gok, of ook wel de gemiddelde afwijking (van een observatie) naar het gemiddelde (voor alle observaties). Je zou misschien kunnen denken: dan tel je toch gewoon alle individuele afwijkingen bij elkaar op en deel je het totaal door het aantal afwijkingen? Maar helaas, het idee is wel juist, maar we komen onderweg een aantal problemen tegen waar we een correctie voor zullen moeten maken.

Opgaven

1.5a Maak een 21 x 4 tabel (21 rijen en 4 kolommen). Zet in de tweede rij vanaf de eerste kolom de nummering 1 t/m 20 (de i-tjes) en daarachter in de tweede kolom de bijbehorende X-scores. Op de eerste rij in het eerste vakje zet je een *i* en in het tweede vakje zet je X_i. Bereken nu voor iedere persoon (observatie) in onze steekproef zijn individuele afwijking naar het gemiddelde, dus de gokfout die je zou maken als je voor die persoon het gemiddelde zou nemen voor zijn voorspelling en dan zou berekenen hoeveel hij er dus naast zit. Iedereen die links van het gemiddelde zit, moet uiteindelijk dus ook een negatieve waarde als gokfout krijgen.

1.5b Eigenlijk stuiten we hier al op het eerste probleem. Als we volgens het *verkeerde idee* (de gokfouten optellen en delen door het aantal) de individuele afwijkingen nu zouden optellen, dus de sommatie ervan zouden nemen, welke waarde krijgen we dan? Om dit probleem te omzeilen kwadrateren we eerst alle gokfouten voordat we ze gaan optellen. Zet de gekwadrateerde gokfouten in de vierde kolom. Komen in deze kolom nog negatieven waarden voor?

1.5c Als je de waarden in de vierde kolom zou optellen, hoe zou je die waarde in woorden kunnen omschrijven? Bereken deze waarde.

Vrijheidsgraden, degrees of freedom

Hier stuiten we op het tweede probleem als we het *verkeerde idee* zouden volgen. Volgens het verkeerde idee zouden we nu de gevonden kwadratensom (de som van alle gekwadrateerde afwijkingen) moeten delen door het *aantal* afwijkingen of observaties, n. We hadden immers 20 gokfouten berekend en we willen weten wat de grootte is van de gemiddelde gokfout. Maar zijn er ook daadwerkelijk 20 gokfouten bij 20 observaties?
Stel je bent bij een echtpaar met drie kinderen (drie observaties) thuis en het echtpaar vertelt je dat deze drie kinderen gemiddeld een leeftijd hebben van 10 jaar, maar ze spelen nu buiten. Wanneer het eerste kind thuis komt, roep jij natuurlijk dat dat kind wel 10 jaar oud zal zijn, je bent immers een wetenschapper en gaat dus uit van de beste gok! Het blijkt echter dat het kind 8 jaar is. Voordat het tweede kind binnenkomt roep je natuurlijk weer dat ook dat kind wel 10 jaar zal zijn, maar helaas je hebt het alweer mis want het is 9 jaar oud. Alleen het derde kind moet nog thuis komen, maar welke leeftijd zou je nu 'gokken'? Omdat je weet dat er drie kinderen zijn, kun je nu de leeftijd van het laatste kind uitrekenen en hoef je dus niet meer te gokken! Samen moeten de kinderen 30 jaar oud zijn om een gemiddelde leeftijd van tien te krijgen. Je kan dus voordat het derde (laatste) kind binnenkomt al zeggen dat het 13 jaar oud zou moeten zijn. In dit geval maak je dus eigenlijk maar twee gokfouten omdat de laatste waarneming 'vast' ligt als de voorgaande bekend zijn. Je mag hier zeggen dat 2 observaties de vrijheid hebben en dat er 1 dus vast ligt. In een set of verzameling van 20 observaties hebben 19 observaties dus de 'vrijheid' en ligt er dus 1 vast (die je dus niet hoeft te gokken, maar gewoon kan berekenen). In het algemeen zeggen we: een set observaties van n groot heeft n-1 vrijheidsgraden of *degrees of freedom*. Wanneer we de standaardafwijking willen uitrekenen, delen we de kwadratensom door het aantal vrijheidsgraden en dus niet door het aantal observaties! Dit doen we dus als correctie voor die ene gokfout die we dus eigenlijk niet maken. Blijf qua idee wel denken dat we de kwadratensom 'gewoon' door het aantal observaties delen (alleen dus gecorrigeerd voor die ene gokfout die we niet maken). Later in de statistiek komen we complexere berekeningen tegen voor het aantal vrijheidsgraden bij een bepaalde probleemstelling of analyse. Ik zal je er nu niet mee lastig vallen.

Opgaven

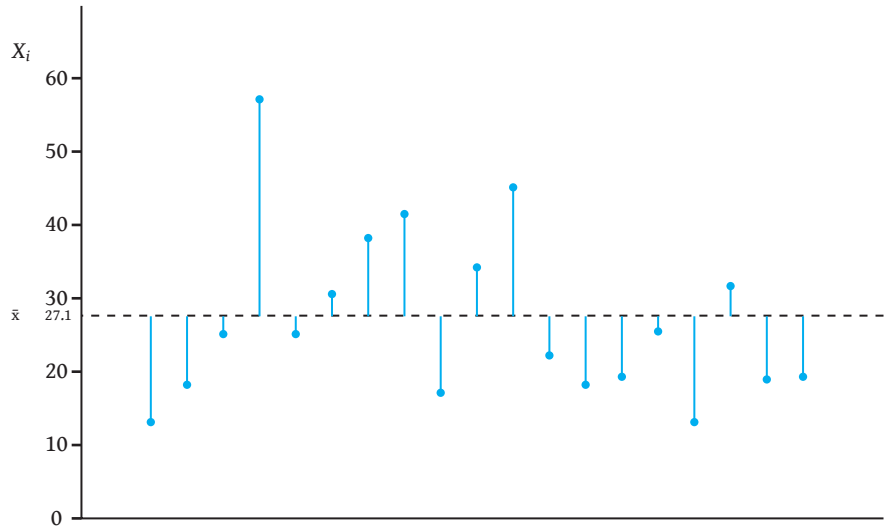
1.5d Deel de kwadratensom door bijbehorend aantal vrijheidsgraden, $df = n - 1$ ('*df*' staat voor *degrees of freedom*). Wat hebben we hier in woorden berekend, wat is hier de officiële benaming voor?

1.5e De variantie en de standaardafwijking zijn verwant aan elkaar. Om van de variantie naar de standaardafwijking te komen, moet je alleen nog de wortel nemen. Als je van de standaardafwijking naar de variantie wil, moet je alleen nog kwadrateren. Bereken de gemiddelde gokfout ofwel de standaardafwijking S_x.

1.5f Hoe groot zou het gemiddelde en de standaardafwijking als *iedereen* in onze steekproef drie minuten hoger had gescoord?

Je kunt je de standaardafwijking ook op een andere manier voorstellen. Als je de scores uitzet zoals in figuur 1.3 hieronder, kun je verticale lijntjes trekken. Elk beginnend van af het gemiddelde (27.1, de horizontale stippel lijn) tot aan de hoogte van een score. De gemiddelde lengte van zo'n lijntje is dan de standaardafwijking (wel gecorrigeerd voor het aantal vrijheidsgraden).

figuur 1C



Uitwerking

1.4a $\bar{X} = 27.1$ zie 1.3a voor uitwerking.

1.4b De geobserveerde scores lopen vanaf 13 (minimum) tot en met 58 (maximum).
spreidingsbreedte = *Max* - *Min* = 58 - 13 = 45

1.4c *individuelegokfout* = $X_i - \bar{X}$
kleinstegokfout = $X_{13} - \bar{X} = 26 - 27.1 = -1.1$
grootstegokfout = $X_{20} - \bar{X} = 58 - 27.1 = 30.1$

Ik heb hier de nummering voor de i-tjes gebruikt van de scores die we al op volgorde hadden gezet.

De grootte van een gokfout hangt dus niet af van of het negatief of positief is, maar wel van de (absolute) afstand naar 0. De kleinste afwijking is dus -1.1 omdat die het dichtst bij 0 ligt. Voor een voorlopige schatting voor de gemiddelde gokfout (standdaardafwijking) kunnen we het beste ergens tussen (in het midden) deze twee gokfouten in gaan zitten, ongeveer 16 dus. Na berekening zullen we zien of we ongeveer goed zaten.

1.5abc

TABEL 1C

i	X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	13	-14.1	198.1
2	14	-13.1	171.61
3	17	-10.1	102.01
4	18	-9.1	82.81
5	18	-9.1	82.81
6	19	-8.1	65.61
7	20	-7.1	50.41
8	20	-7.1	50.41
9	21	-6.1	37.21
10	22	-5.1	26.01
11	25	-2.1	4.41
12	25	-2.1	4.41
13	26	-1.1	1.21
14	31	3.9	15.21
15	33	5.9	34.81
16	35	7.9	62.41
17	39	11.9	141.61
18	42	14.9	222.01
19	46	18.9	357.21
20	58	30.9	954.81
	$\bar{x} = 27.1$	$\Sigma(x_i - \bar{x}) = 0$	$\Sigma(x_i - \bar{x})^2 = 2665.8$

1.5b De optelling van de getalletjes in de derde kolom is 0, de afwijkingen heffen elkaar op! Qua berekening ziet de formule en de uitwerking er als volgt uit:

$$\sum(x_i - \bar{x}) = [(13-27.1) + (14-27.1) + \ldots \ldots + (46-27.1) + (58-27.1)] = 0$$

$$\sum(x_i - \bar{x}) = [(-14.1) + (-13.1) + \ldots \ldots + (18.9) + (30.9)] = 0$$

Je moet dus voor elke observatie of persoon het gehele stuk achter het sommatieteken invullen en uitrekenen, daarna ga je dus pas optellen (sommeren).

1.5c Uiteraard kun je op verschillende manieren uitleggen waar de optelling van wat er in de vierde kolom berekend is, voor staat. In ieder geval is het qua formule:

$\sum(x_i - \bar{x})^2$ Weer zo'n 'ding', maar nu al een stuk complexer dan dat we eerder zijn tegen gekomen. Voordat ik tot benoeming van het ding over ga, eerst maar 'even' de volledige berekening in stapjes:

$$\sum_{i=1}^{i=20}(x_i - \bar{x})^2 = [(13 - 27.1)^2 + (14 - 27.1)^2 + \ldots\ldots + (46 - 27.1)^2 + (58 - 27.1)^2]$$

$$\sum_{i=1}^{i=20}(x_i - \bar{x})^2 = [(-14.1)^2 + (-13.1)^2 + \ldots\ldots + (18.9)^2 + (30.9)^2]$$

$$\sum_{i=1}^{i=20}(x_i - \bar{x})^2 = [198.81 + 171.61 + \ldots\ldots + 357.21 + 954.81] = 2665.8$$

Bij deze, 2665.8 is dus het antwoord, nu nog een benaming. Je kunt zeggen dat de optelling van alle gekwadrateerde gokfouten 2665.8 is. Of iets moeilijker, maar hetzelfde: 2665.8 is de

sommatie van de gekwadrateerde individuele afwijkingen naar het gemiddelde voor i=1 tot en met i=20. Kort weg noemen we dit een kwadratensom (ten opzichte van het gemiddelde) of ook wel in het engels: *the Sum of Squares* (*due to Total*).

1.5d De uitkomst van wat we hier berekenen noemen we de variantie (*variance*) van de variabele X of ook wel S_x^2. Maar wat betekent het? Het is de gemiddelde gekwadrateerde afwijking naar het gemiddelde. Of in meer normalere woorden: de gemiddelde gekwadrateerde gokfout.

$$S_x^2 = \frac{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \cdot \sum_{i=1}^{i=n}(x_i - \bar{x})^2$$

Delen door een getal (19 in ons geval) is hetzelfde als vermenigvuldigen met het omgekeerde ($\frac{1}{19}$).

$$S_x^2 = \frac{2665.8}{20-1} = \frac{2665.8}{19} = \frac{1}{19} \cdot 2665.8 \approx 140.3053 \quad \text{(ik heb hier afgerond)}$$

Blijkbaar is de gemiddelde gekwadrateerde gokfout dus (afgerond) gelijk aan 140.3053. Dit getal is nog niet handig.Veel handiger en makkelijker te interpreteren is de *gewone* gemiddelde gokfout. Omdat de optelling van de gewone afwijkingen 0 was, hadden we ze eerst gekwadrateerd. Er volgt dus nog 1 stap, we zijn er dus bijna!

1.5e $S_x = \sqrt{S_x^2} \approx \sqrt{140.3053} \approx 11.845$

Eindelijk: we kunnen nu dus zeggen dat de mensen in onze steekproef gemiddeld 27.1 minuten bellen (in de afgelopen week welliswaar) en dat de personen in onze steekproef gemiddeld 11.8 minuten daar vandaan zitten (zowel daaronder als daarboven). We hebben nu dus eindelijk een idee in hoeverre de respondenten in onze steekproef verdeeld zijn op de variabele X, ' aantal belminuten in de afgelopen week'. De waarde 11.8 komt redelijk overeen met de eerder geschatte waarde voor de standaardafwijking in opgave 1.4b. Daar had ik ongeveer 16 gezegd. De berekende waarde 11.8 is kleiner omdat er naar verhouding meer kleine gokfouten zitten dan grote, vandaar de iets kleinere waarde.

1.5f Het gemiddelde schuift drie punten op naar boven/rechts.

$$\bar{X}_{\text{nieuw}} = 27.1 + 3 = 30.1$$

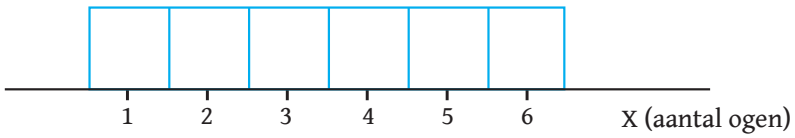
Als *iedereen* qua score hetzelfde verandert, verandert het gemiddelde dus gewoon mee, op dezelfde manier.
Voor de standaardafwijking maakt het niks uit, die blijft gelijk. De individuele gokfouten blijven hetzelfde, de gemiddelde gokfout dus ook. Het enige wat verandert is *waar* iets gebeurt (zich de scores bevinden), maar de onderlinge verschillen tussen de scores niet. Vandaar dat een centrummaat dus wel verandert maar een spreidingsmaat niet.

De normaal verdeling.

2§1 De normaal verdeling, *the normal distribution*

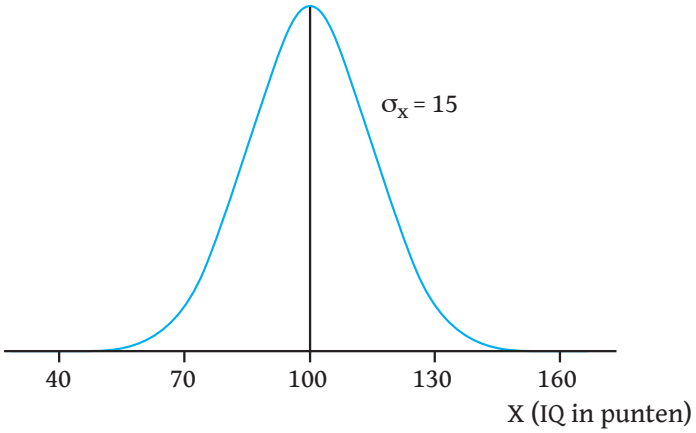
Als je heel vaak met een dobbelsteen gooit, weet je dat je ongeveer even vaak alle mogelijke scores zult gooien, de zes verschilllende 'gebeurtenissen' op een dobbelsteen zullen even vaak voorkomen. Omdat de kans op - of de frequentie van - elke gebeurtenis even groot zal zijn, zeggen we ook wel dat de dobbelsteen 'uniform' verdeeld is. Maar hoe zit dat met scores zoals belminuten, iemands lengte of IQ-scores? Is de kans dat iemand 170 centimeter is net zo groot als de kans op iemand die 210 centimeter is, komen deze gebeurtenissen net zo vaak voor? Zijn er op deze wereld net zoveel reuzen als gemiddelde mensen? Het mag misschien duidelijk zijn dat de scores voor lengte niet uniform verdeeld zijn. De verschillende waarden die je tegenkomt, komen niet allemaal evenveel voor, sommige waarden of observaties kom je vaker tegen dan anderen. Mensen met ongeveer een gemiddelde lengte zijn talrijker dan met mensen met een extreme lengte (reuzen of lilliputters).

figuur 2A



dobbelsteen volgt een uniforme verdeling

figuur 2B



IQ volgt een normaal verdeling

In de statistiek komen we vaak variabelen tegen die zogezegd 'normaal' verdeeld zijn. Als een variabele normaal verdeeld is, 'gedraagt' deze variabele zich volgens een bekend patroon. Zoals bij een dobbelsteen bekend is dat ie zich uniform gedraagt, is bijvoorbeeld van IQ-scores bekend dat deze variabele zich 'normaal' gedraagt met een gemiddelde van 100 (punten) en een standaardafwijking van 15 (punten). Sowieso kunnen we aan de hand van voorgaande paragrafen al stellen dat de beste gok voor een willekeurig persoon 100 IQ-punten is en dat men gemiddeld 15 punten van het gemiddelde verwijderd zit. Maar omdat IQ-scores zich 'normaal' gedragen, kunnen we nog veel meer uitspraken doen. Omdat we weten (aannemen) dat de variabele IQ normaal verdeeld is, kunnen we voor elke willekeurige waarde of score uitrekenen hoeveel procent van de populatie boven of onder die waarde zou moeten vallen. We kunnen dus uitrekenen wat de kans is dat iemand links of rechts van een bepaalde waarde zal vallen. Zoals je bij een dobbelsteen kunt uitrekenen wat de kans is op 3 of 4 gooien, zo kun je nu bij IQ scores uitrekenen wat de kans is dat iemand een IQ heeft tussen de 120 en 130. We mogen deze berekeningen alleen doen als we ook *daadwerkelijk* mogen aannemen dat de variabele normaal verdeeld is. Vaak in onderzoek worden tal van berekeningen gedaan op basis van onterechte 'normaliteits-' aannames en worden dus conclusies getrokken die nooit hadden mogen worden

gedaan. Ik zal je mijn frustraties hierover besparen en zo snel mogelijk verder gaan met een aantal verlichtende opgaven.

Opgaven

2.1a Stel, ik ga naar mars en kom in gesprek met een marsmannetje. Hij heeft nog nooit kennis gemaakt met iets of iemand van de planeet aarde, het enige wat hij wel beheerst, is een gezonde dosis statistiek-kennis. Het enige dat ik hem uitleg, is dat IQ een maat is voor iemands intelligentie en dat een hogere score ook staat voor een hogere intelligentie en dat deze score een normaal verdeling volgt. Verder vertel ik hem dat mijn IQ-score 130 is. Wat kan het marsmannetje nu zeggen over mijn score (in relatie tot de rest van de mensen op aarde)?

2.1b Ok, dat was wel erg weinig informatie. Ik vertel hem nu ook dat het gemiddelde voor IQ-scores 100 is op aarde. Welke informatie voegt het gemiddelde toe? Wat kan het marsmannetje nu zeggen over mijn score in relatie tot de rest van de wereld?

Nu vertel ik hem ook dat de standaardafwijking 15 (IQ-punten) is. Aan de hand van deze informatie (gemiddelde, standaardafwijking, een bepaalde score voor X en de wetenschap dat deze variabele normaal verdeeld is) moet hij in staat zijn om precies uit te zoeken hoeveel procent van de bevolking dommer is en hoeveel procent slimmer is dan ik (met een IQ van 130). Dat is handige informatie, als je precies weet waar ik ten opzichte van de rest sta, dus mijn positie qua IQ-score ten opzichte van de rest van de mogelijke observaties in de wereld. Alleen aan de score 130 hebben we dus niet genoeg. Om uit te zoeken hoeveel procent van de populatie boven of onder mij valt, moeten we eerst de 'ruwe' score ($X_{benjamin}$=130) 'standaardiseren' om vervolgens hierbij een 'overschrijdingskans' op te zoeken. Bij deze heb ik de variabele IQ omgedoopt tot de variabele X. In de komende opgave gaan we over tot standaardisatie van de ruwe score X en gaan we bekijken hoeveel procent van deze wereld nou echt slimmer is dan ik. We gaan het op twee verschillende manieren oplossen. Eerst aan de hand van een aantal vuistregels vooral om inzicht te vergaren wat hier aan de hand is en zodat je in de toekomst altijd een redelijke schatting kun maken voordat je echt gaat rekenen. De tweede manier is aan de hand van de standaard-normaal verdeling (ook wel z-tabel), die ons een preciezere benadering geeft.
Maar voordat we verder gaan, eerst nog even een stukje notatie. We hebben dus een variabele X die normaal verdeeld is met een gemiddelde van 100 en een standaardafwijking van 15. we kunnen dit ook korter noteren:

$X \sim N(100;15)$ Spreek uit als: 'De variabele X volgt een normaal verdeling met een gemiddelde van 100 en een standaardafwijking van 15'.

Of in het algemeen:

$X \sim N(\bar{X}, S_x)$ $\bar{X}$ refereert naar een steekproefgemiddelde en S_x slaat op de steekproef-standdaardafwijking, Dit zijn dus statistieken die je echt zelf kunt berekenen.

Parameters

Eigenlijk omdat ik geen informatie heb gegeven over één enkele steekproef, maar over de gehele populatie, gebruiken we geen statistieken voor het gemiddelde en standaardafwijking, maar parameters. Parameters lijken eigenlijk dezelfde dingen als statistieken, maar met één verschil: ze verwijzen naar waarden die eigenlijk niemand echt zou kunnen weten. Wie weet er nou echt wat het gemiddelde voor IQ is als het gaat om de gehele populatie? Afgezien dat een populatie oneindig groot is, heb je nooit tijd en of geld om iedereen op de gehele wereld op te meten qua IQ-score om dan vervolgens het ware gemiddelde te berekenen. Ik zeg altijd maar zo (als geheugensteun): wij kunnen alleen maar dromen over het ware gemiddelde, laat staan uitrekenen. De Enige Echte die kan weten wat er zich echt afspeelt qua gemiddelde en standaardafwijking wat dus de echte waarde ervan is, is God. Ik als agnost (atheïst durf ik niet

te hard te zeggen) vind dat behoorlijk vaag en wanneer iets vaag is, gebruik je ook vage letters! Vandaar dat we voor parameters ook vage letters gebruiken, Griekse wel te verstaan.

$X \sim N(\mu_x, \sigma_x)$

μ_x, spreek je uit als 'mu x' en staat ook voor het gemiddelde, maar refereert naar het populatie-gemiddelde en dus niet naar een 'berekend' steekproefgemiddelde.
σ_x, spreek je uit als 'sigma x' en staat voor de standaardafwijking van de variabele X, maar dan voor scores van de gehele populatie.

Het onderscheid tussen statistieken en parameters is qua gebruik nu nog niet heel belangrijk, maar ik zal het vanaf nu toch consequent toepassen, zodat je er vast aan kunt wennen.

Opgaven

2.2a Teken een normaal verdeling met een gemiddelde van 100 en een standaardafwijking van 15. Als je vanuit het gemiddelde (het midden van je verdeling) 1 standaardafwijking naar links wandelt, op welke waarde kom je dan uit? En als je 1 standaardafwijking naar rechts wandelt? Volgens de vuistregel vind je altijd ongeveer 68 procent van alle waarnemingen tussen deze twee waarden, dus tussen de waarden die je bereikt, als je 1 standaardafwijking naar links en naar rechts 'wandelt'. Hoeveel procent van de mogelijke observaties valt boven of rechts van de score $X = 115$? Wat is dan dus de kans dat iemand een score heeft hoger of gelijk aan 115? In formule ook wel:

$P(X \geq 115)$ De kans dat de variabele X een waarde aan neemt groter of gelijk 115.
Of ook wel de rechteroverschrijdingskans op $X = 115$.

2.2b Als we niet 1 maar 2 standaardafwijkingen naar links en naar rechts wandelen, hebben we volgens de vuistregel ongeveer 95 procent van de mogelijke waarden te pakken. Op welke waarden komen we uit als we 2 standaardafwijkingen naar links en naar rechts wandelen? Teken ook voor dit probleem een normaal verdeling met de nodige waarden erin. Wat is de kans dat iemand tussen de scores 115 en 130 valt, of ook wel:

$P(115 \leq X \leq 130)$ Spreek uit als: 'de kans op 115 kleiner gelijk X, X kleiner gelijk 130'

2.2c 'Toevallig' had ik gezegd dat mijn IQ gelijk was aan 130. Wat is volgens de vuistregel ongeveer de kans dat iemand in de populatie lager of gelijk scoort? En wat is de kans dat iemand gelijk of hoger scoort? Respectievelijk zijn dit de kansen:

$P(X \leq 130)$ en
$P(X \geq 130)$

2.2d Nu draaien we de vraag om. Hoe slim is ongeveer de slimste persoon van de domste 2,5 procent van de populatie. Besef dat ik hier een kans geef in plaats van (specifieke waarde van) een gebeurtenis voor X waarop ik een overschrijdingskans wil weten. Gek maar waar: dit is de formulering die bij deze vraag hoort.

$P(X \leq x) = 0.025$

Voor welke waarde van x is de kans dat X een waarde aanneemt kleiner of gelijk aan (dan?) die x, gelijk aan 0.025?

2.2e Nog een omgekeerde vraagstelling. Volgens de vuistregel valt 99,7 procent van alle waarnemingen tussen 3 standaardafwijkingen naar links en naar rechts. Voor welke waarde(n)

van x is de kans ongeveer 0.0015 ofwel 0.15 procent dat X een waarde aanneemt groter dan x? De vraag is hier hoe slim de domste persoon (meest linkse waarneming) van de 0.15 procent slimste mensen (de rechterstaart met een oppervlakte van 0.0015).

$$P(X \geq x) = 0.0015$$

In de opgave 2.2 hebben we gekeken naar hoe vaak bepaalde scores binnen een verdeling voorkomen (uitgedrukt in kansen of percentages) aan de hand van de vuistregels. In de volgende opgave gaan we hetzelfde doen, maar dan aan de hand van z-scores en de z-tabel. We kunnen dan ook scores evalueren die niet precies 1, 2 of 3 standaardafwijkingen van het gemiddelde vandaan liggen (de wereld bestaat niet alleen uit mooie ronde getallen).

2.3a Zie de standaardafwijking als een soort liniaal, we hebben in geval van IQ scores te maken met een 'liniaaltje van' 15 IQ-punten lang, onze standaardafwijking voor de variabele X. Om een ruwe score te standaardiseren, is er slechts een vraag die je moet beantwoorden: Hoe vaak past ons liniaaltje tussen het gemiddelde en de ruwe score (waar we een overschrijdingskans op willen weten)? De vraag is dus hoeveel standaardafwijkingen ligt de score 130 (mijn IQ-score) verwijderd van 100 (het gemiddelde). Teken een normaalverdeling met de relevante getallen (100, 15 en 130) en waarschijnlijk zie je het (intuïtief of invullend) vrij snel. Probeer ook te bedenken hoe het sommetje eruit ziet als je de drie getallen in één formule gooit zodanig dat het goede antwoord eruit komt. Vaak als je gemakkelijke getallen gebruikt en je het antwoord dus ook wel gewoon ziet, kun je juist aan de hand van diezelfde getallen ook beredeneren wat de formule had moeten zijn en hem dus ook onthouden!

2.3b We hebben de ruwe score dus omgezet (getransformeerd) naar een nieuw soort score die ons alleen nog vertelt hoeveel standaardafwijkingen de oorspronkelijke score verwijderd is van het gemiddelde. We noemen de getransformeerde score een gestandaardiseerde score of een z-score. We zijn dus van een ruwe 'gebeurtenis' naar een gestandaardiseerde 'gebeurtenis' over gegaan.

$X_{Benjamin} = 130$ ruwe score
$Z_{Benjamin} = 2.00$ gestandaardiseerde score (*standardized score*)

Aan de hand van de berekende z-waarde kunnen we nu opzoeken in de z-tabel wat de kans is dat z een hogere – of een lagere – waarde kan aannemen dan de opgezochte waarde voor Z. We zijn in dit geval dus geïnteresseerd in de *rechter* overschrijdingskans op z is 2.00, omdat ik wilde weten hoeveel procent van de mensheid slimmer is dan ik, dus boven (of gelijk aan) de 130 scoort. De tabel geeft z-waarden (afgerond op 2 decimalen) en geeft de daarbij behorende, *vaste* linker overschrijdingskansen. Zoek de waarde $z = 2.00$ op en kijk welke overschrijdingskans de tabel daarbij geeft. Als we de kans willen weten dat iemand 130 of hoger scoort, wat moeten we dan nog doen met de opgezochte kans?

2.3c Als ik mijn eigen ego een beetje extra zou willen strelen, zou ik dan de z-tabel of de vuistregel beter kunnen gebruiken om de overschrijdingskans te berekenen. Met andere woorden: In hoeverre verschilt de uitkomst van 2.3b op basis van de z-tabel in vergelijking tot de oplossing volgens de vuistregels (opgave 2.2c)?

2.3d Wat is de kans dat iemand een score heeft lager of gelijk aan $X = 125$?

2.3e Hoeveel procent van de bevolking heeft een IQ hoger (of gelijk aan) 125?

2.3f Wat is de kans dat iemand een score voor X heeft lager of gelijk aan 75?

2.3g Wat is de kans dat iemand een score heeft hoger gelijk 75?

2.3h Stel je bent heel lui (lui is goed, daar word je wiskundig van). Wat is het minimale aantal keer dat je de z-tabel zou moeten raadplegen om vraag 2.3d t/m 2.3g op te lossen?

2.4a Een iPod-fabrikant vraagt zich af welk garantietermijn hij zijn klant mee moet geven. Hij onderzoekt zijn iPods en vindt dat de levensduur van zijn iPods een normaal verdeelde score volgt met een gemiddelde van 20000 uur en een standaardafwijking van 2500 uur. Wat is de proportie (of kans) iPods die langer zal meegaan dan 18500 uur?

2.4b Geef de proportie iPods die langer meegaan dan 18500 uur, maar korter dan 23000 uur.

2.4c Hij wil een niet-goed-nieuwe-iPod garantie geven zodanig dat hij slechts 1 procent van zijn verkochte iPods hoeft te vervangen. Welk garantietermijn (levensduur in uren) moet hij op het etiket zetten zodanig dat hij maximaal 1 procent van de iPods hoeft te vervangen?

Groter-, kleiner-, gelijktekens en kansen in een normaalverdeling.

We gaan terug naar IQ-scores. We nemen weer even aan dat deze scores normaal verdeeld zijn met een gemiddelde van 100 en een standaardafwijking van 15. Eigenlijk wil ik weten wat de kans is op een IQ-score van precies 110. Bij een dobbelsteen kunnen we kansen uitrekenen op

2.5a enkele gebeurtenissen. Bijvoorbeeld: Wat is de kans dat je precies drie ogen gooit bij één worp? Bijna iedereen die vaker dan eens gedobbeld heeft, weet dat het juiste antwoord 1/6 is. Tot nu toe hadden we het bij IQ telkens over de kans op een *aantal* scores. De kans op een score hoger of gelijk X=130 bijvoorbeeld, behelst eigenlijk oneindig veel opties, want *alle* scores op het interval vanaf 130 en groter, tellen mee! Dus ook een score met heel veel cijfers achter de komma zoals zoals de score 131.300234. We beginnen even met een klein interval voor IQ-scores nog niet 1 enkele score. Wat is de kans dat iemand tussen de 109 en de 111 scoort?

2.5b Een dobbelsteen heeft 6 mogelijke gebeurtenis en de kans op precies drie ogen bij één worp is 1/6. Hoeveel mogelijkheden zijn er bij de variabele IQ en wat is de kans op precies X = 110?

2.6a Grade Point Average (GPA) zijn scores die zeggen wat het gemiddelde resultaat van een student is, de scores lopen van 0 tot en met 4. GPA volgt in de populatie een normaal verdeling met een gemiddelde $\mu_{GPA} = 2.75$ en $\sigma_{GPA} = 0.500$. Wat is de kans (dus niet percentage) dat een willekeurige student een score tussen de 2.60 en de 2.70 of ook wel:

$$p(2.6 \leq GPA \leq 2.7)$$

2.6b Stel dat er op een onderwijsinstelling 4500 leerlingen zitten, hoeveel van deze leerlingen hebben naar verwachting een GPA-score tussen de 2 en 3?

2.6c Geef de GPA score van de domste leerling van de 10 procent slimste leerlingen op deze onderwijsinstelling.

2.7a Van een universiteit in Nederland is bekend dat voor het eerste statistiek tentamen gemiddeld een 6.7 wordt behaald. Aangenomen mag worden dat de scores normaal verdeeld zijn met een standaardafwijking van 1.4. Jaarlijks doen 950 studenten het tentamen. Hoeveel studenten zullen naar verwachting een score hebben tussen de 5 en een 7?

2.7b Een score van 5.5 betekent een voldoende, hoeveel procent van de studenten slaagt niet voor het tentamen?

2.7c Hoeveel punten zou de universiteit de studenten kado moeten geven zodanig dat slechts

5 procent zou zakken? Hint: Tekenen en bedenk wat het nieuwe gemiddelde zou moeten zijn zodanig dat slechts vijf procent lager scoort dan 5.5. De standaardafwijking voor het tentamencijfer blijft constant en vraag jezelf af (zoek op) hoeveel standaardafwijkingen de score 5.5 van het gemiddelde af moet liggen.

2.8a Al vijf jaar ben ik samen met mijn zogenaamde levenspartner en nu al zegt ze dat ze genoeg van me heeft. Maar hoe serieus moet ik haar nemen? Onder mijn vrienden weet ik dat ongeveer 20 procent het niet langer dan 3 jaar uit houdt met hun relatie. Gelukkig is 30 procent van mijn vrienden hoopgevend en duurt bij hun een relatie langer dan 9 jaar. Ook blijkt dat de duur van een relatie een normaalverdeling volgt. Uiteraard zijn mijn vrienden niet van invloed op mijn of haar keuzes en gedraagt elke relatie zich onafhankelijk, dus voor elke relatie betekent dit: Nieuwe ronde, nieuwe kansen.
Wat zou de gemiddelde duur van een relatie zijn als 20 procent van de mensen het korter uithoudt dan 3 jaar en 20 procent het juist langer dan 9 jaar uithoudt?

2.8b We hebben te maken met twee overschrijdingskansen die niet gelijk zijn. Het gemiddelde zal dus ook niet in het midden van 3 en 9 jaar liggen. Bij welke grenswaarde ligt het gemiddelde dichterbij en welke grenswaarde is dus het minst extreem?

2.8c Bereken de verwachte duur van mijn relatie ofwel de gemiddelde duur van een relatie en de standaardafwijking voor de variabele 'duur van een relatie'. Ga hier dus uit van de eerder gegeven overschrijdingskansen, de genoemde 20 en 30 procent.

Uitwerking

2.1a Alleen aan een ruwe score als informatie hebben we helemaal niks. Niemand kan op basis van slechts deze informatie zeggen of dit een hoge of een lage score is. Wat betekent een score van 130 nou eigelijk? Om wel een antwoord hier op te kunnen geven heb je meer parameters nodig (die dus de rest van de bevolking ook beschrijven).

2.1b Door toevoeging van het gemiddelde en de informatie dat de score normaal verdeeld is, moet het marsmannetje wel weten dat we dus in de rechter helft van de verdeling vallen. Het gemiddelde is het midden en de score 130 bevindt zich daar rechts van. Het moet dus wel een hogere score zijn (ten opzichte van de 'rest'). Maar *hoe* hoog, daar kunnen nu nog geen antwoord opgeven. Scores hebben of krijgen (vaak) alleen betekenis ten opzichte van elkaar, en zelfs dan is het echte verschil niet altijd duidelijk. Een IQ-score van 128 is lager dan 130, ja, zelfs twee punten om precies te zijn, maar om hoeveel hersencellen gaat het? Overigens, een niet onbelangrijk vraag hier: is IQ überhaupt wel te definiëren (in termen van aantal hersencellen)?

2.2a Eigenlijk hebben we de hele verdeling in drieën gehakt, ik spreek hier dus ook wel van een 'driedeling'. Het linkerstuk (of linker staart) loopt van min-oneindig tot en met $X = 85$. Het middenstuk dat precies in het midden ligt, loopt vanaf $X = 85$ tot en met $X = 115$. Het rechterstuk loopt vanaf $X = 115$ tot en met plus oneindig. Het middenstuk beslaat volgens de vuistregel 68 procent van de gehele verdeling. Je kunt ook zeggen dat de oppervlakte onder de curve tussen 85 en 115 gelijk is aan 0.68. In kansen wil dit zeggen:

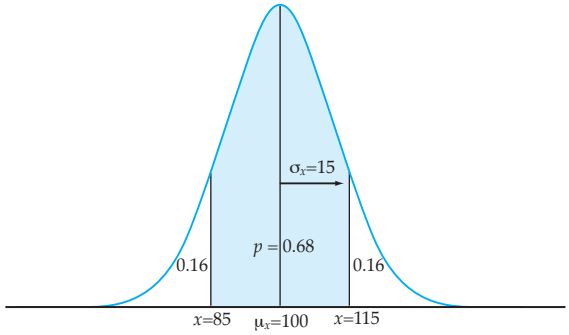
$$P(85 \leq X \leq 115) = 0.68$$

De kans dat X een waarde aanneemt tussen de 85 en de 115 is 0.68 of je kan dus ook zeggen de 68 procent van de bevolking een IQ heeft tussen de 85 en de 115.
Omdat de twee staartjes qua oppervlakte even groot moeten zijn (vanwege de symmetrie in deze verdeling) moeten ze samen de totale oppervlakte min 0.68 hebben. De totale oppervlakte van een verdeling is altijd 1. De twee staarten samen hebben dus een oppervlakte van 0.32

en een enkele staart moet dus wel de helft van 0.32 zijn. De vraag is hoeveel procent van de bevolking 115 of meer scoort.

$$P(X \geq 115) = 0.16$$

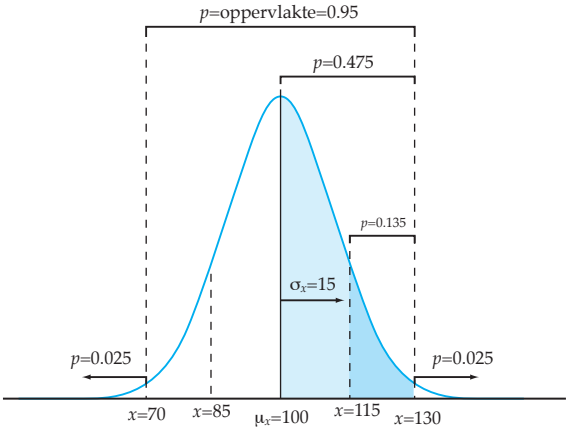
figuur 2C



2.2b Nu kunnen we het beste een 'zes-deling' maken en door optellen en aftrekken, vinden we de juiste oppervlakte. Het stuk vanaf het gemiddelde tot en met X=115 beslaat 34 procent (de helft van 68 procent). Het stuk vanaf het gemiddelde tot en met X=130 beslaat 47.5 procent (de helft van 95 procent). Dus als het om het gevraagde stuk gaat:

$$P(115 \leq X \leq 130) = P(100 \leq X \leq 130) - P(100 \leq X \leq 115) = 0.475 - 0.34 = 0.135$$

figuur 2D



13.5 procent van de bevolking heeft dus een IQ-score tussen de 115 en de 130.

2.2c Kijk alleen even naar de rechter helft van de verdeling, dus vanaf het gemiddelde naar rechts. In de vorige vraag hadden we de rechterhelft in drie stukken gedeeld. Het gaat hier om het laatste stukje, het rechter staartje dus. Vanwege de twee standaardafwijking-vuistregel weten we dat het stuk van 100 tot en met 130 ongeveer 47.5 procent moet beslaan. We houden dus nog slechts 2.5 procent over voor het laatste staartje. Hiermee weten we dus de kans dat iemand hoger scoort dan X=130.

$$P(X \geq 130) = 0.025$$

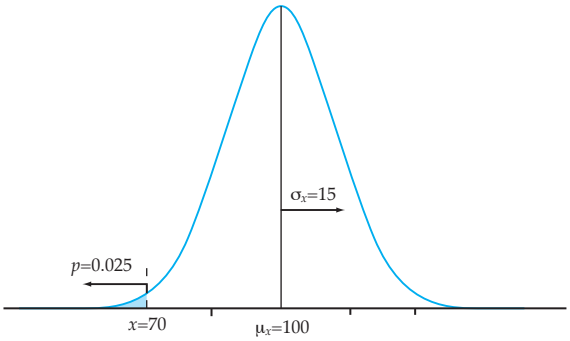
Om de vraag te beantwoorden hoeveel procent van de mensheid lager scoort dan 130, moeten we dus uitvinden hoe groot de oppervlakte onder de curve is links van X=130.

$$P(X \leq 130) = 1 - P(X \geq 130) = 1 - 0.0250 = 0.975$$

97.5 procent van de bevolking scoort 130 of lager.

2.2d We zijn op zoek naar de slimste persoon van de 2.5 procent domste mensen. Waar in de verdeling bevinden zich de domste mensen? Die bevinden zich aan de linkerkant van de verdeling ofwel de linker staart beginnend bij *X = min oneindig* (de allerdomste persoon) tot en met een waarde voor X zodanig dat je 2.5 procent van de bevolking te pakken hebt. Op basis van de vuistregels kunnen we zeggen dat een linkerstaartje van 2.5 procent altijd begint bij min-oneindig en altijd eindigt op twee standaardafwijkingen links van het gemiddelde. De slimste persoon zit precies waar het staartje eindigt, dit is de hoogste waarde voor X van alle mogelijke waarden van X onder het staartje.

figuur 2E



2.2e We zijn op zoek naar de domste persoon van de 0.15 procent slimste mensen. Mijn IQ van 130 is helaas te laag, want dat was juist de ondergrens van de 2.5 procent slimste mensen. 'Mijn staartje' met een oppervlakte van 2.5 is dus groter of langer dan het 'gegeven' staartje van 0.5 procent. Of anders gezegd: mijn score (*X* = 130) is minder ver verwijderd van het gemiddelde en is dus minder extreem dan de gevraagde score, die hoort bij de 'gegeven' overschrijdingskans van 0.15 procent. 99.7 procent van alle mogelijke waarnemingen valt dus tussen drie standaardafwijkingen links en drie standaardafwijkingen rechts van het gemiddelde. Je zou kunnen zeggen dat we nu met een driedeling te maken hebben:

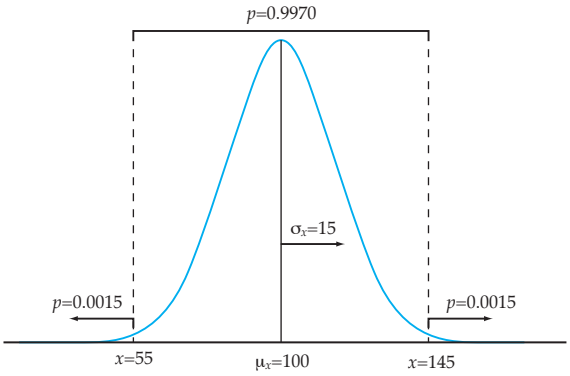
$X \leq 65,\ 65 \leq X \leq 145,\ X \geq 145$ Dit zijn de drie intervallen of stukken (qua scores)

$P(X \leq 65) = 0.0015$ de driedeling qua kansen: 0.0015 - 0.9970 - 0.0015

$P(65 \leq X \leq 145) = 0.9970$

$P(X \geq 145) = 0.0015$

figuur 2F



Toevallig begint het staartje van 0.15 procent (*p*=0.0015) waar wij mee te maken hebben, precies bij *X*=145. Een IQ score van 145 is dus de ondergrens van het staartje die de slimste 0.15 procent van de bevolking representeert.

$P(X \geq x) = 0.0015$

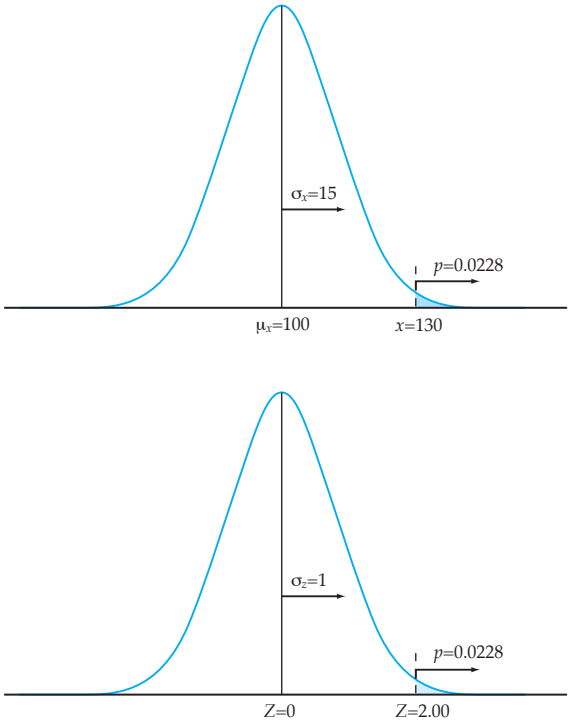
Dit geldt alleen voor x=145. De domste onder de 0.15 procent slimste mensen moet dus ongeveer een IQ hebben van 145 punten.

2.3a $X_{Benjamin} = 130$

$$Z_{Benjamin} = \frac{X_{Benjamin} - \mu_x}{\sigma_x} = \frac{130 - 100}{15} = (130 - 100)/15 = 2.00$$

Berekening in woorden: eerst neem (bereken) je het ruwe verschil tussen de observatie (x=130) en het gemiddelde en daarna deel je het gevonden verschil door de standaardafwijking. Het antwoord geeft je dus het *aantal* standaardafwijkingen dat de waarneming van het gemiddelde verwijderd ligt.

figuur 2G



2.3b De z-tabel geeft alleen linker overschrijdingskansen, dus de kans dat een gebeurtenis of score links van een bepaalde waarde valt:

$$P(Z \leq 2.00) = 0.9772$$

Dit wil dus zeggen dat de kans 0.9772 is dat Z een waarde aanneemt kleiner gelijk 2.00. Deze overschrijdingskans geldt dus ook voor de ruwe waarde (X=130). Omdat de kans die ze geven eigenlijk overeen komt met de oppervlakte (*area*) onder de curve links van Z=2.00, noemen we dit ook wel de '*left tail probability*'.

$$P(Z \leq 2.00) = P(X \leq 130) = 0.9772$$

Met andere woorden: 97.72 procent van de bevolking is dus dommer dan ik! Maar wij wilde eigenlijk weten hoeveel procent slimmer is dan ik ofwel de kans dat iemand 130 of hoger scoort in de populatie. Wanneer je een hele verdeling in tweeën hakt (in ons geval bij Z=2.00 of X=130), hebben de rechter en linker staart samen altijd een oppervlakte van 1. Nu dus nog een stapje om dat laatste op te lossen:

$$P(X \geq 130) = P(z \geq 2.00) = 1 - P(Z \leq 2.00) = 1 - 0.9772 = 0.0228$$

Nu hebben we de oppervlakte van de rechterstaart en deze staat dus voor de kans dat Z een waarde aanneemt groter gelijk 2.00 ofwel de *right tail probability*. Ik kan dus nu zeggen dat slechts 2.28 procent van de bevolking slimmer is dan ik, omdat deze overschrijdingskans dus ook op gaat voor de ruwe score X=130.

2.3c Volgens de vuistregel was de rechter overschrijdingskans gelijk aan 2.5 procent en op basis van de z-tabel 2.28 procent. Mijn ego is gebaat bij zo min mogelijk mensen die slimmer zijn dan ik en dus kies ik voor de z-tabel omdat die overschrijdingskans net iets lager uitvalt.

2.3d De ruwe score (X=120) waarop de linker overschrijdingskans wordt gevraagd, moet eerst worden omgezet naar een gestandaardiseerde score, een z-score.

$$X=125 \qquad Z = \frac{125 - 100}{15} \approx 1.67$$

Nu kun je de bijbehorende overschrijdingskans opzoeken in de z-tabel. In de kolom links vind je de z-scores met het eerste decimaal (de 6 uit 1.67), voor het tweede decimaal (de 7), moet je in de bovenste rij kijken van de tabel. Vervolgens kunnen we zeggen:

$$P(X \leq 125) = P(Z \leq 1.67) = 0.9525$$

De tabel geeft de linker overschrijdingskans. Dus 95,25 procent van de bevolking scoort lager of gelijk aan X=125.

2.3e De (grens)gebeurtenis blijft hetzelfde als bij de vorige vraag (X=125), alleen moeten we nu de rechter overschrijdingskans – eigenlijk percentage – uitrekenen. We hadden de ruwe verdeling in tweeën gehakt bij X=125 of in de gestandaardiseerde verdeling bij Z=1.67. De oppervlakte van het linkerstuk (*left tail probability*) hadden we al gevonden bij de vorige vraag (p=0.9525). Het rechter stuk (*right tail probability*) moet dus wel het resterende gedeelte zijn. In formules:

$$P(X \geq 125) = 1 - P(X \leq 125) = P(Z \geq 1.67) = 1 - P(Z \leq 1.67) = 1 - 0.9525 = 0.0475$$

In de vraag vroeg ik om een percentage, dus moet je ook in percentages antwoorden en niet met een kans komen aandraven. Dus maken we het nog even af: de kans van 0.0475 komt overeen met een percentage van 4.75 procent (de kans dus vermenigvuldigen met 100).

2.3f In plaats van 25 punten rechts van het gemiddelde zitten we nu 25 punten links van het gemiddelde. Er wordt gevraagd naar een linker overschrijdingskans.

$$X=75 \qquad Z = \frac{75 - 100}{15} = (75–100)/15 = –1.67$$

De waarde voor Z is even groot als bij de vorige vraag, alleen nu negatief! We moeten dus zoeken bij de negatieven z-waarden.

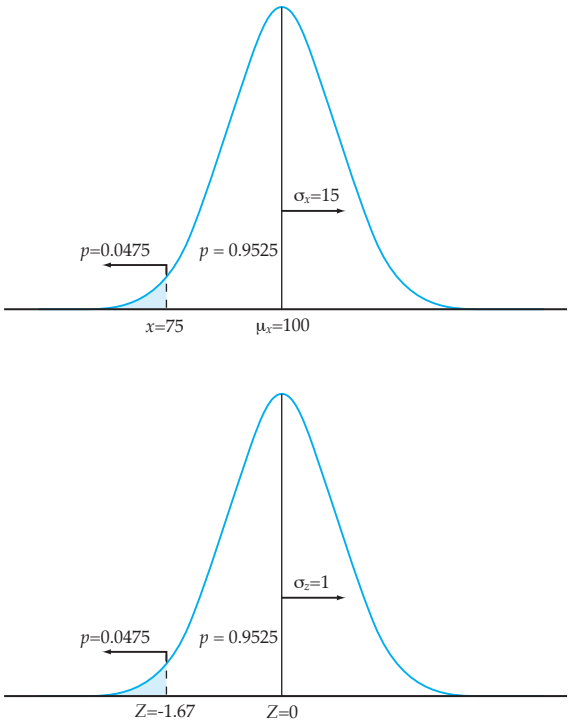
$$P(X \leq 75) = P(Z \leq –1.67) = 0.0475$$

De kans is dus 0.0475 (weer?!) dat iemand lager of gelijk aan een IQ van 75 scoort.

2.3g $$P(X \geq 125) = 1 - P(X \leq 125) = P(Z \geq –1.67) = 1 - P(Z \leq –1.67) = 1 – 0.0475 = 0.9525$$

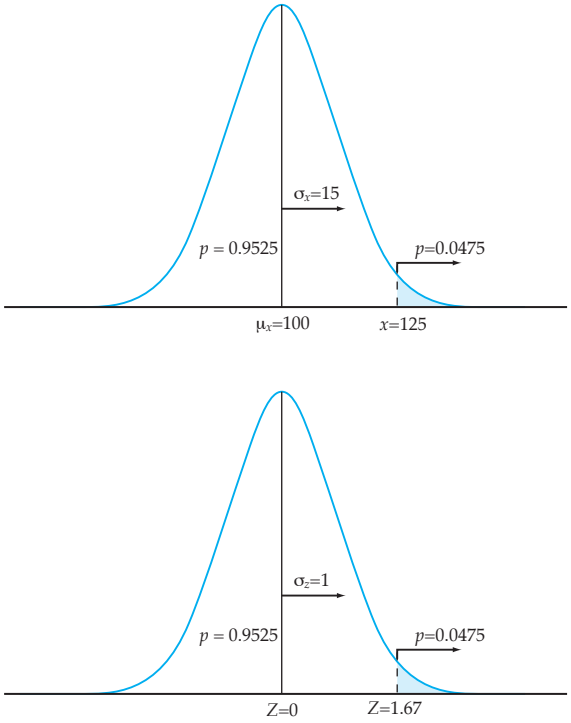
De kans is dus 0.9525 (weer?!) dat iemand hoger of gelijk aan X=75. Zie figuur 2H op de volgende bladzijde.

figuur 2H



2.3h We hadden slechts één keer de tabel te hoeven raadplegen. Elke Z-waarde (in ons geval 1.67) heeft een linker en een rechter overschrijdingskans, de zogenaamde tweedeling. Of de Z-waarde nu positief of negatief is, maakt niet uit voor die tweedeling, het enige wat verandert, is *waar* het grote of het kleine stuk zich bevindt, dus links of rechts van die z-waarde. Als de z-waarde negatief is, zit het kleinere deel van de tweedeling altijd links van die z-waarde en als de z-waarde positief is, zit het kleine deel juist altijd rechts van die z-waarde. Slechts één waarde voor Z leidt tot een tweedeling waar de stukken precies gelijk zijn en dat is de waarde 0. Dit is het geval wanneer de gebeurtenis X dus nul standaardafwijkingen opzij zit van het gemiddelde ofwel dus gelijk is aan het gemiddelde.

figuur 2I



2.4a De variabele levensduur van een iPod is normaal verdeeld met een gemiddelde van 20000 (uur) en een standaardafwijking van 2500 (uur) ofwel:

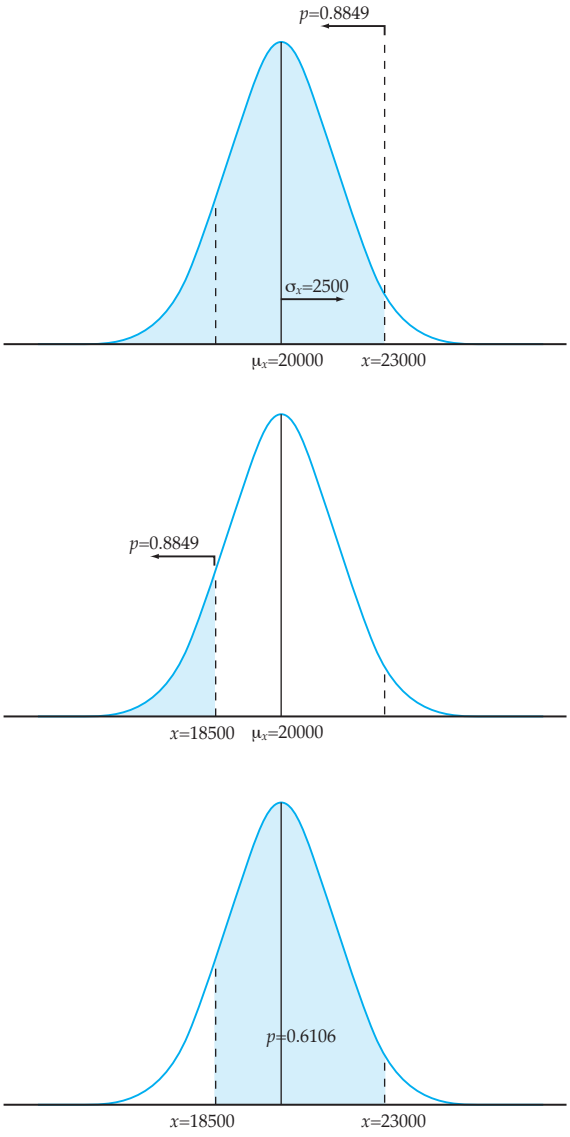
$$X \sim N(20000\ ;\ 2500)$$

De vraag is hoe groot de proportie is die de waarnemingen vanaf 18500 en hoger in beslag nemen ten opzichte van alle waarnemingen. Proporties en kansen zijn eigenlijk gelijk aan elkaar. We willen dus de rechter overschrijdingskans op X=18500.

$$P(X \geq 18500) = P(Z \geq \frac{18500 - 20000}{2500}) = P(Z \geq -0.60)$$

De z-tabel geeft een overschrijdingskans van 0.2743 ofwel de tweedeling 0.2743 tegen 0.7257. Hier werd gevraagd om de rechter overschrijdingskans op Z = –0.60, we hebben dus het grote stuk nodig. De proportie iPods die langer meegaat dan 18500 uur is dus 0.7257 (72,57 Procent).

figuur 2J



2.4b We zijn nu geïnteresseerd in de waarden voor X tussen de 18500 en de 23000. Door voor beide waarde een (linker) overschrijdingskans te berekenen, komen we tot een oplossing.

$$P(X \leq 18500) = P(Z \leq -0.60) = 0.2743$$

$$P(X \leq 23000) = P(Z \leq \frac{23000 - 20000}{2500}) = P(Z \leq 1.20) = 0.8849$$

Als 88.49 procent van de iPods korter meegaat dan 23000 uur en 27.43 procent korter dan 18500 uur, dan moet het wel zijn dat 88.49 - 27.43 = 61.06 procent van de iPods tussen de 18500 uur en 23000 uur meegaat. Vergelijk deze redenering met het volgende: Als 80 procent van een groep kinderen jonger is dan 9 jaar en 20 procent jonger dan 5 jaar dan moet dus 80 - 20 = 60 procent van alle kinderen in die groep tussen de 5 en de 9 jaar zijn. Ons sommetje in formules:

$P(18500 \leq X \leq 23000) = P(X \leq 23000) - P(X \leq 18500)$ Maar dit is ruw, dus:

$P(-0.60 \leq Z \leq 1.20) = P(Z \leq 1.20) - P(Z \leq -0.60) = 0.8849 - 0.2743 = 0.6106$

2.4c De levensduur varieert, sommige iPods gaan korter mee dan andere. Sommige zelfs zo kort dat de klant dus een nieuwe zou willen hebben (niet-goed-nieuwe-iPod garantie). De fabrikant wil hooguit 1 procent van zijn iPods vervangen omdat ze 'te slecht' waren. Als hij dus weet wat de beste iPod is qua levensduur van zijn 1 procent slechtste iPods, kan hij dat als garantie termijn meegeven. Hij zal dan hooguit 1 procent van zijn verkochte iPods terug krijgen en moeten vervangen.
Hier is weer sprake van zo'n omgedraaide vraag zoals bij vraag 2.2d en e. Bij de gewone 'berekeningen' zetten we een ruwe score om naar een z-waarde om vervolgens de overschrijdingskans op te zoeken of even schematisch:

$X \rightarrow Z \rightarrow Probability$

Bij deze vraag beginnen we met de overschrijdingskans (die is gegeven in de opgave) om die vervolgens weer terug te transformeren - omzetten - naar een ruwe score ofwel:

$Probability \rightarrow X \rightarrow Z$

De slechtste 1 procent van de iPods bevinden zich onder de linkerstaart met een oppervlakte van 0.01. Zoek dus in de z-tabel naar 0.01 en kijk welke waarde voor Z daarbij hoort. Omdat 0.01 er net niet precies in staat is het het handigst om te kijken welk getalletje er het dichtstbij in de buurt komt. Twee getalletjes komen in aanmerking. De linker overschrijdingskans van 0.0099 ligt er net iets onder en 0.0102 ligt net iets boven de 0.01 (denk voor het kijk-gemak dat ons staartje een oppervakte heeft van 0.0100, hetzelfde dus, maar dan in vier decimalen zoals de tabel ook geeft). De fabrikant wilde *maximaal* 1 procent vervangen dus laten we vooral niet het grotere staartje nemen, maar die van 0.0099! Lees nu af welke z-waarde bij de linker overschrijdingskans van 0.0099 en als het goed is vind je dan Z=-2.33 want:

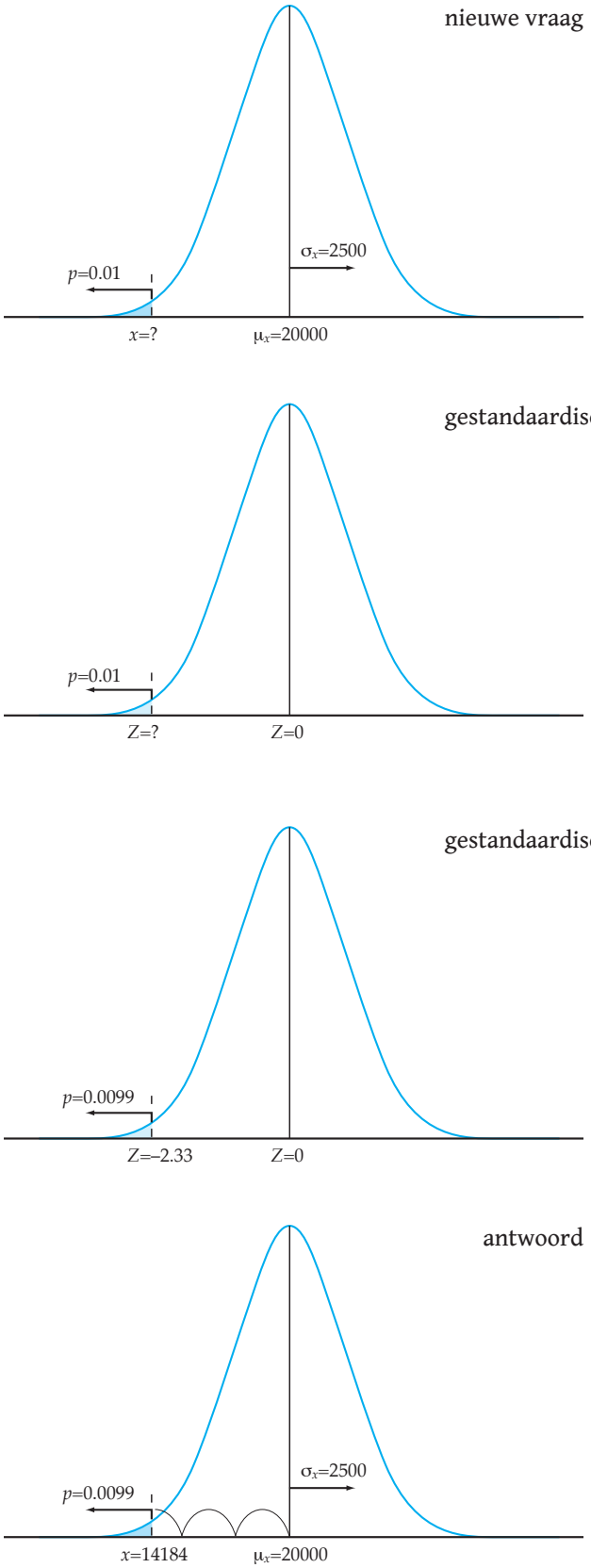
$P(Z \leq -2.33) = 0.0099$

Bij een linker staartje met oppervlakte van 0.0099 hoort dus een z-waarde van -2.33

Dit betekent dus dat de ruwe score *X* dus 2.33 standaardafwijking links van het gemiddelde zit in de ruwe verdeling. Dus we gaan een stukje wandelen vanuit het gemiddelde ($\mu_x = 20000$). Niet 1, niet 2, maar 2.33 standaardafwijking naar links dus.

$X_{garantie} = 20000 - 2.33 \cdot 2500 = 14184$

figuur 2K



Deze berekening kun je beredeneren, zie de standaardafwijking weer als een liniaaltje: je start vanuit het middelpunt van de verdeling (μ_x = 20000) en omdat je in dit geval naar links moet, doe je 'min' het aantal standaardafwijkingen (2.33) keer de lengte van ons liniaaltje (2500). Je legt het liniaaltje dus 2.33 keer naar links vanuit het gemiddelde. Het kan ook moeilijker. Zodra je de z-waarde weet, vul je die in in de z-formule. Je vult ook de andere waarden in die je al

weet en je houdt 1 onbekende over, de waarde die we willen weten, namelijk X:

$$Z = \frac{X - \mu_x}{\sigma_x}$$

de Z-formule, nu nog invullen, onze z-waarde was negatief dus vergeet nu niet het minteken.

$$-2.33 = \frac{X - 20000}{2500}$$

Ik pak hier even een uitgebreidere uitwerking, snellere manieren zijn ook mogelijk uiteraard, maar op deze manier komen we wat basis rekenregels tegen.

$$\frac{X - 20000}{2500} = -2.33$$

$$\frac{X - 20000}{2500} \cdot 2500 = -2.33 \cdot 2500$$ Beide kanten vermenigvuldigen met 2500,

$$X - 20000 = -2.33 \cdot 2500$$ Daardoor vallen ze links allebei weg.

$$X - 20000 + 20000 = -2.33 \cdot 2500 + 20000$$ Aan beide kanten 20000 optellen,

$$X = -2.33 \cdot 2500 + 20000$$ daardoor vallen ze links tegen elkaar weg.

$X = 20000 - 2.33 \cdot 2500$ Omgedraaid mag ook, dan lijkt het op ons 'beredeneerde' sommetje.

$X = 14184$ *solution*

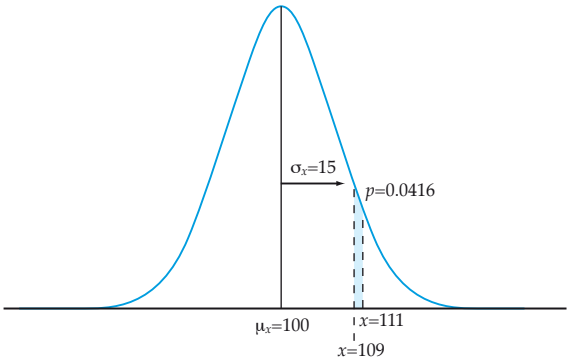
De fabrikant moet zijn klant dus vertellen dat als zijn iPod sneller kapot gaat dan 14184 uur, dat de klant dan een nieuwe kan komen halen.

2.5a $P(109 \leq X \leq 111) = P(X \leq 111) - P(X \leq 109)$

$P(0.60 \leq Z \leq 0.73) = P(Z \leq 0.73) - P(Z \leq 0.60) = 0.7673 - 0.7257 = 0.0416$

De kans is 0.0416 dat iemand 109 of hoger maar lager gelijk 111 scoort.

figuur 2L



2.5b Tot nu toe hebben we telkens de kans op een verzameling scores tussen twee waarden bekeken dus de de kans op scores behorend bij een bepaald interval. Als we de rechter overschrijdingskans op een score van bijvoorbeeld 120 bekijken, heeft die ook een tweede waarde of grens, namelijk plus-oneindig. Het interval met de scores waarop we een kans willen berekenen, loopt dan dus van X=120 tot en met plus oneindig. Eigenlijk is het interval dus oneindig lang aan de rechter kant. We zullen nu de kans op een enkele waarde bekijken. Ik bekijk het op twee verschillende manieren.

Bij een dobbelsteen zou je kunnen zeggen dat er in totaal zes verschillende scores mogelijk zijn. De kans op bijvoorbeeld precies 3 ogen bij één worp, bereken je door het aantal juiste mogelijkheden (in dit geval dus maar 1) door het totale aantal mogelijkheden te delen:

$$P(\textit{precies 3 gooien}) = \frac{\textit{aantal juiste gebeurtenissen}}{\textit{totaal aantal gebeurtenissen}} = \frac{1}{6}$$

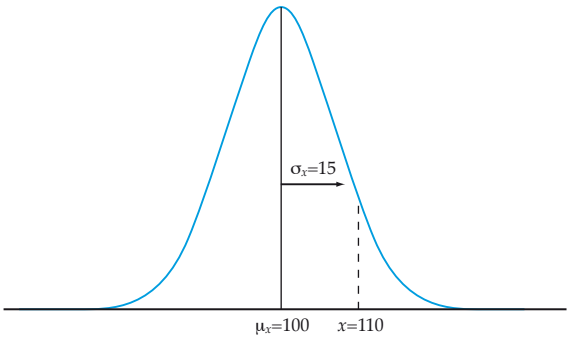
Wij zijn op zoek naar de kans dat X precies 110 is (dat iemand dus een een IQ heeft van precies 110). Er is dus maar een 'juiste' gebeurtenis die bevredigt (voldoet aan onze eis), maar hoeveel gebeurtenissen zijn er in totaal? Oneindig! We krijgen dus

$$P(X = 110) = \frac{\textit{aantal juiste gebeurtenissen}}{\textit{totaal aantal gebeurtenissen}} = \frac{1}{\textit{oneindig groot getal}} = \frac{1}{\infty}$$

Omdat er eigenlijk oneindig veel mogelijkheden zijn, moeten we dus delen door een oneindig groot getal (∞) en als je 1 deelt door oneindig krijg je 0. Anders gezegd is de kans dus gelijk aan 0 dat iemand precies een score van 110 heeft.

We kunnen het ook aan de hand van een normaalverdeling bekijken. Je zou kunnen zeggen dat het interval begint op X=110 en eindigt op X=110 ofwel een lengte van 0 heeft. We zijn op zoek naar de oppervlakte boven het interval en als je in het plaatje hieronder kijkt, zie je dat die oppervlakte eigenlijk alleen uit het verticale lijntje boven X=110 bestaat. En een lijntje is in de wiskunde oneindig dun en heeft dus ook een oppervlakte van 0.

figuur 2M



Bij een dobbelsteen kun je dus wel een kans berekenen maar bij scores die normaal verdeeld zijn dus niet. Of beter gezegd: Het aantal ogen bij het gooien met een dobbelsteen gedraagt zich als een *discrete* variabele wat wil zeggen dat er slechts een aantal mogelijk waarde zijn en als je van één waarde naar de dichtstbij zijnde en mogelijke waarde wil wandelen, moet je springen! Met een dobbelsteen kun je geen 2.5 gooien! Maar als we berekeningen los laten op een variabele als IQ, behandelen we IQ alsof het een continue variabele is. Een continue variabele loopt door, elke waarde is mogelijk, ook een getal met 20 cijfers achter de komma. Toch is het vreemd, want dit alles suggereert dat er niemand zou *kunnen* zijn die precies een score van 110 heeft. In de praktijk is dit natuurlijk wel het geval.

Eigenlijk zijn hier twee punten die ik wil maken.

Als eerste dat de kans op één specifieke waarde van een continue variabele altijd gelijk aan nul is.

$$P(X{=}110) = 0$$

Als tweede, of eigenlijk leidt het eerste punt tot het volgende:

$$P(X{\leq}110) = P(Z{\leq}0.67) = 0.7486$$

$P(X<110) = P(Z<0.67) = 0.7486$

Bij berekeningen met de normaal verdeling maakt het dus niet uit of we een 'kleiner dan'- of een 'kleiner gelijk' teken gebruiken. Dit geld ook voor de vraagstelling. 'Wat is de kans dat iemand *lager dan* 110 scoort?' heeft hetzelfde antwoord (0.7486) als de vraag: 'Wat is de kans dat iemand *lager of gelijk aan* 110 scoort?'. Dit gaat dus niet op voor discrete scores, maar alleen voor continue scores, waar dus alle waarden tussen twee punten mogelijk zijn.

2.6a $p(2.6 \leq GPA \leq 2.7)$

$p(2.6 \leq GPA \leq 2.7) = P(GPA \leq 2.7) - P(GPA \leq 2.6) =$
$P(Z \leq -0.1) - P(Z \leq -0.3) = 0.4602 - 0.3821 = 0.0781$

GR:*normalcdf* = (2.6,2.7,2.75,0.5) = 0.0781

2.6b $P(2 \leq GPA \leq 3) = P(GPA \leq 3) - P(GPA \leq 2) = 0.6247$
Aantal leerlingen $0.6247 \cdot 4500 \approx 2811$

2.6c $P(X \geq x) = 0.1$
$P(Z \geq z) = 0.1$ opzoeken in de z- of t-tabel geeft:
$z = 1.282$
$x = 2.75 + 1.282 \cdot 0.5 = 3.391$
GR:*invNorm* = (0.9,2.75,0.5) = 3.391

2.7a $P(5 \leq X \leq 7) = P(X \leq 7) - P(X \leq 5)$

$P(-1.21 \leq Z \leq 0.21) = P(Z \leq 0.21) - P(Z \leq -1.21) = 0.5832 - 0.1131 = 0.4701$

Dus de kans (of proportie) is dus 0.4701 dat iemand een cijfer tussen de 5 en 7 zal halen, maar er werd om een *aantal* mensen gevraagd:

$0.4701 \cdot 950 = 446.6$ *studenten*

(eigenlijk naar beneden afronden anders zouden er mensen bijzitten die lager dan een 5 of hoger dan een 7 scoren).

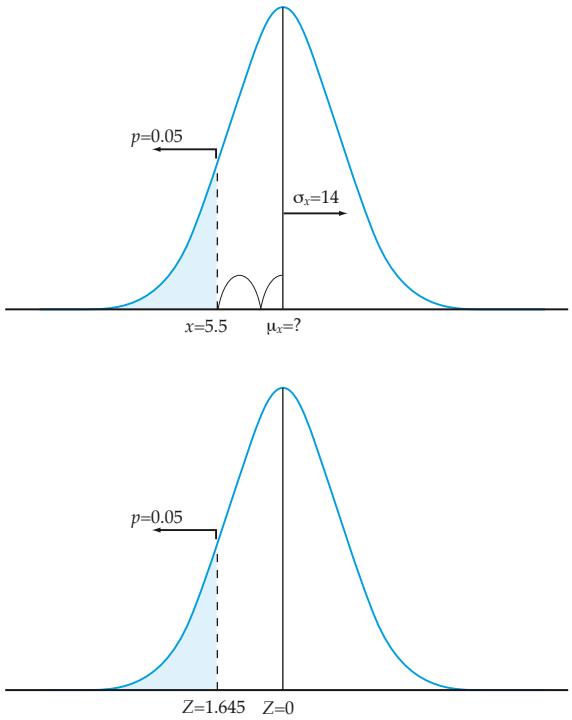
normalcdf(5,7,6.7,1.4) = 0.4725
47,25% van de leerlingen scoort dus tussen de 5 en de 7.
Dit zijn dus 448.875 leerlingen

2.7b $P(X \leq 5.5) = P(Z \leq -0.86) = 0.1949$

Dus 19,5% van de leerlingen heeft een onvoldoende.
normalcdf(-10^99,5.5,6.7,1.4) = 0.1957
19,57% van de leerlingen scoort dus lager dan een 5,5.

2.7c μ of het nieuwe gemiddelde is dus – nu nog – onbekend, maar op of bij de score 5,5 moet de tweedeling voor 5 en 95 procent vallen. 5 procent zou dan links van 5.5 vallen en 95 procent boven de 5.5 en het tentamen dus halen.

figuur 2N



Bij deze tweedeling hoort de vaste z-waarde van ongeveer 1.645 (de z-tabel geeft *Z=1.64* of *Z=1.65*. De waarde 1.645 komt uit de t-tabel. Anders gezegd: vanuit het – nu nog – onbekende gemiddelde zouden we 1.645 standaardafwijking naar links moeten wandelen om op de juiste score uit te komen (X=5.5). Of omgedraaid: We moeten vanuit die X=5.5 1.645 standaardafwijking naar rechts wandelen om bij het nieuwe gemiddelde te komen.

$$\mu_{nieuw} = 5.5 + 1.645 \cdot 1.4 \approx 7.8$$

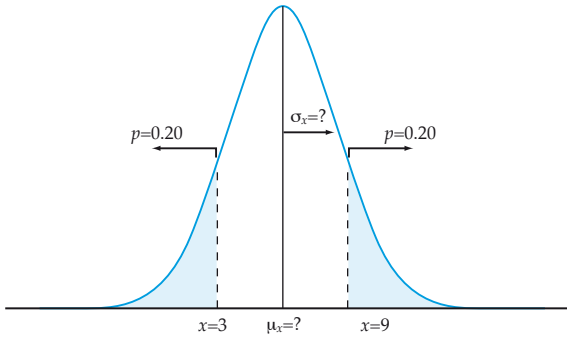
Gezien het oude gemiddelde 6.7 was, moet dus iedereen er (minimaal) 1.1 punt bij krijgen, dan zal (maximaal) 5 procent van de studenten zakken.

(Voor de GR-nerd)
Invoeren y1= normalcdf(-10^99,5.5,X,1.4)
y2=0.05
Optie intersect geeft X=7,8
De universiteit moet de studenten dus 7.8-6.7=1.1 punt kado geven willen als ze willen dat slechts 5 procent zakt.

2.8a Omdat beide staarten van de normaal verdeling even groot zijn, beide hebben een oppervlakte van 0.20, moet het gemiddelde wel precies in het midden liggen als het om een normaal verdeelde variabele gaat. Het midden van 3 en 9 (de twee grenswaarden van de staarten, waarop de overschrijdingskans gegeven is) is ook wel het gemiddelde van die twee:

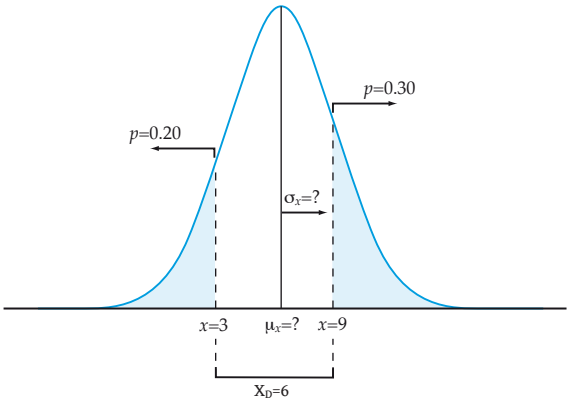
$$\mu_x = \frac{3+9}{2} = 6$$

figuur 2O



2.8b Omdat er een grotere overschrijdingskans is op 9 jaar betekent dat deze waarde makkelijker te overschrijden valt (vanuit het gemiddelde gezien). Dit betekent dat het gemiddelde dichter bij 9 zal liggen dan en dat 9 dus minder extreem is. In het algemeen: Hoe verder een score verwijderd ligt van het gemiddelde, des te extremer is deze score en zal deze score moeilijker te 'bereiken' zijn en heeft daarom een kleinere overschrijdingskans.

figuur 2P

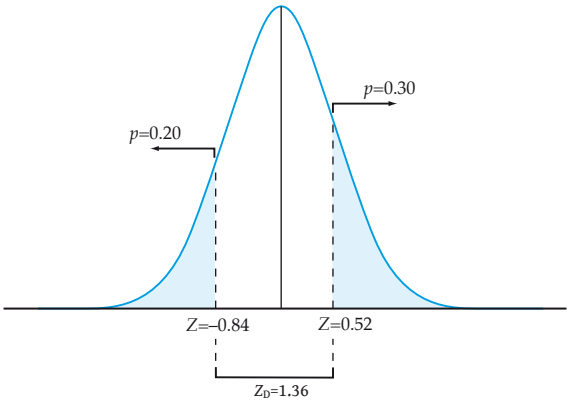


2.8c Eigenlijk is dit een van de allermoeilijkste vragen, met een aantal stapjes die we zullen moeten nemen om het antwoord te vinden. Ook hier zijn meerdere manieren om dit probleem op te lossen. We zullen het zoveel mogelijk visueel proberen op te lossen.

Gegeven in de vraag:
$P(X \leq 3) = 0.20 P(X \geq 9) = 0.30$ (Zie ook figuur bij opgave 2.8b)

We hebben dus alleen twee ruwe x-waarden hun bijbehorende overschrijdingskans. Via de z-verdeling vinden we de oplossing. Omdat we de overschrijdingskans weten op X=9, weten we dus eigenlijk ook hoeveel standaardafwijkingen X=9 verwijderd zou moeten liggen van het gemiddelde. Zo ook voor X=6. Als je de z-verdeling tekent met de twee staarten, kun je de bijbehorende z-waarden opzoeken in de z-tabel en erbij zetten.

figuur 2Q



De rechter overschrijdingskans op X=9 is 0.30, de linker overschrijdingskans op X=9 is dus automatisch 0.70. Opzoeken in tabel geeft Z=0.52 (of 0.53). Voor X=3 vind je ongeveer Z=–0.84. Zie de standaardafwijking weer als een liniaaltje met een (nog on-) bepaalde lengte. Als je op X=3 zou 'staan' en je wandelt naar het gemiddelde, dan moet je dus 0.84 standaardafwijking naar rechts wandelen. Als je dan verder zou wandelen tot X=9, moet je nog 0.52 standaardafwijking verder naar rechts wandelen.

$0.84\sigma_x + 0.52\sigma_x = 1.36\sigma_x$ of

$Z_{difference} = 0.52 - -0.84 = 0.52 + 0.84 = 1.36$

In totaal moet je dus 1.36 standaardafwijkingen wandelen om van het ene punt (X=3) naar het andere punt (X=9) te wandelen. Tot nu toe weten we alleen nog niet hoe groot de standaardafwijking is, maar we weten wel dat als je het ruw bekijkt, dat je van X=3 naar X=9 moet wandelen en dat dat overeenkomt met een wandeling van 6 jaar.

$X_{difference} = 9 - 3 = 6$

Het ruwe stuk of verschil van zes jaar groot, moet overeen komen met het aantal standaardafwijkingen dat afgelegd moet worden tussen de twee punten.

$1.36\sigma_x = 6 jaar$ Een vergelijking met één onbekende! Dus los op!

$\sigma_x = \frac{6\ jaar}{1.36}$

$\sigma_x \approx 4.412\ jaar$ Dus we weten nu de grootte van de standaardafwijking!

Om vanuit X=3 naar het gemiddelde te wandelen moesten we 0.84 standaardafwijking naar rechts wandelen, laten we dat dan ook maar doen:

$\mu_x = 3 + 0.84 \cdot 4.412 \approx 6.71$

Blijkbaar duurt een gemiddelde relatie onder mijn vrienden 6.71 jaar en heb ik nog 1.71 jaar te gaan.

Of een meer wiskundige oplossing, ik gebruik hier de substitutie methode:

$Z = \frac{X_i - \mu_x}{\sigma_x}$ Voor beide grensen zoveel mogelijk invullen.

$-0.84 = \frac{3 - \mu_x}{\sigma_x}$ en $0.52 = \frac{9 - \mu_x}{\sigma_x}$

Ik pak nu een van de twee vergelijkingen en gooi die om (druk μ_x uit in σ_x):

$0.52 = \frac{9 - \mu_x}{\sigma_x}$

$9 - \mu_x = 0.52 \cdot \sigma_x$

$-\mu_x = 0.52 \cdot \sigma_x - 9$

$\mu_x = -0.52 \cdot \sigma_x + 9$

Ik heb nu μ_x uitgedrukt in σ_x en kan dit weer invullen in de andere vergelijking. Ik vervang

(substitueer) dan μ_x voor $-0.52\cdot\sigma_x + 9$.

$$-0.84 = \frac{3 - \mu_x}{\sigma_x}$$

$$-0.84 = \frac{3 - (-0.52\cdot\sigma_x + 9)}{\sigma_x}$$ Een vergelijking met één onbekende, dus op te lossen!

$$-0.84 = \frac{3 + 0.52\cdot\sigma_x - 9)}{\sigma_x}$$ Beide kanten vermenigvuldigen met σ_x :

$$-0.84\cdot\sigma_x = 3 + 0.52\cdot\sigma_x - 9$$

$$-1.36\cdot\sigma_x = -6$$ Dit gaat er al bekend uitzien!

$$\sigma_x = \frac{-6}{-1.36} \approx 4.412\ jaar$$ Eén antwoord.

En voor het gemiddelde nog een keer een vergelijking pakken en zoveel mogelijk invullen:

$$Z = \frac{X_i - \mu_x}{\sigma_x}$$

$$-0.84 = \frac{3 - \mu_x}{4.412}$$ Een vergelijking met één onbekende.

$$3 - \mu_x = -0.84\cdot 4.412$$

$$-\mu_x = -0.84\cdot 4.412 - 3$$ Ziet er ook redelijk bekend uit.

$$-\mu_x \approx -6.71$$

$$\mu_x \approx 6.71$$ Klaar!

Zo zie je maar dat visueel en even denken vaak de voorkeur verdient voor het oplossen van een probleem, het geeft je zelfs al van te voren een idee van wat de uitkomst ongeveer zou moeten zijn. Bij de wiskundige oplossing sluipen en er snel foutjes in en is het vaak heel onduidelijk waar we nu eigenlijk mee bezig zijn.

Samenhang tussen 2 variabelen, correlatie.

Tot zo ver hebben we in hoofdstuk 0, 1 en 2 telkens naar één variabele gekeken (en die dus beschreven aan de hand van een aantal statistieken zoals het gemiddelde en de standaardafwijking). In dit hoofdstuk gaan we een stap verder en kijken we naar het verband tussen twee variabelen. Denk bijvoorbeeld aan het verband tussen motivatie en prestatie, mensen die minder gemotiveerd zijn (dus onder-gemiddeld op motivatie scoren) zullen hoogstwaarschijnlijk ook minder presteren (onder-gemiddeld presteren). En mensen die hoog gemotiveerd zijn, zullen waarschijnlijk ook vaker hoger dan gemiddeld presteren. Hier zijn het dus de twee variabelen motivatie en prestatie die een samenhang (verband) vertonen. Natuurlijk zijn er uitzonderingen op deze uitspraken, zoals die luie (ongemotiveerde) nerd die toch hoog presteert vanwege zijn hoge intelligentie, maar bij *algemene* verwachtingen laat je - natuurlijk - die extreem rare of bijzondere gevallen, even buiten beschouwing.

De begrippen *verband, correlatie, covariatie, covariantie, associatie, samenhang, afhankelijkheid* en *relatie* (tussen twee variabelen) betekenen allemaal hetzelfde.

Het draait natuurlijk allemaal om het voorspellen of verklaren van verschillen (variatie) of waarden, dus natuurlijk ook bij het begrip 'verband'. Het enige verschil is dat we *nu* bij een voorspelling qua waarde voor de ene variabele, rekening houden met een waarde op een andere variabele. Omdat die twee waarden dus 'iets met elkaar te maken hebben' of dus 'samen' 'hangen'. Als je weet dat iemand een *man* (waarde op de ene variabele) is, zal die wel wat *langer* (waarde op de andere variabele) zijn. Dus de voorspelling (of beste gok) qua lengte - als je weet dat iemand een man is - zal hoger (anders) zijn, dan als je weet dat die persoon een vrouw is. Vrouwen zijn over het algemeen (meestal) kleiner. Als we (op grond van de waarden van de ene variabele) alleen maar juiste (dus zonder fouten) voorspellingen doen voor de andere variabele, zeggen we ook wel dat het verband *perfect* is. Maar als het dus *totaal geen* zin heeft om (waarden van -) de ene variabele te gebruiken om de andere te voorspellen, zeggen we dat er *geen* verband is tussen die twee variabelen.

- Sommige zaken zijn nou eenmaal *perfect* aan elkaar gerelateerd, bijvoorbeeld wanneer je temperatuur in graden Celsius omrekent naar temperatuur in graden Fahrenheit)
- Sommige zaken zijn *gedeeltelijk* aan elkaar gerelateerd, bijvoorbeeld de lengte van een persoon en zijn gewicht, lange mensen zijn over het algemeen zwaarder maar er zijn genoeg uitzonderingen.
- sommige zaken zijn in zijn geheel *niet* aan elkaar gerelateerd, zoals de lengte van je schoenveter en de mate waarin je van oliebollen houdt.

Laten we eerst maar even kijken hoe het bij onze aapjes zit, voordat we verder gaan.

TABEL 3A

i	Y_i	X_i
1	120	1.0
2	130	1.0
3	140	1.0
4	140	1.5
5	150	1.5
6	160	1.5
7	160	2.0
8	170	2.0
9	180	2.0

De meest basale gok qua lengte (als één van onze aapjes binnen zou komen wandelen) zou het gemiddelde van de variabele Y zijn (ook wel het nul-model of intercept-model genoemd), het gemiddelde van alle aapjes in onze steekproef wordt ook wel het '*grote* gemiddelde' genoemd. Hier dus $\bar{Y} = 150$. Maar stel nou dat ik je extra informatie zou geven over de leeftijd (X_i) van een aapje dat binnenkomt. Ik vertel je bijvoorbeeld dat het aapje - voordat hij binnenkomt - een leeftijd heeft van 2 jaar. Je mag nog steeds naar alle gegevens kijken die je in de tabel ziet staan. Zou je dan nog steeds zeggen dat de beste gok 150 is? Ik hoop het niet, want als je kijkt naar de drie aapjes die 2 jaar oud zijn, zie je dat die allemaal langer zijn dan 150 cm (namelijk 160, 170 en 180 cm). Anders gezegd: Het gemiddelde qua Y voor alléén de aapjes die de waarde 2 op X hebben, is 170 cm. Ik hoop dat je dus 170 zou gokken als je weet dat een aapje dat binnenkomt, 2 jaar oud is. Als je zou weten dat een aapje 1 jaar is, zou 130 de beste gok of verwachting zijn en als een aapje 1,5 jaar is, dan zou je je voorspelling *niet* aanpassen en gewoon nog steeds 150 cm zeggen.

Het gokspelletje nog een keer, maar nu ook met de 'nieuwe' gokfouten:

- als je dus *niet* weet dat 'toevallig' aapje nummer 9 gaat binnen komen lopen én je zou ook niet weten welke leeftijd dat aapje heeft, zou je dus 150 cm als beste gok geven. Als dan aapje nummer 9 binnenkomt lopen, heb je dus een gokfout van 30 cm gemaakt.
- Als je dus *niet* weet dat 'toevallig' aapje nummer 9 binnenkomt lopen, maar je weet *wel* dat het aapje dat binnenkomt 2 jaar oud is, zou je dus 170 cm als beste gok geven. Aangezien aapje nummer 9 weer 'toevallig' binnen komt lopen, heb je nu dus maar een gokfout van 10 cm gemaakt. Hij ($Y_9 = 180$) zit namelijk 10 cm boven je nieuwe – of *specifiekere* – verwachting (170 cm). Ik zeg hier 'specifieker' omdat het gemiddelde van 170 alleen voor de laatste drie aapjes geldt en dat is specifieker (of minder algemeen) dan het gemiddelde van alle negen aapjes. De score van aapje nummer negen zelf (180) is weer een specifiekere waarde dan het gemiddelde voor de drie aapjes van 2 jaar (170) en daarom is ook de nieuwe afwijking positief omdat hij (180) rechts van dat laatste gemiddelde (170) zit. Besef dat met je specifiekere model (een model of voorspelling op basis van leeftijd) in het geval van aapje nummer 9 dus nog maar een gokfout maakt van 10 cm en niet meer van 30 cm. Het specifiekere model voorspelt bij dit aapje dus beter dan het algemenere model. Zou dat voor ieder aapje gelden? Als het specifiekere model – gemiddeld gezien – beter voorspelt dan het grote gemiddelde, kan je dus zeggen dat het specifiekere model de scores van y beter (terug) verklaart (omdat de gemiddelde gok-fout kleiner worden).

Verband in woorden

Samenhang of verband heel algemeen: Als jou voorspelling (hier qua lengte) afhangt van de waarde op een andere variabele (hier leeftijd) dan zegt men dus dat er sprake is van *samenhang tussen* de twee variabelen. We zeggen dan ook wel dat de twee variabelen *afhankelijk van elkaar* zijn, omdat de informatie over het één (informatie qua leeftijd) iets zegt over de mogelijke informatie van het andere (lengte). Simpel gezegd:

- Naar mate een aapje ouder is, zal die ook wel wat langer zijn. Maar ook:
- Hoe jonger een aapje is, des te kleiner zal die waarschijnlijk zijn. En:
- Als een aapje gemiddeld is qua leeftijd? Dan zal die, qua voorspelling, ook gemiddeld zijn op lengte. Gemiddeld op het één betekent vaak ook gemiddeld op het andere (als variabelen dus samenhangen).

Hoe sterker het verband is, des te zekerder je bent van deze drie uitspraken (en je dus minder uitzonderingen zal tegenkomen zoals een oud 'lilliputter' aapje of een heel jonge 'king kong' aap).

Verband tussen twee variabelen is niet hetzelfde als een oorzaak/gevolg relatie tussen die variabelen.

We weten allemaal dat jonge aapjes langer worden naarmate ze ouder worden. Maar zou je deze uitspraak ook kunnen omdraaien? Dus: Naarmate aapjes langer worden zullen ze ook ouder worden. Ja dat kan! Technisch of statistisch gezien is dit *precies* hetzelfde, maar op de een of andere manier klinkt dit in onze taal toch niet zo lekker, maar het is dus wel goed! Die laatste uitspraak 'klinkt' minder fijn, omdat wij als mensen (naïevelingen dus) - op de een of andere manier - leeftijd als een soort oorzaak (begin, basis of reden) zien van lengte. Maar laten we wel wezen: Natuurlijk is lengte *niet echt* het gevolg van leeftijd. Leeftijd *veroorzaakt geen* lengte. Leeftijd *beïnvloedt* lengte niet. Niemand groeit *vanwege* de tijd, je groeit hooguit *met* de tijd. De echte redenen (oorzaken) dat mens en dier groeit, zijn natuurlijk zaken zoals voedsel, genen, zuurstof en een beetje liefde misschien! Echte oorzaak/gevolg relaties (causale relaties) tussen variabelen zijn heel moeilijk te bewijzen, mocht je dat toch willen doen, zul je een heus experiment moeten doen, en dus de mogelijke 'oorzaak' van groeien, zoals voedsel, weg moeten nemen (Natuurlijk stoppen aapjes met groeien als je ze geen eten geeft!) Denk dan dus ook echt aan een laboratorium gevoel met experimentele en controle condities. Overigens is er een tal van grote denkers (waaronder ik) die zelfs zover gaan dat ze stellen dat causale relaties überhaupt *niet* bestaan. Causatie is dan ook meer een onderwerp voor de filosofie. Als je een opera zangeres heel hard en hoog laat gillen naast een wijnglas dan veroorzaakt het harde en hoge geluid dat het wijnglas breekt. Maar als je tegelijkertijd een hamer op het wijnglas slaat, dan is het dus onduidelijk wie de veroorzaker was van het kapotte glas. Misschien een beetje flauw en gelukkig hoef je hier natuurlijk niet echt over na te denken, maar onthoud voorlopig wel dat causale uitspraken (het één veroorzaakt of *beïnvloedt* het ander) binnen de statistiek veel te voorbarig zijn en dus meestal ongegrond (onbewezen) zijn. Maar we maken dus vaak wel - voor het gemak - een keuze over welke variabele *wij* als oorzaak *willen* zien (die noemen we dan de X variabele) en welke variabele we als gevolg willen zien (de Y variabele).

Correlatie is dus niet hetzelfde als causatie, maar causatie *behelst* of omvat wel correlatie.

Dus als twee variabelen een samenhang vertonen (gecorreleerd zijn), is dat nog niet hetzelfde als een causaal verband tussen de twee variabelen (dat zou dus verder onderzocht moeten worden). Maar als (je weet dat) twee variabelen een causale (oorzaak/gevolg) relatie hebben met elkaar, dan zijn deze twee variabelen - sowieso - gecorreleerd met elkaar. Het bestaan van een correlatie tussen twee variabelen is dus een *noodzakelijke voorwaarde* voor causatie, maar het is nog niet genoeg voor causatie, correlatie is dus geen *voldoende voorwaarde* voor causatie. Dus puur een correlatie tussen twee variabelen is nog niet voldoende bewijs dat die twee variabelen ook echt causaal aan elkaar gerelateerd zijn.

Extreem voorbeeld: Als er op een dag (object, case) veel ijsjes worden gegeten (waarde op de variabele ijs-eten) dan weten we ook dat er op diezelfde dag ook meer mensen zullen verdrinken (waarde op aantal verdrinkingen die dag). De naïeveling (Geert Wilders) zou roepen dat ijs eten dus de oorzaak is van het aantal verdrinkingen. Maar natuurlijk is er geen causale relatie tussen de hoeveelheid gegeten ijs en het aantal verdrinkingen op een dag. De waarden die de twee variabelen aannemen noemen we ook wel een *common response*, een gelijk of algemeen antwoord op iets anders, een derde variabele. Het is alleen het zonnetje (gebrek aan wolken) of de temperatuur waarvan mensen trek krijgen in ijs en natuurlijk zin hebben om te gaan zwemmen (en dus misschien wel verdrinken). Het is dus het zonnetje dat het (gerelateerde) gedrag van beide variabele (of van de corresponderende waarden op beide variabelen) verklaart.

predictor, criterium, respons

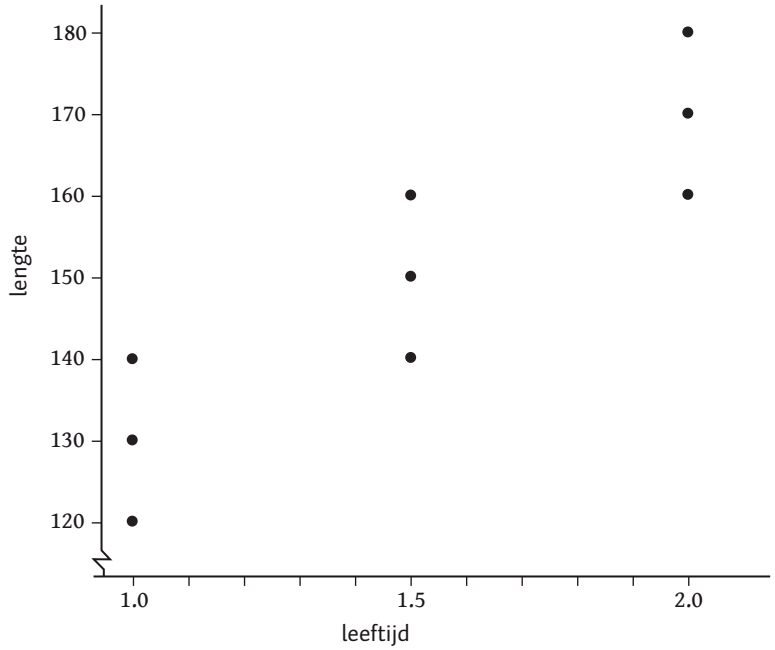
Omdat ík uiteindelijk (in hoofdstuk 5) vooral geïnteresseerd ben in de *verklaring* van de variatie (verschillen) op lengte aan de hand van, of op basis van, leeftijd, heb ik dus voor leeftijd de

variabele X genomen en voor lengte Y. In hoofdstuk 5 gaan we een 'regressiemodel' bouwen om aan de hand van X de variatie op Y te voorspellen (verklaren) en ik heb er dus voor gekozen om leeftijd dus als oorzakelijk te zien (maar het is dus niet *echt* de 'oorzaak') en leeftijd als gevolg (is dus niet *echt* het gevolg). Als ik het hier over oorzaak en gevolg heb, bedoel ik dus alleen de keuze qua richting van voorspelling: van X naar Y. 'Oorzakelijke variabelen' worden vaak ook wel de onafhankelijke of predictor (voorspeller) variabelen genoemd. Voor de 'gevolg variabelen kiezen we dus vaak Y als naampje, maar deze gevolg variabelen worden ook wel, afhankelijke, criterium (bereik), uitkomst of respons (antwoord) variabelen genoemd. Nogmaals: bij het begrip verband boeit de richting van voorspelling dus *totaal* niet en is de voorspel-richting (of volgorde qua benoeming van variabelen) puur een kwestie van wat je leuk vindt of wat handig is voor een lopend verhaal of probleemstelling.

Het verband grafisch weergeven.

Om zaken zoals plaatjes en figuren uit te leggen, denk ik vaak aan een blinde. Dan kan je dus niet een plaatje voor zichzelf laten spreken en leuk met je vingertje wijzen. Hoe zou je een blinde moeten uitleggen wat jij ziet? Wat ik hier eigenlijk mee wil zeggen, is dat als twee mensen naar hetzelfde figuur kijken, wil dat nog niet zeggen dat ze ook daadwerkelijk hetzelfde zien (of de zelfde informatie eruit halen). Ik zal niet elk detail benoemen van de figuur, maar doe wel een kleine of grove poging.

figuur 3A



Als je een mogelijke samenhang tussen twee variabelen grafisch wilt weergeven of bekijken, kunnen we dat dus doen aan de hand van een grafiek. Omdat we slechts kijken naar het verband tussen twee variabelen hebben we maar twee assen nodig, een horizontale en een verticale as, ook wel de X (leeftijd) en de Y (lengte) as. De variabele leeftijd had als laagst voorkomende waarde een 1 jaar en als hoogste waarde twee jaar. Voor lengte was het minimum 120 cm en het maximum 180 cm. We hoeven dus ook niet oneindig lange assen te tekenen, alleen die (lijn) stukken die onze waarden dekken. In een standaard assen-stelsel vind je altijd de waarden 0 voor X en 0 voor Y waar de assen elkaar snijden of kruisen (met een hoek van 90 graden). Hier heb ik dus gesmokkeld omdat de telling niet bij nul begint voor beide variabelen of assen. Omdat je per case (aapje) twee datapunten hebt (een waarde op leeftijd en een waarde op lengte), kun je die twee scores (ook wel coördinaten) tegen elkaar uitzetten. Die puntjes in zo'n grafiek zetten of tekenen noemen we ook wel 'plotten'. Elk punt in het figuur is dus opgebouwd uit twee waarden, een X-waarde en een Y waarde (ook wel de twee componenten-waarden of coördinaten van een punt). De plek of positie van één (of ieder) punt is dus te definiëren door slechts twee coördinaten, de X en de Y coördinaten. De X waarde van één punt vertelt ons hoe

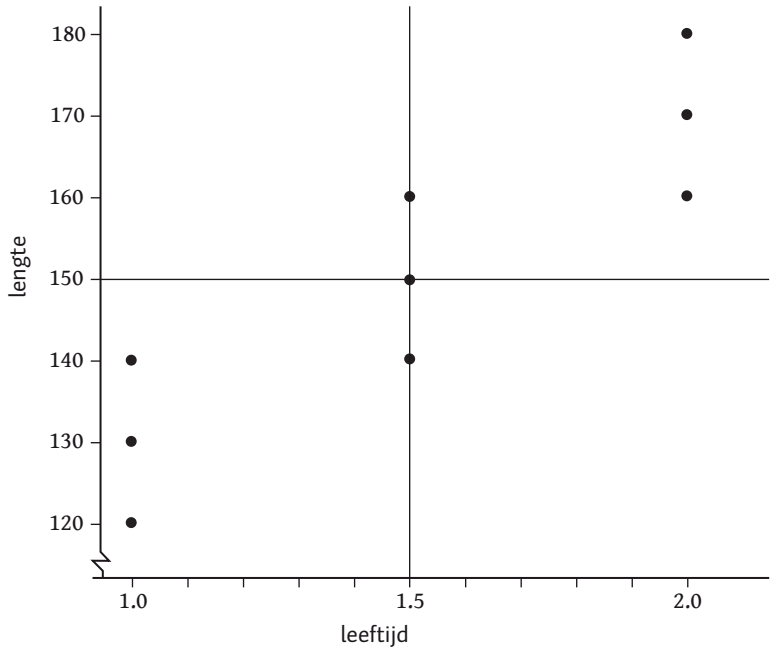
links of rechts het punt ligt (horizontale richting) en de Y-waarde van dat zelfde punt, vertelt ons hoe hoog of hoe laag (verticale richting) dat zelfde puntje zich bevindt. Even moeilijk gezegd, maar probeer het toch te begrijpen, als je van een willekeurig punt wil weten wat zijn X-waarde is projecteer je dat punt op de X-as (je laat hem verticaal omlaag vallen) en kijk je waar dat punt terecht komt op de X-as. Zo ook voor de Y-waarde van een punt: door het punt op de Y-as te projecteren (door het punt horizontaal naar links of naar rechts te duwen, tot dat het puntje op de Y-as ligt), zie je dus de Y-waarde van het punt.

covariantie en *Pearson* correlatie

Als je naar de hele puntenwolk kijkt, zie je dat de wolk omhoog loopt, of beter: de ligging of het verloop van de wolk is vanaf linksonder in de grafiek en gaat dan rechtlijnig schuin omhoog naar rechtsboven in de grafiek. Als een wolk *rechtlijnig* omhoog of omlaag kruipt (en dus juist niet een horizontale ligging heeft), spreken we van een lineair (rechtlijnig) verband. De maat voor een rechtlijnig of lineair verband noemen we ook wel de *Pearson* correlatie. Deze maat (statistiek of parameter) is officieel alleen bedoeld voor variabelen die minimaal een interval meetniveau hebben (of je moet tenminste aannemen dat de variabelen op intervalniveau behandeld kunnen worden) Voorlopig kijken wij alleen naar rechtlijnige verbanden en laten we de kromlijnige verbanden (als een puntenwolk bijvoorbeeld een HEMA-worst vorm heeft) even buiten beschouwing. De sterkte van het verband hangt van twee dingen af: enerzijds de steilheid van de wolk, en anderzijds de dikte of breedte van de wolk. Hoe steiler (omhoog of omlaag) de wolk loopt, des te sterker zal het verband zijn. Als de wolk dunner of smaller wordt (de wolk gaat steeds meer op een rechte lijn lijken), wordt het verband ook sterker. Des te sterker het verband des te meer zal de waarde (getal) van het verband van 0 afwijken. Om het lineaire verband tussen twee (interval) variabelen te beschrijven gebruiken we twee maten, een ruwe maat ook wel de covariantie genoemd (S_{xy} of $cov(xy)$) en een gestandaardiseerde maat: de *Pearson* correlatie, die als statistiek de letter r krijgt, dus voor de variabelen X en Y noemen we hem r_{xy}.

Omdat een verband tussen twee variabelen positief (de puntenwolk kruipt omhoog) of negatief (de wolk kruipt omlaag) willen we vaak weten hoeveel cases bijdragen aan geen, een positief of een negatief verband. Hiervoor heb ik het volgende plaatje gemaakt. je kan dit dus grafisch maar ook numeriek (qua getallen) bekijken. Eerst grafisch.

figuur 3B



Je ziet nu dat ik door het middelpunt van de wolk een horizontale en een verticale lijn heb getrokken. Dat middelpunt is altijd bij het (denkbeeldig) punt ($\bar{X}, \bar{Y}$), bij ons dus het punt (1.5, 150). Bij ons is er toevallig een aapje die voldoet aan die waarden en is het dus ook toevallig een

echt punt in onze wolk, maar dat hoeft dus niet. Ik heb dus het plaatje in vieren gedeeld, ook wel in vier kwadranten. De drie punten in het kwadrant linksonder hebben alledrie gemeen dat ze onder-gemiddeld op X zijn (allemaal 1 jaar) en onder-gemiddeld op Y zijn (120, 130 en 140 cm). Je zou kunnen zeggen dat deze drie cases op beide variabelen in de zelfde richting scoren (laag, laag). En omdat deze drie aapjes op beide variabelen in de *zelfde* richting bewegen (*co*-variëren) zeggen we ook wel dat ze alledrie bijdragen aan een positief verband. Zo ook voor de drie aapjes in het kwadrant rechtsboven. Op beide variabelen scoren de drie aapjes boven gemiddeld, dus de combinatie 'hoog, hoog'. Omdat ook deze drie op de twee variabelen dus in *dezelfde* richting bewegen, zeggen we natuurlijk hier ook dat ze bijdragen aan een positief verband. Dus tot zover hebben we zes aapjes die bijdragen aan een positief verband. Omdat we geen aapjes in het kwadrant linksboven of rechtsonder vinden, draagt er geen één aapje bij aan een negatief verband. Een punt in het kwadrant linksboven zou beteken dat iemand - op de twee variabelen - juist in tegengestelde richting beweegt, laag op X en hoog op Y . In het kwadrant rechtsonder, zou je juist 'hoog op X en laag op Y combinaties' tegenkomen, ook tegengesteld qua richting, dus ook een bijdrage aan een negatief verband. De punten van de middelste drie aapjes liggen allemaal op de verticale lijn (dus hun X waarde is precies gelijk aan het gemiddelde). Eén ervan (nummer 5) is zelfs ook gemiddeld op Y, omdat ie ook op de horizontale lijn ligt. Omdat er dus *geen* sprake is van variatie (beweging) op *beide* variabelen tegelijk (geen sprake van co-variatie dus), dragen deze drie apen dus ook niet bij aan een positief dan wel negatief verband.
Genoeg geouwehoerd, aan de slag met wat berekeningen, we beginnen met de ruwe samenhangsmaat, de 'covariantie' voor X en Y, ook wel S_{xy} of $cov(xy)$ met de formule:

$$S_{xy} = cov(xy) = \frac{1}{n-1} \cdot \sum_{i=1}^{i=n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Berekening in woorden:

- eerste voor iedere case de afwijkingen op X en op Y uitrekenen.
- dan voor iedere case de x-afwijking met de y-afwijking vermenigvuldigen, dus het product nemen.
- dan de producten optellen (in plaats van een *sum of squares* hebben we dus hier een *sum of products*)
- dan delen door het aantal vrijheidsgraden, dus n-1, (of dus vermenigvuldigen met het omgekeerde, zoals ik hem graag zie)

Zoals de variantie S^2 voor een variabele voor het 'gemiddelde kwadraat' (of vierkant) staat, zo staat de covariantie S_{xy} voor twee variabelen voor het gemiddelde product.

Berekening aan de hand van een tabel:

TABEL 3B

i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	1.0	120	-0.5	-30	15
2	1.0	130	-0.5	-20	10
3	1.0	140	-0.5	-10	5
4	1.5	140	0	-10	0
5	1.5	150	0	0	0
6	1.5	160	0	10	0
7	2.0	160	0.5	10	5
8	2.0	170	0.5	20	10
9	2.0	180	0.5	30	15
			$\sum_{i=1}^{i=n} (X_i - \bar{X}) = 0$	$\sum_{i=1}^{i=n} (Y_i - \bar{Y}) = 0$	$\sum_{i=1}^{i=n} (X_i - \bar{X})(Y_i - \bar{Y}) = 60$

Aan de laatste kolom is dus te zien wat de individuele producten zijn en je kan dus aan de waarden zien of ze bijdragen aan geen, een positief of een negatief verband. Hier dragen dus zes apen bij aan een positief verband, drie aan geen verband en nul aapjes aan een negatief verband. Je kunt nu ook meteen zien dat aapje nummer 1 en 9 het meest bijdragen omdat hun producten het meest van 0 afwijken (absoluut gezien het grootst zijn). Deze twee apen liggen dan ook beide het verst van het 'middelpunt' van de puntenwolk vandaan.

Als je de de *sum of products* dus deelt door het aantal vrijheidsgraden, hier 8, ben je klaar:

$$S_{xy} = \frac{60}{8} = 60 / 8 = 7.5$$

$$S_{xy} = \frac{1}{8} \cdot 60 = 1/8 * 60 = 7.5 \qquad \text{(of je vermenigvuldigt dus met } \frac{1}{8}\text{)}$$

De berekening meteen uitgewerkt aan de hand van de formule:

$$S_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{i=n} (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{8} \cdot [(1-1.5)(120-150) + (1-1.5)(130-150) + (1-1.5)(140-150) + (1.5-1.5)(140-150) + (1.5-1.5)(150-150) + (1.5-1.5)(160-150) + (2-1.5)(160-150) + (2-1.5)(170-150) + (2-1.5)(180-150)]$$

$$S_{xy} = \frac{1}{8} \cdot [-.5 \cdot -30 + -.5 \cdot -20 + -.5 \cdot -10 + 0 \cdot -10 + 0 \cdot 0 + 0 \cdot 10 + .5 \cdot 10 + .5 \cdot 20 + .5 \cdot 30]$$

$$S_{xy} = \frac{1}{8} \cdot [15 + 10 + 5 + 0 + 0 + 0 + 5 + 10 + 15]$$

$$S_{xy} = \frac{1}{8} \cdot [60] = \frac{1}{8} \cdot 60 = 1/8 * 60 = 7.5$$

correlatiecoëfficiënt

We hebben dus nu de waarde van de covariantie berekend (S_{xy} = 7.50), maar aangezien de covariantie een ruwe samenhangsmaat is, kunnen we nog niet echt zeggen of het gevonden verband sterk is (behalve dat het verband tussen x en y dus positief is). De eenheden voor leeftijd en lengte waar we nu mee gerekend hebben, stonden respectievelijk in jaren en *centimeters*, maar als je bijvoorbeeld je berekeningen in *meters* en jaren had gedaan, was de uiteindelijke waarde van de covariantie honderd maal zo klein (leuke oefening, transformeer centimeters naar meters en bereken opnieuw de covariantie). De correlatiecoëfficiënt (r_{xy}) is dus een gestandaardiseerde maat voor samenhang die wel in één keer is te interpreteren qua sterkte! Laten we hem eerst maar berekenen. Natuurlijk zijn er tal van manieren (formules) om de correlatie te berekenen, maar we pakken even degene die voor *nu* het snelst werkt, we hebben immers al heel wat statistieken berekend:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Met de gegevens die we eerder hebben berekend:

$$S_{xy} = 7.5 \qquad S_x^2 = .1875 \qquad S_y^2 = 375$$

dus om de standaardafwijkingen voor x en y te krijgen, moet je eerst nog de wortel nemen:

$$r_{xy} = \frac{7.5}{\sqrt{.1875} \cdot \sqrt{375}} = 7.5 / (\sqrt{.1875} \cdot \sqrt{375}) = .89$$

Omdat de berekende waarde voor de correlatie positief van nul afwijkt, moeten we dus zeggen dat er een positief verband is. Omdat de waarde dicht bij 1 ligt kunnen we ook meteen zeggen dat het een 'sterk' verband is (maar wat is sterk?). De correlatie-coëfficiënt kan elke waarden aannemen vanaf -1 tot en met 1, waarbij de uiterste waarden (-1 en 1) betekenen dat het verband perfect is en alle punten precies op één rechte lijn liggen, als je aan de puntenwolk denkt. Bij

een waarde van -1 zeggen we dus dat het verband perfect negatief is en zal de punten-lijn dus omlaag lopen, bij + 1 spreken we dus van een perfect positief verband. Bij een waarde van precies 0 zeggen we dus dat er *geen* verband is. Later, wanneer we regressie-analyses gaan doen, kan ik jullie meer vertellen over de sterkte van het verband en hoe je dus de sterkte van het verband op andere manieren kunt uitdrukken, benoemen of interpreteren (of zelfs voelen). Maar voordat we daaraan beginnen, gaan we in het volgende hoofdstuk, eerst nadenken over de significantie van onze berekende steekproef-resultaten. Het begrip 'significantie' is een veel gebruikte term in de statistiek, de 'significantie' van jouw steekproef-resultaat (statistiek) vertelt je hoe *serieus* je de berekende waarde mag nemen, bijvoorbeeld bij een gemiddelde of een correlatie – of die nou klein of groot is – . Door de 'significantie' van jou statistiek, weet je of je die waarde mag generaliseren. Als je generaliseert, maak je van een uitspraak (of stelling), specifiek bedoeld voor één groep, een uitspraak die voor een *grotere* groep bedoeld is. Bijvoorbeeld dat de correlatie tussen x en y in onze steekproef (r_{xy} = .89) ook geldt voor de gehele populatie aapjes. Niet alles wat je binnen je gewone leventje tegenkomt, neem je serieus en zal je *meteen* als *altijd waar* aannemen. In zo'n geval zeg je vaak 'Oh, dat is toeval' en 'Normaal gaat het anders', bijvoorbeeld als je een keer te laat komt (niet doen, geen excuus geven). Zo dus ook bij steekproef-trekkingen: Niet elk gevonden verschil of effect neem je serieus. Je gaat niet zomaar roepen dat onze aapjes – van nu - langer zijn dan aapjes van 50 jaar geleden (die waren toen gemiddeld 135 cm namelijk), tenzij je genoeg vertrouwen hebt dat je, heel vaak, tot diezelfde conclusie zou komen, als je *nog een keer* een steekproef zou trekken en dan het *liefst* wel uit die zelfde populatie.

En nog een keer, in eindeloze herhalingen.

Van Steekproef naar Populatie.

Generalisatie van statistieken naar parameters aan de hand van betrouwbaarheidsintervallen en significantie-toetsen.

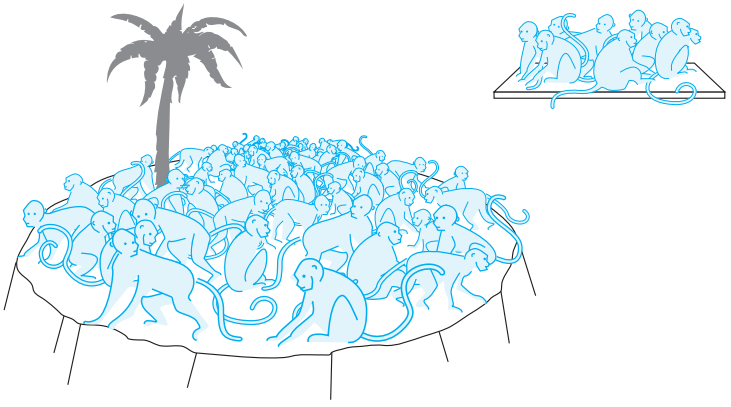
4§1 Populaties kun je niet meten, steekproeven wel.

Je hoort wel is zeggen: 'Meten is weten'. Misschien is het leuk om te weten wat er in je steekproef gebeurt, maar dat is niet wat we willen weten of waar we uiteindelijk uitspraken over doen. Het *doel* van (inferentieële) statistiek is om uiteindelijk een uitspraak te doen over wat er gebeurt in de populatie. Elke keer wanneer je een steekproef uit een populatie trekt, is het maar weer de vraag, in hoeverre jouw specifieke waarden (van je berekende statistieken) zoals het gemiddelde, overeenkomen met de echte waarden (zoals ze in de populatie zijn). Ook al gebruiken we een statistiek als puntschatting voor een parameter, wil dat nog niet zeggen dat deze twee *precies* dezelfde waarde zullen hebben. Laat alsjeblieft één ding duidelijk zijn: Niemand weet precies wat de waarde is van het gemiddelde van een populatie, zoals μ_y (behalve een *alziend* oog zoals die van een echte God zou daar zicht op hebben). Niemand heeft ooit de tijd, geld of zin gehad om een hele populatie aapjes (dat zijn *oneindig* veel aapjes) op te meten om het ware gemiddelde (qua lengte bijvoorbeeld) van de gehele populatie te achterhalen. En toch is dat het doel van onderzoek en statistiek, iets zeggen over populatie-parameters. We kunnen geen hele populatie (alle scores van die populatie) aanschouwen en daaraan rekenen, maar toch zouden we dat dus wel willen. Omdat we alleen dan honderd procent zekerheid zouden hebben over de uitspraken die we doen (over de waarden van populatie-parameters). Helaas, absolute zekerheid is (ook) in de statistiek dus niet te vinden, maar met iets minder zekerheid – zeg met 95 procent zekerheid - kun je ook bakken met geld verdienen (of de wereld verfijnen en gezelliger maken)! Wat nu komen gaat, namelijk de generalisatie van een gevonden steekproef-gemiddelde, doen we enkel en alleen om de volgende onderzoeksvraag te beantwoorden:

Wat zou de waarde van een populatie gemiddelde zijn?

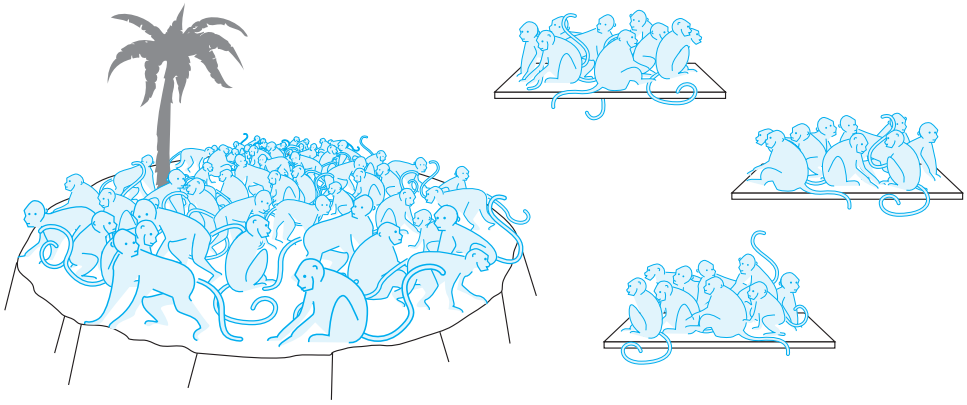
4§2 De steekproeven-verdeling.

Terug naar onze aapjes, wij hebben slechts één *echte* steekproef genomen (met n = 9 aapjes) en daar rolde een gemiddelde uit van $\bar{y}$ = 150 en een standaardafwijking (S_y = 19.36). Omdat we *geen* andere steekproeven hebben genomen, is deze steekproef de enige informatiebron over (9) mogelijke data-punten (die uit de populatie komen) en zullen we het dus hiermee moeten doen! Hiervoor gaan we nog een keer het voorspél-spelletje doen, maar nu gaan we geen individuele data-punten voorspellen, maar (steekproef-) statistieken! Om dit nieuwe spelletje te spelen, moet je je afvragen wat er *zou* gebeuren als je meer steekproeven zou nemen.



Stel dat je de lengte-score van *ieder* aapje uit de populatie zou hebben en al die aapjes (of hun scores, maar aapjes leek me gezelliger) in één grote grabbelton zou doen (of denk aan een heel mooi eilandje met alle mogelijke aapjes uit de populatie). Vervolgens trek je - random - 9 aapjes uit die ton en je berekent het gemiddelde over hun scores (jouw steekproef-statistiek). Je

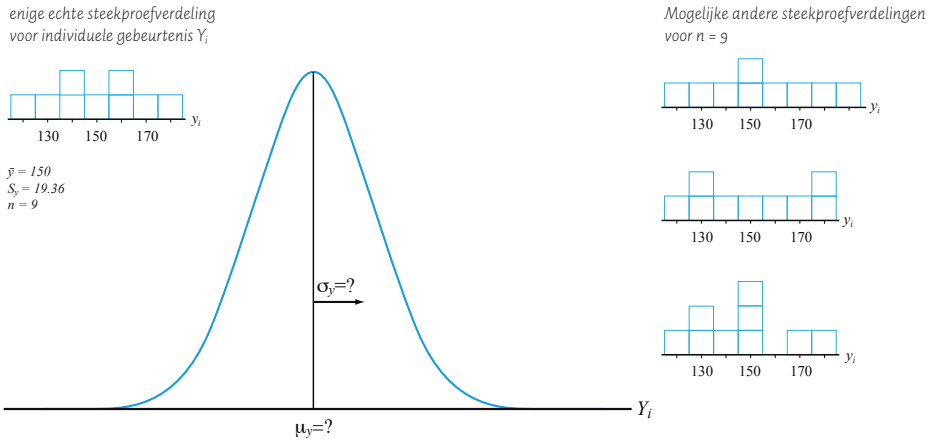
hebt dan dus een random (iedere aapje uit die grabbelton of populatie had een gelijke kans om getrokken te worden, heel eerlijk dus) steekproef getrokken. Eigenlijk hebben we dat dus al één keer gedaan met onze eigen, 9, enige echte aapjes. En bij ons kwam er 'toevallig' 150 cm uit qua gemiddelde (statistiek).



Maar stel je voor dat we dit proces van steekproeftrekking zouden herhalen, dus negen *nieuwe* aapjes uit de grabbelton trekken en opnieuw het nieuwe steekproef-gemiddelde zouden berekenen. Wat is op dit moment jou beste gok qua nieuw gemiddelde? Je weet wat het gemiddelde en spreiding was van onze eerste steekproef en dat is alles wat je tot zover weet, dus is dat ook de enige informatie, die je überhaupt kan gebruiken, om vast een voorspelling te doen over een nieuwe steekproefgemiddelde.

Als je een nieuwe steekproef zou trekken, zou je hopelijk verwachten (of gokken) dat het nieuwe steekproef gemiddelde óók waarschijnlijk ergens rond 150 cm ligt. Natuurlijk, het kan *lager*, ongeveer *gelijk* of *hoger* zijn dan je eerste steekproef-gemiddelde. Hoeveel lager of hoger? Het kan natuurlijk zijn dat als je een nieuwe steekproef trekt, je - *heel toevallig* - de negen grootste apen uit de grabbel-ton trekt en dus een heel hoog gemiddelde krijgt. Maar hoe groot is die kans? Als we *aannemen* dat de populatie normaal verdeeld is qua lengte-score, betekent dat, dat er veel meer aapjes rond het gemiddelde zitten (denk aan vorm van de normaal-verdeling) en maar relatief weinig aapjes, zullen zich qua lengte, ver weg van het gemiddelde bevinden. De kans is natuurlijk uiterst klein dat je - precies - de negen grootste apen in je steekproef treft. Zeg maar gerust dat die kans één op oneindig veel is. Stel dat je niet alleen een tweede steekproef uit je ton pakt, maar door blijft gaan tot in oneindigheid (gelukkig hebben we computers die dit soort langdradige spelletjes voor ons kunnen doen). Hoogstwaarschijnlijk zullen de meeste steekproef-gemiddelden (statistieken) dan rond de 150 cm liggen en relatief minder gemiddelden daar ver vandaan. Nu zijn we *slechts* begonnen met 9 aapjes en dat aantal hebben we *telkens* herhaald. Maar als je ditzelfde spelletje zou doen, maar met een grotere steekproef, zeg n = 20, wat verwacht je dan? Het steekproef-gemiddelde zal bij nieuwe trekkingen (van n = 20) natuurlijk ook variëren, maar wel (veel) minder. Wat moeilijker gezegd, zeggen we ook wel dat het steekproef-gemiddelde zal fluctueren bij herhaling. De mate van deze fluctuatie (voor het gemiddelde) hangt onder andere af van de steekproef-grootte (n) en de spreiding van de individuele scores in de populatie (σ). Als er veel spreiding (grote individuele verschillen) is een populatie, dan zal bij herhaaldelijk steekproef trekken, het gemiddelde ook meer - of heftiger - fluctueren. Maar als je de steekproef-grootte *n* vergroot, zal het gemiddelde zich weer stabieler gaan gedragen en dus minder fluctueren. Sterker nog: we hebben daar natuurlijk berekeningen voor zodat we van te voren kunnen bepalen wat er zou gebeuren (qua spreiding van het gemiddelde) als we dit langdradige spelletje echt zouden spelen.

figuur 4A



Populatieverdeling voor individuele gebeurtenis Y_i

De steekproeven-verdeling van het gemiddelde, *sampling distribution of the mean.*

Omdat we maar één echte steekproef hebben genomen en daar de twee statistieken ($\bar{y}$ = 150, S_y = 19.36 met n = 9) uitrolde en we dus echt geen andere info kunnen gebruiken, zeggen we dus ook wel dat beide statistieken de enige en dus de beste *punt-schatters* zijn voor het populatie-gemiddelde, μ_y en standaardafwijking σ_y. Natuurlijk hoeven deze schattingen niet precies te kloppen (en dat zal ook wel niet), dus we houden rekening met deze onzekerheid. Hoe groter de steekproef, hoe zekerder je bent dat je schattingen (ongeveer) kloppen. Denk maar eens aan een extreem grote steekproef (bijvoorbeeld een steekproef van 99 procent van je gehele populatie). Het zou wel heel raar zijn dat die laatste procent (die je dus niet in je steekproef hebt) aapjes totaal anders zouden zijn dan je steekproef. Wat hier een hele dikke vinger in de pap heeft, is ook wel de wet van heel grote *aantallen* (soms zegt men 'getallen'). Denk maar eens aan een gewone dobbelsteen. Als je die 1 miljoen keer zou gooien, wat zou dan de waarde zijn van de gemiddelde worp? Hoe groot is dan de kans dat je gemiddeld, precies 1 zou gooien? Om precies een gemiddelde van 1.00000 (ja, op 6 decimalen nauwkeurig) te krijgen, moet je dus elke keer weer, dus een miljoen keer, precies een 1 gooien. Ik zou mijn geld er niet op in zetten. Liever gok ik op 3.5 (het gemiddelde van een dobbelsteen als je heel vaak gooit, precies in het midden van de waarden 1 en 6). Overigens is de kans dat je precies 1 miljoen keer een 1 gooit ook wel:

$$p(precies\ 1\ miljoen\ keer\ een\ 1) = (\frac{1}{6})^{1000000}$$

Behoorlijk klein dus (je mag het zelf in-typen, hoogstwaarschijnlijk raakt je rekenmachientje ervan in de war of rond het af op 0)

Het blijkt (uit wiskundige bewijzen en computer simulaties) dat we precies kunnen berekenen hoe een steekproef-gemiddelde (of andere statistieken) zal variëren bij herhaaldelijk steekproef trekken. Als we weten hoe een variabele (een statistiek is ook een variabele) verdeeld is, kunnen we bijvoorbeeld uitrekenen wat de kans is op een range van mogelijke waarden voor een statistiek. Je kunt dan (onderzoeks-) vragen beantwoorden zoals: 'Wat is de kans dat het steekproef-gemiddelde hoger zal zijn dan bijv. 160 (bij herhaaldelijke steekproeftrekkingen), gegeven dat we in onze eerste en enige steekproef, een gemiddelde vonden van 150.' Dus we kunnen uitspraken doen over mogelijke waarden die kunnen optreden (als we dus nog een steekproef zouden trekken) en met welke kans deze *range* van gebeurtenissen zullen optreden. We vinden deze kansen door gebruik te maken van de steekproeven-verdeling van het gemiddelde (denk dus herhaaldelijk steekproef trekken). Laten we maar gaan knallen en tot echte uitspraken komen. Als we *aannemen* (dat doen we dus maar even) dat de lengte-scores in populatie een normaal-verdeling volgen en we vervolgens meerdere steekproeven zouden trekken (en dus opnieuw het gemiddelde uitrekenen), dan zullen *al* die steekproef-gemiddelden, ook normaal verdeeld zijn. *Elke* normaal verdeling heeft een verwachting (het gemiddelde) en een standaardafwijking. De *standaardafwijking* voor een *statistiek* zoals het gemiddelde ($\bar{y}$, bij

herhaling dus) geven we een nieuwe naam, namelijk de 'standaardfout' of '*standard error*' voor het gemiddelde. Denk dus - nog steeds - de gemiddelde gokfout wanneer je het spelletje speelt, maar nu voorspel je dus de waarde van een statistiek. Een *standard error* geven we aan met de afkorting of symbool '*SE*'. Omdat wij de standaardfout voor het gemiddelde nodig hebben, noemen we hem dus $SE_{\bar{y}}$. Zijn waarde kan berekent worden met de volgende formule:

$$SE_{\bar{y}} = \frac{s_y}{\sqrt{n}}$$

Aangezien we de waarde voor S_y (19.36) al eerder zijn tegen gekomen en de steekproefgrootte 9 is, kunnen we de standaard error dus uitrekenen:

$$SE_{\bar{y}} = \frac{19.36}{\sqrt{9}} = 19.36 \,/\, \sqrt{9} \approx 6.45$$

Nu we dus weten wat de waarde voor de *standard error* is (6.45) en aannemen dat ook het steekproef-gemiddelde ($\bar{y}$) zich – *een soort van* - volgens de normaal-verdeling gedraagt, kunnen we vast wat grove uitspraken doen over de mogelijke waarden van het steekproef-gemiddelde die zich zullen voordoen als we nog veel meer steekproeven zouden nemen, laat het - voor het gemak – even duizend steekproeven zijn. Volgens de *vuistregels* voor standaardafwijkingen (in ons geval dus *standard errors*) zou je ongeveer 95 procent van alle mogelijke gemiddelden tussen twee specieke waarden vinden, de laagste waarde (of ondergrens) vind je door twee keer de *standaard error* van het steekproef-gemiddelde af te halen en de hoogst mogelijke waarde, door twee keer de *standard error* bij het steekproef-gemiddelde, op te tellen:

$$Lower\ Bound = \bar{y} - 2 \cdot SE_{\bar{y}}$$

$$Upper\ Bound = \bar{y} + 2 \cdot SE_{\bar{y}}$$

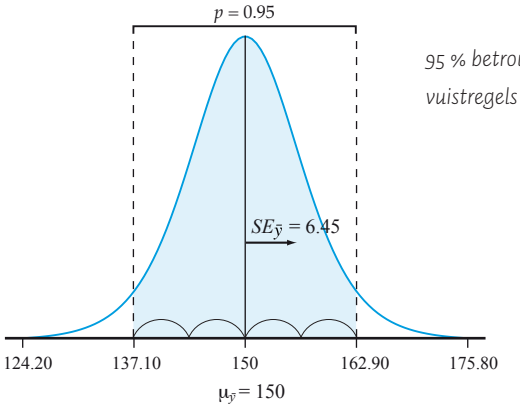
in ons geval dus:

$$Lower\ Bound = 150 - 2*6.45 = 137.10$$ (alsjeblieft, geen haakjes gebruiken hier)

$$Upper\ Bound = 150 + 2*6.45 = 162.90$$

We hebben dus nu een betrouwbaarheidsinterval berekent (dat dus loopt van de *lower* naar *upperbound*). Je mag dus nu zeggen, dat als je duizend steekproeven neemt (na onze enige echte), 950 van die steekproef-gemiddelden (95 procent), een waarde zullen hebben tussen de 137.10 en 162.90. Maar ook dat 25 (2.5 procent) steekproef-gemiddelden, een waarde zullen hebben onder de 137.10 en dat dus 2.5 procent een waarde zal hebben die boven de bovengrens (162.90) ligt. Op grond van deze bevindingen mogen we ook een uitspraak doen over het ware gemiddelde van de gehele populatie. Hèhè, Eindelijk!

figuur 4B



95 % betrouwbaarheidsinterval op basis van vuistregels voor standaardafwijkingen

We *zeggen* dan ook wel dat het ware gemiddelde van de populatie (μ_y) met 95 procent zekerheid ergens tussen 137.10 en de 162.90 ligt. Officieel (maar niet belangrijk voor nu) moet je eigenlijk zeggen dat als we *echt* honderd keer een steekproef zouden nemen en *echt* honderd keer de *lower* en *upperbound* zouden berekenen (en dus elke keer iets andere waarden zullen vinden) dat 95 procent van die betrouwbaarheidsintervallen het ware gemiddelde behelst. Maar goed, laten we het vooral praktisch houden en dus grofweg stellen dat *ons* 95 procent betrouwbaarheidsinterval dus met 95 zekerheid het ware gemiddelde bevat (ik kan het toch niet laten om te zeggen dat het wel een beetje raar is want het gemiddelde is een getal en kan natuurlijk niet ergens voor een gedeelte in of tussen liggen). Afgezien van welke interpretatie je kiest (maakt mij niet uit) hebben we ook gesmokkeld met het 'aantal' *standard errors* (dat we naar links en rechts zijn opgeschoven om bij die grenzen te komen).

4§3 Betrouwbaarheidsinterval voor een populatie gemiddelde aan de hand van de t-verdeling.

Dus nu de berekening voor een betrouwbaarheidsinterval (*confidence interval*) volgens de precieze regels en verdelingen. We maken hierbij gebruik van een nieuwe kansverdeling, ook wel de *Student's t-distribution*, of simpel weg de *t*-verdeling.

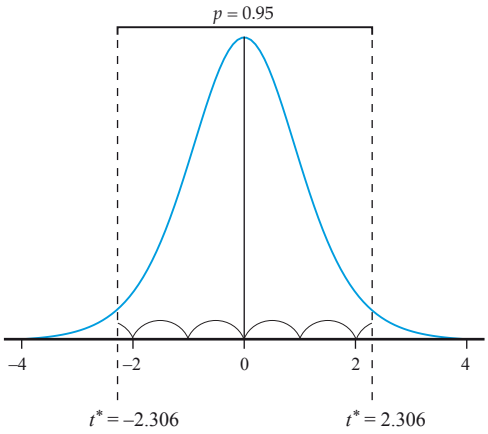
Definitie:
Een z – score is het aantal standaardafwijking dat een ruwe gebeurtenis verwijderd zit van het gemiddelde (of andere verwachting).

Definitie:
Een t – score is het aantal *standard errors* dat een ruwe gebeurtenis (een statistiek) verwijderd zit van de verwachting (vaak, maar niet altijd, het gemiddelde).

Omdat wij de *lower* en *upper-bound* voor het ware gemiddelde (μ_y) willen uitrekenen (bepalen), moeten we dus weten hoeveel *standard errors* we precies naar links en naar rechts moeten vanuit ons gevonden gemiddelde ($\bar{y}$ = 150).

Het aantal *standard errors* hangt dus af van de t-waarde en de betrouwbaarheid (bijv, 90, 95 of 99 %), die we kunnen opzoeken in de t-tabel. Gelukkig doen de statistiek-programma's dit allemaal voor je. Maar om het beter te kunnen begrijpen, doen we het een keer handmatig. Als we de normaal-verdeling zouden gebruiken voor kans berekeningen en betrouwbaarheidsintervallen, ga je er eigenlijk van uit dat sigma (σ_y in ons geval) bekend is, en laten we wel wezen; de waarde van de standaardafwijking van een populatie is per definitie *onbekend*! Omdat we ook niet zeker weten wat de waarde van σ_y is, moeten we de waarde van σ_y dus schatten op basis van de standaardafwijking die we in de steekproef gevonden hebben (s_y = 19.36). Als de steekproef heel groot zou zijn geweest, heb je meer zekerheid dat de waarde van de steekproef-standaardafwijking niet ver van de ware waarde af ligt (dus meer lijkt op σ_y) en heb je dus meer zekerheid dat de (punt-) schattingen voor het gemiddelde, terecht zijn. De t-verdeling houdt rekening met deze extra onzekerheid. Hoe kleiner de steekproef, hoe minder betrouwbaar je schattingen voor σ_y. Eigenlijk is er dus - voor elke waarde van steekproefgrootte *n* - een aparte verdeling. Je zou dus ook wel kunnen zeggen dat de t-verdeling een familie van vele verdelingen is. Welke t-verdeling je moet gebruiken om betrouwbaarheidsintervallen uit te rekenen, hangt dus af van je steekproefgrootte. Vrijheidsgraden zeggen indirect iets over je steekproefgrootte en bij het berekenen van de standaardafwijking gebruiken we altijd $df = n - 1$. Nu dus ook, Als we de benodigde t-waarde (t^*) willen opzoeken in een t-tabel (in ons geval voor df = 9-1=8 en 95 % zekerheid), kijken we dus naar $df = 8$ (vaak eerste kolom, links) en het 95 procent betroubaarheids-niveau (boven of onderaan de tabel, het *confidence level*). In ons geval heeft t* dus de waarde 2.306. Dit betekent niks meer dan dat we dus precies 2.306 *standard errors* naar links en naar rechts moeten, om vanuit het gevonden steekproef-gemiddelde, de *lower* en *upperbound* te bereiken!

figuur 4C



t-verdeling voor 8 vrijheidsgraden

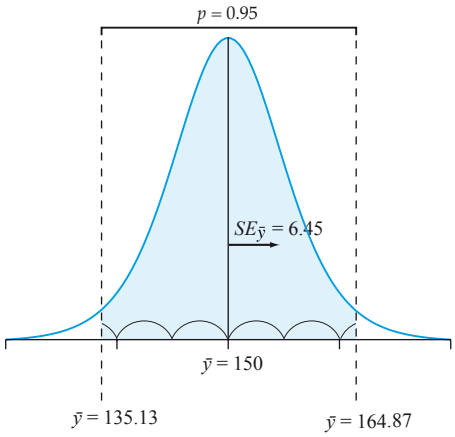
$$Lower\ Bound = \bar{y} - t^* \cdot SE_{\bar{y}}$$

$$Lower\ Bound = 150 - 2.306 \cdot 6.45 = 135.13$$

$$Upper\ Bound = \bar{y} + t^* \cdot SE_{\bar{y}}$$

$$Upper\ Bound = 150 - 2.306 \cdot 6.45 = 164.87$$

figuur 4D



margin of error

We mogen nu dus concluderen dat op basis van onze berekeningen, het ware gemiddelde (μ_y), met 95 procent zekerheid, ergens tussen de 135.13 en de 164.87 ligt. Het hele stuk dat je naar links (of naar rechts) moet wandelen vanaf $\bar{y}$ = 150 (2.306 * 6.45 = 14.87), noemt men ook wel de *margin of error*. Voor ons verhaal heeft de *margin of error* dus een waarde van 14.87 cm. Hoe kleiner de *margin of error*, des te stabieler (minder variatie) zal het steekproef-gemiddelde zijn (bij herhaling). Bij een kleinere *margin of error* (*m*) hoort een minder breed betrouwbaarheidsinterval, en zul je dus preciezere schattingen kunnen doen, wel zo wenselijk dus.

$$Margin\ of\ Error = m = t^* \cdot SE_{\bar{y}} = t^* \cdot \frac{s_y}{\sqrt{n}}$$

Valt je op dat het betrouwbaarheidsinterval nu breder (en *m* groter) is geworden in verhouding tot onze eerste benadering in paragraaf 4.2? Dit ligt in de lijn der verwachting, juist omdat we niet precies weten wat de echte waarde van σ_y is. We moesten die (ook) schatten! 'Schat op schatfout, *small estimated cutie*!' (ik hoorde mezelf zeggen dat dit niet grappig is, en hopelijk alleen Inzichtelijk voor Nederlandstalige statistici). Nu we in deze paragraaf de *correcte* berekeningen hebben uitgevoerd, kunnen we dus ook tot valide uitspraken komen over het populatie-gemiddelde. We hebben dus eindelijk de volgende (onderzoeks-) vraag beantwoord:

Waar ligt het gemiddelde van een populatie?

Dus ons antwoord is:

Het populatie gemiddelde qua lengte ligt met 95 procent zekerheid ergens tussen de 135.13 en 164.87 cm.

of

De beste punt-schatting voor het populatie-gemiddelde is 150 maar waarbij we rekening moeten houden met een *margin of error* (fout-marge) van 14.87 cm (gebruikmakend van een 95 procent *confidence level*)

Besef dat het berekenen van betrouwbaarheidsintervallen voor elke parameter mogelijk is, of je het dus nu doet voor een correlatie (ρ), regressiegewichten (β_0, β_1), standaardafwijking (σ), echt *anything* dus.

4§4 Toetsing van een *idee* aan de hand van een stelsel van Hypotheses.

Als je bij de bakker een brood gaat kopen, maar je krijgt een fiets, vraag jij je dan ook af of 'de bakker' nog wel 'de bakker' is? Maar als je een kopje koffie bij 'de slager' krijgt, vermoed je dan meteen dat hij 'een barista' is?

Sommige extreme gebeurtenissen verwacht je op grond van jou ideeën (modellen) - simpelweg- niet. Sommige extreme gebeurtenissen treden soms – of toevallig – op.
Wanneer zou jij je ideeën aanpassen?

In paragraaf 4.3 hebben we gekeken naar de meest basale onderzoeksvraag mogelijk. 'Waar ligt iets, wat is de waarde van iets?'. Ik hoop dat je beseft dat je voor deze vraagstelling, geen enkel idee (verwachting, theorie of norm) over de populatie hoeft te hebben, omdat dit soort vraagstelling heel open zijn, kun je ze dus ook vrij eenvoudig beantwoorden (aan de hand van een betrouwbaarheidsinterval). Veel vaker zit er een *vergelijking* in een onderzoeksvraag. Misschien is het wel bekend (of aangenomen) dat aapjes van vroeger (zeg 50 jaar geleden) een gemiddelde lengte van 135 cm hadden. En dat een overschot aan verdwaalde groeihormonen of andere gekkigheid (vanwege de bio-industrie) in het oerwoud, misschien wel heeft geleid tot een verandering of verschuiving van het gemiddelde en dat ze dus, gemiddeld gezien, groter zijn geworden. Of misschien vermoed je juist wel dat - door schaarste in de rimboe – ze juist kleiner zijn geworden (gemiddeld gezien). Bij dit soort vraagstellingen heb je dus wel degelijk een idee of vermoeden over hoe de werkelijkheid (populatie) eruit ziet, en wil je dus de huidige situatie met jou idee vergelijken, zodat je achteraf een uitspraak kunt doen of jou vermoeden (van verandering of verschil) klopt of niet met de werkelijkheid waarin we leven. Natuurlijk, ook dit zal weer gaan aan de hand van kans-verwachtingen en een steekproeven-verdeling.

Goed, stel dat ik dus het idee heb, dat de gemiddelde lengte van aapjes de afgelopen jaren *veranderd* is (door te veel groeihormonen, schaarste of wat dan ook) en dat ik die eventuele verandering wil onderzoeken. Welke (statistische) procedures moet we dan doorlopen? We gaan een significantie-toets doen om te kijken welk idee (geen verandering *versus* wel verandering) het meest waarschijnlijk is.

Op onze aarde in ons universum (ik laat de ongeveer 10^{500} andere mogelijke universa even buiten beschouwing) kan toch maar echt één van de twee ideeën, waar zijn. Twee stelling die elkaar ontkennen *kunnen* niet allebei waar zijn (de quantum-mechanica en de *string*-theorie uit de moderne natuurkunde leren ons hier heel iets anders). In ons geval hebben we dus de twee stellingen, hypotheses, theorieën of ideeën die elkaar tegenspreken (ontkennen):

H_0 : *Aapjes van nu hebben hetzelfde gemiddelde qua lengte als aapjes van 50 jaar geleden.*

H_1 : *Aapjes van nu hebben een ander gemiddelde dan 50 jaar geleden.*

Twee ideeën of stellingen, de Nul Hypothese en de Alternatieve Hypothese.

De *Null-hypothesis* (H_o, de nul-hypothese of zeg: H-nul) stelt *altijd* dat er *geen* verschil (tussen twee of meer dingen) of geen effect (verband, associatie, samenhang, allemaal hetzelfde) is tussen twee of meer variabelen. De *Alternative Hypothesis*, (H_1, alternatieve hypothese of zeg: H-één) zegt altijd dat er *wel* een verschil is tussen twee dingen of *wel* een effect tussen twee of meer variabelen is. Hypotheses die we opstellen, doen altijd een uitspraak over populatie-gegevens en dus *nooit* over steekproef-gegevens (daar hoef je immers niet aan te twijfelen, dat heb je al *berekend*) We kunnen de nul-hypothese ook iets technischer (of wiskundiger) formuleren:

$$H_0 : \mu_{\text{lengte moderne aapjes}} = \mu_{\text{lengte aapjes van 50 jaar geleden}}$$

$$H_0 : \mu_{\text{aapjes}} = 135 = (\mu_0)$$

μ_0 staat voor de verwachting (135) voor het gemiddelde onder de voorwaarde dat de H_o waar is, of kortweg:

$$H_0 : \mu_{\text{aapjes}} = \mu_0$$ of uiteindelijk het meest praktisch in gebruik:

$$H_0 : \mu_{\text{aapjes}} - \mu_0 = 0$$

Als twee dingen dezelfde waarde hebben, dan is het *verschil* natuurlijk nul.

Voor de alternatieve hypothese hoef je alleen maar het is-gelijk-teken (=) te veranderen in een on-gelijk-teken:

$$H_1 : \mu_{\text{lengte moderne aapjes}} \neq \mu_{\text{lengte aapjes van 50 jaar geleden}}$$

$$H_1 : \mu_{\text{aapjes}} \neq 135 \neq (\mu_0)$$

$$H_1 : \mu_{\text{aapjes}} \neq \mu_0$$

$$H_1 : \mu_{\text{aapjes}} - \mu_0 \neq 0$$

Het verschil wijkt af van nul, aangezien de waarde onder of boven nul kan zijn, noemen we dit een *tweezijdige* alternatieve hypothese.

Denk in verschil, natuurlijk kan ik zeggen dat mijn ouders even lang zijn, maar beter zeg je gewoon (wiskundig) dat *het verschil* in lengte nul is. Misschien is dat hier een beetje vreemd, maar als er tussen twee dingen *wel* een verschil is (mijn pappa heeft een lengte van 173.43 cm en mijn moeder is 163.56 cm) dan willen we veel liever weten hoe groot dat verschil is (173.43-163.56 = 9.87). Aan het verschil van 9.87 cm kunnen we makkelijker zien hoe heftig (belangrijk) dat verschil is, en waarom zou je de lezer niet verwennen met één handeling minder? Ook voor de toetsings- procedure, die we zo gaan uitvoeren, denk je in verschil en niet in twee losse waarden! Dus voor alle duidelijkheid: onder (waarheid van) H_o verwachten we dus dat het verschil 0 is. Omdat we (statistici en de meeste studenten) lui zijn, formuleren we het stelsel van hypotheses dan ook vaak als volgt:

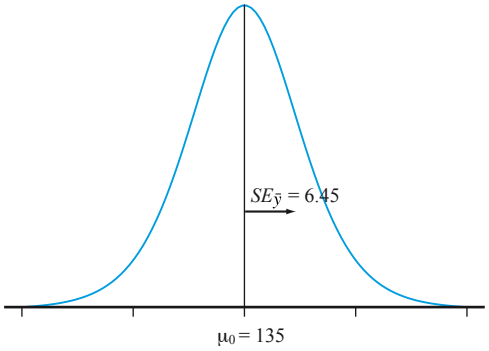
$$H_0 : D = 0$$

D natuurlijk voor *Difference*, in ons geval dus het verschil in gemiddelden.

$$H_1 : D \neq 0$$

Het gokspelletje komt ook hier van pas, dus laten we daar maar mee beginnen. *Stel* dat de H_0 waar is en aapjes van nu, gemiddeld dezelfde lengte hebben als aapjes van vroeger. Stel dat we nog *geen* steekproef zouden hebben genomen, en we dat nu pas gaan doen. Wat is in dit geval jou verwachting voor het nieuwe steekproef-gemiddelde? We voorspellen dus *weer* een statistiek (en dus niet een individueel data-puntje), maar nu onder de voorwaarde dat de nul-hypothese waar is! Nou, wat zou je gokken? Ik hoop dat je zegt: '135 cm', want dat was vroeger ook zo en als aapjes niet veranderd zijn, zou je datzelfde getal weer moeten vinden, nou vooruit, iets rond de 135 dan, waarschijnlijk dus niet precies (lang leve steekproef-fluctuatie!). We kunnen natuurlijk de steekproeven-verdeling van het gemiddelde onder de H_0 tekenen. En wel met de zelfde *standard error* die we eerder tegen zijn gekomen ($SE_{\bar{y}}$, gebaseerd op s_y en n). Ook hier maken we weer gebruik van de t-verdeling en wederom voor 8 vrijheidsgraden:

figuur 4E



P-Values

Aan de hand van deze verdeling, kunnen we weer kans-uitspraken doen over een *range* van mogelijke waarden voor het steekproef-gemiddelde die kunnen optreden, *gegeven* dat de nul-hypothese waar is (behoorlijk hypothetisch dus). Natuurlijk gebeurt het heus wel eens dat je steekproef echt niet op je populatie lijkt (niet representatief is) en dat je heel iets anders vindt dan dat je verwacht, ook al is de H_0 waar. Maar gelukkig niet vaak, er zijn nou eenmaal meer steekproeven die wel representatief zijn. Denk maar aan de grabbel-ton of dobbelsteen, maar nu weten we dat het gemiddelde 135 cm is voor die populatie in die ton. Ook hier gaan we ervan uit dat die aapjes nog steeds normaal verdeeld zijn qua lengte en dat de verdeling een (onbekende) standaardafwijking heeft (die we dus moeten schatten). Nu komt ie:

Wat zou de kans zijn dat als we een steekproef zouden nemen uit deze nul-hypothese-ton, dat we een steekproef-gemiddelde vinden, dat heftiger afwijkt van 135 cm dan de ENIGE ECHTE steekproef die we hebben getrokken?

In onze enige echte steekproef vonden we een gemiddelde van $\bar{y}$ = 150 de technische vraag kan dus als volgt geformuleerd worden:

$$P(\bar{y} \geq 150 \mid \mu_0 = 135)$$

De kans dat het steekproef-gemiddelde 150 of hoger is, gegeven dat het echte populatie gemiddelde 135 cm is. Dit noemen we een rechts-zijdige kansvraag, want eigenlijk vraag je naar de oppervlakte van de rechter staart van je verdeling.
Of:

$$P(\bar{y} - \mu_0 \geq 15 \mid \bar{y} - \mu_0 = 0)$$

De kans dat het verschil tussen het gevonden gemiddelde en de H_0 verwachting groter of gelijk aan 15 cm is (éénzijdige of rechtszijdige kans-vraag).
Of tweezijdig:

$$P(|\bar{y} - \mu_0| \geq 15 \mid \bar{y} - \mu_0 = 0)$$

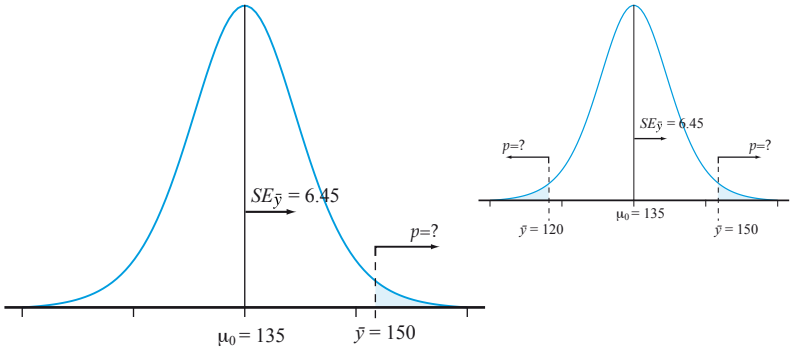
De kans dat het *absolute* verschil tussen het gevonden gemiddelde en de H_0 verwachting groter of gelijk aan 15 cm is. Deze kans-vraag is tweezijdig omdat je dus ook de mogelijkheid meeneemt dat het gevonden steekproef-gemiddelde ook wel eens meer dan 15 cm lager zou kunnen uitvallen dan de nul-verwachting (μ_0 = 135). Het gemiddelde zou bijvoorbeeld ook 118 kunnen zijn (bij een volgende trekking). Deze waarde verschilt 17 cm van je verwachting, maar dan negatief, en is absoluut gezien dus ook groter (extremer) dan 15 cm.
Of kortweg:

$$P(|D| \geq 15 \mid D_0 = 0)$$

De kans dat het absolute verschil tussen steekproef-gemiddelde en de nul-verwachting groter is dan 15 cm gegeven dat de H_0 waar is.

In de statistiek noemen we dit soort kans-vragen ook wel kortweg de p-waarde, *p-value* of de significantie van onze toetsings-procedure. En het is de p-waarde die we nodig hebben om een beslissing te maken over de houdbaarheid (waarschijnlijkheid) van de H_0 hypothese. Laten we vooral eerst bij ons die tweezijdige p-waarde uitrekenen. We beantwoorden dan dus de vraag wat de kans is dat we, absoluut gezien, een extremer verschil (onze D was 15 cm) zullen vinden (bij herhaling) dan dat wij hebben gevonden.

figuur 4F



Je ziet dat het ruwe verschil tussen de nul-verwachting en ons gemiddelde precies 15 cm is, maar dat is ruw en 'ruw is ruk'. Om de kans te bepalen, moeten we eerst weten hoeveel *standard errors* ($SE_{\bar{y}}$ = 6.45) er in dat ruwe stuk (het interval dat loopt van 135 naar 150) passen . Iets meer dan twee keer, maar hoeveel precies?

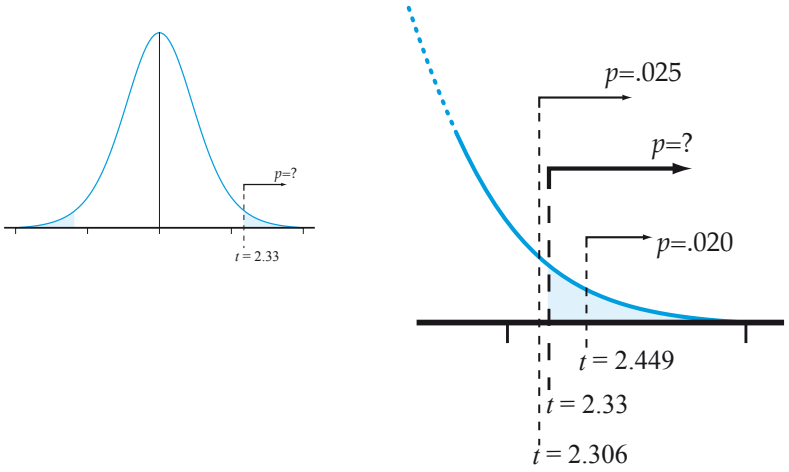
$$15/6.45 = 2.33$$

Het *aantal* keer dat de *standard error* dus tussen de nul en onze vondst past, is dus 2.33 keer. Weet je de definitie van 't' nog? We hebben dus nu een t-toets uitgevoerd voor het vergelijken van een gemiddelde met een verwachting (H_0). De berekende t-waarde (2.33) noemen we de bijbehorende toets-statistiek (deze t beschrijft de heftigheid qua (gestandaardiseerd) verschil tussen onze vondst ten opzichte van de nul-verwachting. Het enige wat we nu nog moeten doen, is de bijbehorende tweezijdige p-waarde achterhalen (opzoeken in t-tabel of a.d.h.v. software). Ook dit laat ik één keer zien aan de hand van de tabel, maar gelukkig rekenen onze programma's dat dus voor ons uit en hoeven wij het in de praktijk dus nooit zelf op te zoeken.

Omdat de vorm van t-verdeling afhangt van het aantal vrijheidsgraden zijn er eigenlijk oneindig veel t-verdelingen. De *hele* tabel zou groter (veel groter) dan een heel boekwerk zijn. Wij maken dus gebruik van een tabel waar één en ander, in is samengevat: In de t-tabel vind je alleen voor de belangrijkste (meest praktische) p-waarden en een beperkt aantal mogelijke waarden voor de vrijheidgraden (*df*), de *bijbehorende* t-waarden. Je kan dus heen en terug, soms weet je een t-waarde met vrijheidsgraden en kun je dus de bijbehorende p-waarde opzoeken. Soms is je handeling omgekeerd: je weet een p-waarde (en *df*) en zoek je juist de bijbehorende t-waarde op.

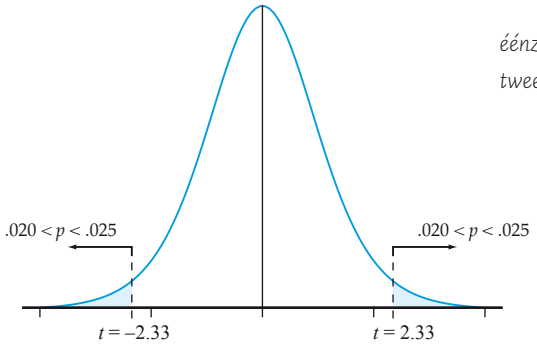
Onze t-waarde heeft een waarde van 2.33 en was gebaseerd op 8 vrijheidsgraden ($t(8) = 2.33$) en bij deze combinatie van waarden moeten we dus een p-waarde opzoeken. Het komt maar zelden voor dat jou t-waarde (2.33 in ons geval) precies in de tabel staat. We moeten hem dus vergelijken met andere t-waarden die op onze t-waarde lijken. Als je dus bij 8 vrijheidsgraden kijkt, zie je dat in die rij een t-waarde staat die net iets kleiner is dan de onze ($t(8) = 2.306$) en eentje die net iets groter is ($t(8) = 2.449$). Onze t-waarde (2.33) ligt dus tussen deze twee t-waarden in. Voor beide t-waarden uit de tabel geeft de tabel dus ook de bijbehorende (eenzijdige) rechter overschrijdingskans (p-waarde). De grotere t-waarde (2.449), die dus verder ligt van het midden (t=0) heeft een staartoppervlakte (*upper tail probability*) van p = .02. De minder extreme *t*-waarde uit de tabel (2.306) heeft een net iets grotere p-waarde, namelijk p = .025. Omdat onze *t*-waarde 2.33 was, zal zijn staart-oppervlakte ook tussen deze twee p-waarden in moeten liggen. Onze p-waarde zal dus groter zijn dan het kleinste staartje (die vast zit aan de extremere *t*) van .02 maar kleiner zijn dan het grootste staartje (.025).

figuur 4F



Best wel verwarrend allemaal. Hoe extremer (groter) de *t*-waarde, des te kleiner de p-waarde. Vergeet nooit dat Einstein (extreem slim) het kleinste staartje heeft, juist omdat qua intelligentie zo super extreem is. Er zijn maar heel weinig personen die zijn intelligentie overschrijden (slimmer zijn). De de kans (p-waarde) om slimmer te zijn, moet daarom dus wel super klein zijn. Maar goed, onze conclusie is dus hier dat de kans, om ons steekproef-gemiddelde bij herhaling aan de rechter kant te overschrijden, ergens tussen de 2 en 2.5 procent ligt ($.02 < p < .025$). Dit was de eenzijdige overschrijdingskans, nu de tweezijdige. Omdat we ook de kans dat een steekproef-gemiddelde net zo extreem of extremer is – maar dan aan de linker of onderkant – erbij willen betrekken, zijn we dus nog niet klaar. De p-waarde die hoort bij een overschrijding van 15 cm vanuit de H_0 verwachting, maar dan naar links is het zelfde als de vorige p-waarde, de twee gebeurtenissen zijn immers net zo extreem, dus net zo 'moeilijk' te bereiken vanuit het midden.

figuur 4H



éénzijdige overschrijdingskans: .020 < p < .025
tweezijdige overschrijdingskans: .040 < p < .050

Als ik naar Den Haag reis dan is dat ongeveer tussen de 15 en 20 km, maar precies weet ik het dus niet. Maar als ik dus heen en terug reis, tussen welke twee uiterste waarden ligt die afstand dan?
Tussen de 30 en de 40 km, beide waarden verdubbelen, dus ook bij p-waarden als je geen precieze schattingen hebt. De twee-zijdige p-waarde ligt dus tussen de 4 en de 5 procent ($.04 < p < .05$).
Nu hebben we het rekenwerk wel gehad en kunnen we dus echt overgaan tot een keuze tussen de H_0 en de H_1 . Welk van de twee stellingen zou nu het meest aannemelijk (of het meest waarschijnlijk) zijn? Volgens onze kans-berekening, zal bij herhaling van steekproeven, de kans kleiner dan 5 procent zijn, dat een steekproef-gemiddelde nog heftiger zal afwijken dan die van ons deed (wij weken al 15 cm af van de H_0 verwachting). Best vreemd allemaal.

- als (gegeven dat) de H_0 waar is, zou je verwachten dat je ongeveer een gemiddelde van 135 cm zou moeten vinden (als je een steekproef trekt).
- Wij hebben maar één steekproef getrokken en we vonden meteen 15 cm verschil tussen $\bar{y}$ en μ_0 sterker nog, bij herhaling zou de kans om weer zoiets extreems te vinden, gegeven dat de H_0 waar is, klein zijn, namelijk: ($.04 < p < .05$)
- Wat denk je, als de H_0 echt waar was geweest (en laat hem nog steeds waar zijn), en je zou weer een steekproef mogen trekken? Zou je dan voorspellen dat je weer zo'n (of extremer) verschil zult vinden? Of is de kans dan groter dat je een gemiddelde dichter bij 135 cm zult vinden?
- JA, want als de kans maar tegen de 5 procent aan zit om een extremere afwijking te vinden, is de kans dus ruim 95 procent dat je bij herhaling juist iets minder extreems zou vinden (minder dan 15 cm bij μ_0 = 135 vandaan)
- Is het dan niet raar dat we toch 150 hebben gevonden als de H_0 waar zou zijn?
- JA! of in ieder geval onwaarschijnlijk(er)
- Omdat ons eigen resultaat (150) moeilijk te overschrijden is als de H_0 *echt waar zou* zijn, schrikken we eigenlijk een beetje van onze vondst
- Vanuit 'de bakker' – verwachting, is ons steekproef-gemiddelde als een fiets in plaats van brood.
- Natuurlijk ga je dan twijfelen of de H_0 dan wel *echt* waar is.

Significantie niveau.

Als we de gevonden *p-value* klein genoeg vinden, kunnen we concluderen dat het gevonden steekproef resultaat, significant (belangrijk genoeg) afwijkt van wat we onder de H_0 verwacht hadden en verwerpen we de H_0. We zeggen in dit geval dat de H_0 onhoudbaar is of ook wel niet meer te verdedigen en verwerpen we dit standpunt. In een scriptie of artikel, mag wat mij betreft het woord 'nul-hypothese', niet eens voorkomen. Tegen je publiek spreek je uiteraard in *gewoon* Nederlands:

'Het idee dat de gemiddelde lengte van die aapjes het zelfde is gebleven, is ontzettend onwaarschijnlijk en kan beter verworpen worden. We kunnen dus beter concluderen dat de

gemiddelde lengte van die aapjes tegenwoordig hoger ligt dan 135 cm.'

Of om je claim ook nog eens statistisch of wetenschappelijk te ondersteunen:

'Om te onderzoeken of aapjes van tegenwoordig gemiddeld een andere lengte hebben dan 50 jaar geleden, is een *one-sample t-test* uitgevoerd. Het gemiddelde van de populatie van 50 jaar geleden lag op 135 cm. Het steekproef-gemiddelde (M = 150.00 cm, Sd = 19.36, n = 9) lag 15.00 cm hoger en bleek significant af te wijken van de oude norm ($t(8) = 2.33$, $p = .049$). We mogen dus concluderen dat de gemiddelde lengte van hedendaagse aapjes, is toegenomen t.o.v. 50 jaar geleden.'

Wanneer vinden we een p-waarde klein genoeg om tot verwerping van de H_0 over te gaan? Dat hangt er helemaal van af hoe erg (kwalijk) je het vindt om een foute beslissing te maken. Wat is het risico van een verkeerde beslissing? Stel je voor dat we onze aapjes toevallig rond een stortplaats met groei-hormonen hadden gevonden, natuurlijk zijn ze dan groter (causaal zelfs) en dan hadden we - slechts op grond van hun *data-puntjes* - het echte populatie gemiddelde (het idee van ouwe vertrouwde kleine aapjes) flink ontkent. Onterecht zeggen of beslissen dat er wel een verandering, wel verschil of wel een effect is, terwijl dat niet waar is, heeft *soms* heel heftige implicaties. Als je, als arts beslist dat iemand ziek is, de patiënt dus een positieve diagnose geeft, wil dat nog niet zeggen dat die diagnose ook daadwerkelijk klopt. Als je onterecht zegt dat aapjes gemiddeld langer zijn geworden, of – onterecht – dat iemand ziek of schuldig is, maak je een type I fout (*type I error*). Ik heb een *donker* en *bruin* vermoeden, dat er in de Verenigde Staten meer mensen onterecht vast zitten dan in Nederland. Maar hoeveel mensen gaan er eigenlijk onterecht vrij-uit? Zouden ze het daar beter doen dan hier en wat is 'beter'? Er bestaat dus ook een type II fout (*type II error*). Een fout, van de tweede soort, bega je als je onterecht de H_0 behoudt. In dit geval beslis je dus dat er geen verschil, effect of verandering heeft opgetreden, terwijl in de werkelijkheid dit toch echt wel het geval was. Iemand onterecht vrijspreken, of gezond verklaren, valt dus ook onder het maken van een type II fout. Welk *soort* – of *type* van – fout vind je erger om te maken? Dit zijn vaak praktisch en of ethische kwesties en levert mega-spannende discussies op. Maar onthoud, voor nu, vooral één ding: In al deze situaties kun je stellen dat hoe vaker je een type I fout maakt, des te minder maak je een type II fout, maar natuurlijk ook anders om. Het Nederlands rechtssysteem is, statistisch gezien, conservatiever dan in de V.S., in Nederland zijn we (hopelijk) voorzichtiger en hebben heel veel bewijs (fiets, verschil, effect) nodig voordat we tot een veroordeling of beslissing over gaan.

When P is low, *H-O* must Go!

Bij het maken van beslissingen over het al dan niet verwerpen van de H_0, hanteren we meestal een significantie niveau van 5 procent. Dit wil zeggen dat we het nog net acceptabel vinden om een kans van $p = .05$ te hebben op het maken van een type I fout. In de rechtspraak zou dit dus betekenen dat je dus maximaal 5 procent van de *onschuldige* (H_0 is waar) mensen toch (onterecht) veroordeelt (beslissing). We gebruiken het symbool α (spreek uit als 'alfa') om aan te geven hoe conservatief we beslissen (of hoe streng we toetsen). Als je strenger of conservatiever wilt toetsen, zet je de α op een lagere waarde, bijvoorbeeld $\alpha = .01$. In dit geval zal de kans op een type I fout dus kleiner zijn. In de medicijn-studies, waar foutieve beslissingen vaak ernstige gevolgen kunnen hebben, is het bijvoorbeeld zaak om op een strenger niveau te toetsen, zeg bijvoorbeeld, $\alpha = .001$. Als de p-waarde van jou test lager of gelijk is aan het α– niveau, dat je hanteert, dan verwerp je de H_0 (waarmee de H_1 dus aannemelijk wordt, maar nog niet waar hoeft te zijn). Samengevat:

$p \leq \alpha$ H_0 verwerpen (H_1 is aannemelijk)

$p > \alpha$ H_0 *niet* verwerpen.

Als jou p-waarde dus groter is dan 5 procent, betekent dit dus eigenlijk dat, gegeven de H_0 er een redelijke grote kans is, dat je bij herhaling van steekproef-trekking, een nog meer afwijkend resultaat (van de H_0 verwachting) vindt dan dat je in jouw steekproef hebt gevonden. Alsof je dus koffie bij de slager krijgt (niet héél vreemd, en je concludeert dus dat de slager gewoon nog je slager is). Extremere verschillen of effecten zullen dus regelmatig optreden, ook al is de H_0 waar. Als de p-waarde heel klein is, is de kans om iets extremers te vinden, gegeven dat de H_0 waar is, dus heel laag. Alsof je dus een fiets krijgt bij de bakker (héél vreemd, heel onverwacht en dus twijfel je aan de hoedanigheid van de bakker, misschien is hij dus toch stiekem fietsenmaker).

4§5 Significantie-Toets voor Verband.

Voor elke parameter bestaat een test of toets om te kijken of de waarde voor deze parameter in de populatie afwijkt van een bepaalde waarde (vaak de waarde 0) of niet. Als je dus wilt toetsen of het verband tussen leeftijd en lengte (de correlatie-coëfficiënt ρ_{xy}) van nul afwijkt in de populatie aapjes, kan dat! En natuurlijk laat ik dat nog even zien aan de hand van onze steekproef waarbij wij een correlatie van r_{xy} = .89 hadden gevonden. Ik doe de uitwerking even zo kort mogelijk (eindelijk) zodat je ziet dat er eigenlijk maar een paar handelingen nodig zijn om tot een conclusie te komen. Eerst stellen we het stelsel van hypotheses op, die gaan uiteraard, enkel en alleen, over de populatie:

$H_0 : \rho_{xy} = 0$ Er is geen verband tussen leeftijd en lengte.

$H_1 : \rho_{xy} \neq 0$ Er is wel verband tussen leeftijd en lengte.

We kunnen de bijbehorende toets-statistiek '*t*' (dus ook hier een t-test) uitrekenen met de volgende formule:

$$t = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1 - r_{xy}^2}} \quad \text{met } df = n - 2$$

Onze benodigde steekproef-gegevens: r_{xy} = .89, n = 9

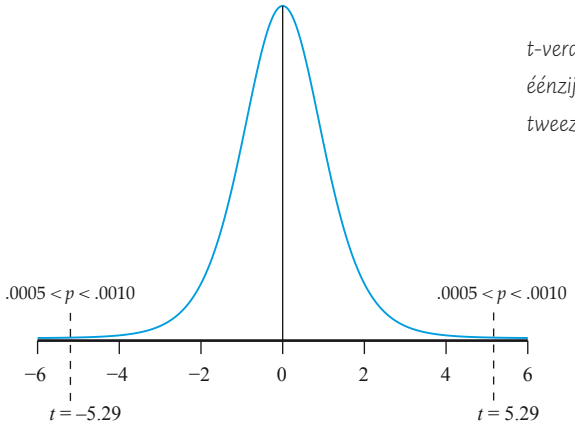
invullen geeft:

$$t = \frac{.89 \cdot \sqrt{9-2}}{\sqrt{1 - .89^2}} = .89 * \sqrt{7} / \sqrt{(1 - .89^2)} = 5.29$$

Dus onze toets-statistiek *t* heeft een waarde van 5.29, met – of gebaseerd op - 7 vrijheidsgraadjes. Afhankelijk van je rekenmachientje, zijn de haakjes onder de wortel *irrelevant* (denk daar maar is (es) over na (sorry, ik kon het niet laten)).

Nog even de (tweezijdige) *p*-waarde opzoeken...
Dit doet *R* – normaal gezien – voor je, maar uit de *t*-tabel blijkt:

figuur 4I



t-verdeling voor 7 vrijheidsgraden
éénzijdige overschrijdingskans: .0005 < p < .0010
tweezijdige overschrijdingskans: .0010 < p < .0020

$.010 < p < .025$ Dus (in ieder geval) kleiner dan (of gelijk aan) $\alpha = .05$

Conclusie: *P* is low, so *H-O* must go!
Conclusie in woorden : Geen verband is onwaarschijnlijk dus een (positief) verband tussen leeftijd en lengte is aannemelijk.
Conclusie voor het volk:
'Uit onderzoek naar aapjes rond de leeftijd van 1 à 2 jaar, bleek dat relatief jongere aapjes over het algemeen wat kleiner zijn dan oudere aapjes.'

Conclusie voor ons (wij willen *keiharde* cijfers zien):
Om te onderzoeken of er een samenhang bestaat tussen de leeftijd van aapjes en hun lengte, is een steekproef ($n = 9$) genomen (op een heel mooi én representatief eiland). Uit analyse bleek een positieve, lineaire samenhang (zie figuur XX). De Pearson correlatie ($r = .89$) week significant af van nul ($t(7) = 5.29$, $p = .011$ (God Zij Dank, dat die *p*-waarde kleiner was dan onze Heilige α)), waaruit we kunnen concluderen dat jonge aapjes over het algemeen kleiner zijn dan oudere aapjes (binnen de leeftijden van 1 en 2 jaar).

Heel *Leuk* die samenhang, maar mogen we nu eindelijk ook tot voorspellingen over gaan?

Check! (Without Residuals Deviating From Zero),

Alles is immers Regressie.

Hoe Fijn.

Enkelvoudige Lineaire Regressie-Analyse.

5§0 De voorspelling van een belangrijke variabele op basis van één andere (toch óók wel een beetje belangrijke) variabele.

Inmiddels ben je al zo ver gekomen, dat je het volgende ook nog wel aankan (is *mijn* voorspelling); Door mijn gebrabbel (Error) heen, mag ik hopen (Model), dat je – let op – één ding inmiddels duidelijk is geworden: In de wetenschap draait het uiteindelijk om tast-bare (dus repliceerbare) en zo correct mogelijke *voorspellingen* (jou theorie werkt en anderen kunnen dat checken) – Maar dit geldt dus ook voor het gewone, en dus ONS leven! Denk alleen al aan je motoriek, toch wel handig als je weet (voorspeld hebt) *wanneer* je je hand moet dicht doen, als je een balletje aan het overgooien bent en hem probeert te *vangen*. Of als je inschat *hoelang* het fietsen is naar school, om op tijd te komen. Misschien is de allerbelangrijkste wel de inschatting qua uren studietijd, die je nodig hebt om het tentamen te halen!
Heb je ooit wel eens van 'regressie-therapie' gehoord? Hierin wordt je (onder hypnose) terug gebracht naar je essentie of het begin van je leven in de baarmoeder. Je weet namelijk maar nooit, stel je voor, dat je in die periode de nodige frustraties hebt opgelopen... Die frustraties (of zegeningen) zouden zomaar een verklaring kunnen zijn voor datgeen dat jij vandaag de dag zoal doet.
Elke keer, als we - zo goed als het kan - een inschatting maken, is het telkens weer de vraag: 'Op basis *waarvan* kunnen we de beste inschatting maken?'. *Wij* slaan de baarmoeder weliswaar over en zoeken naar *voorspellers* (x-en) voor de te verklaren variabele (y), met een iets *tastbaardere* voorspel-kracht!

5§1 Beter voorspellen aan de hand van een lineaire vergelijking.

Denk even aan een kaars, zo'n romantische, op tafel tijdens een diner. Hoe lang zou die zijn qua lengte? Het antwoord hangt af van *minimaal* drie zaken. De eerste van de drie belangrijkste zaken, is hoelang de kaars was *op het moment* dat hij werd aangestoken. Het tweede aspect, is *hoe snel* de kaars *korter* word (hoe snel het kaarsvet smelt en dan verbrand). *Hoe snel* die kaars brand, hangt weer van een heleboel andere zaken af, zoals temperatuur in de kamer, luchtdruk en bijvoorbeeld tocht. Maar ik wil het alleen even hebben over de *begin* lengte van die kaars, en de *snelheid* waarmee hij korter wordt. Het derde aspect is de tijd dat de kaars gebrand heeft, hoe langer hij brand hoe korter die zal zijn. Stel dat deze kaars 30 cm was, toen hij net werd aan gestoken (toen de je dus net aan tafel ging zitten en besloot om hem aan te steken). En zeg dat deze kaars - per uur – ongeveer 6 cm korter wordt. Met deze informatie kunnen we voor – ieder tijdstip tijdens dat etentje – dus uitrekenen of voorspellen wat (ongeveer) de lengte is van die kaars met de volgende formule (model):

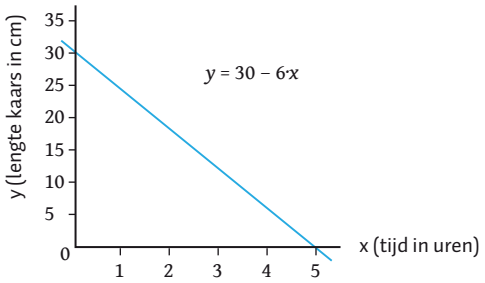
$$lengte\ kaars = 30\ cm - 6\ cm \cdot aantal\ uur$$

of in termen van x en y:

$y = 30 - 6 \cdot x$ y als lineaire functie van x

Een formule van de vorm zoals hierboven, noemen we ook wel een 'lineaire vergelijking'. Als je de formule als grafiek of lijntje tekent (zie figuur 5A), wordt duidelijk waarom dit model (de vergelijking), een *lineair* model wordt genoemd.

figuur 5A



Moeilijk gezegd, kunnen we nu de 'lengte van de kaars' uitdrukken (beschrijven in termen van) in het 'aantal uren' dat die kaars gebrand heeft. De lengte van de kaars is dus afhankelijk van drie dingen:
- Hoelang de kaars was toen hij werd aangestoken op tijdstip *x* = 0
- De brand-snelheid, gemeten in cm per uur
- De tijd dat de kaars gebrand heeft (gemeten in aantal uren, *x*)

startgetal en hellingsgetal, *intercept en richtingscoëfficiënt (slope)*

We zeggen ook wel dat dit model twee parameters heeft, namelijk het startgetal en het hellingsgetal en we hebben voor onze kaars dus twee vaste waarden (respectievelijk 30 en -6). Het startgetal noemen we vaak het *intercept* (ook wel snijpunt van de lijn met de y-as) en staat voor de begin waarde van y als x een waarde van nul heeft. Onze kaars heeft een (start-) lengte van 30 cm op tijdstip *x* = 0. Het hellingsgetal noemen we vaak de *richtingscoëfficiënt* (of de *slope*) van de de lijn en verteld ons hoe snel de lijn daalt of stijgt. Onze kaars wordt 6 cm per uur *korter* en omdat de kaars dus korter wordt, noemen we het verband, tussen x en y negatief. We kunnen nu dus zeggen dat voor elk uur dat het diner vordert (x dus met één eenheid toeneemt), de lengte van de kaars met 6 cm afneemt.

Stel dat we nog niet zouden weten hoelang de kaars was bij het begin, maar toch een algemene formule willen opstellen, dan ziet die formule er als volgt uit:

$y = intercept + slope \cdot x$ of omdat we graag symbolen gebruiken:

$y = a + b \cdot x$ *hierin* is *a* het *intercept* en *b* de *slope*, of in een andere vorm:

$y = a \cdot x + b$ *hierin* is a de *slope* en b het *intercept*

voorspeller (onafhankelijke variabele) en respons (afhankelijke variabele).

Hierin is x de voorspeller (of ook wel de 'onafhankelijke variabele' of denk 'oorzaak') en y is hier de afhankelijke variabele (het antwoord, de respons of criterium variabele, of denk: 'gevolg'). Voorlopig mag je wel even *voelen* alsof y het gevolg is van x, of we zeggen ook wel: y als functie van x, y uitgedrukt in (termen van) x. Ook in het geval van de kaars is de lengte niet een direct causaal gevolg van de tijd. Het feit dat de kaars korter wordt, komt puur door de verbranding en niet door de tijd. Maar omdat de kaars systematisch (met de zelfde snelheid) korter wordt, kunnen we dus toch tijd (x) gebruiken om de lengte (y) uit te drukken of te voorspellen.

Je ziet nu dat we met een model te maken hebben, dat eigenlijk gewoon een recht (lineair) lijntje is, dat schuin naar beneden loopt. De *hoogte* (y) van deze lijn staat voor de lengte van de kaars en wordt steeds minder naarmate het aantal uren toeneemt. Op tijdstip x = 0 is de kaars 30 cm en zal elk uur, 6 cm korter worden. Naar mate de kaars *langer* brandt, zal hij dus *kleiner* worden (*duh*!). Aan de hand van de grafiek (of formule) kunnen we nu *aflezen* (of *uitrekenen*) hoe lang de kaars op een bepaald (specifiek) moment ongeveer zal zijn. Dit betekent dus dat we een nieuw voorspel-model hebben. Natuurlijk is dit nieuwe voorspel-model beter dan het meest simpele model of voorspelling, het grote gemiddelde (de gemiddelde lengte berekend over alle verschillende lengtes van de kaars gedurende het hele diner). We weten natuurlijk (nog) niet precies hoe goed ons nieuwe model zal kloppen. In het (onverwachte) geval van – een veel te – romantisch etentje, kan die kaars misschien die hitte niet aan en smelt dan toch echt sneller dan verwacht op basis van *ons* model (6 cm per uur). *Any way,* we hebben nu dus de lengte van de kaars uitgedrukt in tijd en dus een complexer (en vast wel beter) model gebouwd dan het grote gemiddelde. Nu maar wel hopen dat het het diner niet langer duurt dan vijf uur (x = 5), want volgens mij is het dan toch echt gedaan met de pret (of je huis staat in vlammen). Laten we ons richten op de aapjes en kijken *hoe goed* we hun geobserveerde lengtes kunnen *voorspellen* op basis van hun leeftijden aan de hand van een lineair regressie-model.

De enkelvoudige lineaire regressie-vergelijking, de regressie-lijn.

In de statistiek en de wetenschap draait alles om het verklaren en het voorspellen van zaken in, en om ons heen. Als er verband is tussen twee variabelen, kunnen we een beter model maken dan het nul-model. Met een nieuw (en iets ingewikkelder) model, kunnen we – gemiddeld gezien - beter gokken. Door deze nieuwe manier van voorspellen wordt (hopelijk) de gemiddelde gok-fout dus substantieel kleiner. Het meest basale model (verwachting, voorspelling van de werkelijkheid) dat we inmiddels kennen, is het grote gemiddelde. We noemen dit vaak ook wel het *intercept*-model of het *nul*-model. Bij de aapjes was het grote gemiddelde 150 cm en dus onze beste gok als we nog geen gebruik maken van informatie over hun leeftijd. We hebben gezien dat er tussen leeftijd en lengte bij onze aapjes een positief verband is (r_{xy} = .89) en dat de puntenwolk dus van linksonder naar rechtsboven loopt, via een rechtlijnig (lineair) verband. Je zou dus ook wel kunnen zeggen dat de essentie (of hier ook wel samenvattende beschrijving) van de puntenwolk, een rechte lijn, schuin omhoog is. Deze lijn wordt zodanig door de puntenwolk heen getrokken dat alle negen punten er zo dicht mogelijk bij liggen (de verticale afstanden van een punt tot de lijn, zijn *gemiddeld* zo klein mogelijk). We gebruiken deze lijn nu als nieuwe (hopelijk betere) voorspelling dan het grote gemiddelde op y. We zeggen nu dus dat de *hoogte* van de regressielijn (die bij elke waarde van x anders is, vanwege het verband tussen x en y) voor de *voorspelde waarde* van y (lengte) staat, met als symbool:

voorspelde waarde, *predicted value*

$\hat{y}_i$ de voorspelde waarde van y (voor subject nummer i), of *predicted value* of y

en $\hat{y}_i$ kun je uitreken met de volgende algemene formule, de lineaire regressie-vergelijking (als je dus de schattingen van je parameters weet):

$\hat{Y}_i = b_0 + b_1 \cdot X_i$ als regressie-vergelijking voor je steekproef met b_0 als *intercept* en b_1 als *slope*. b_0 en b_1 zijn dus statistieken.

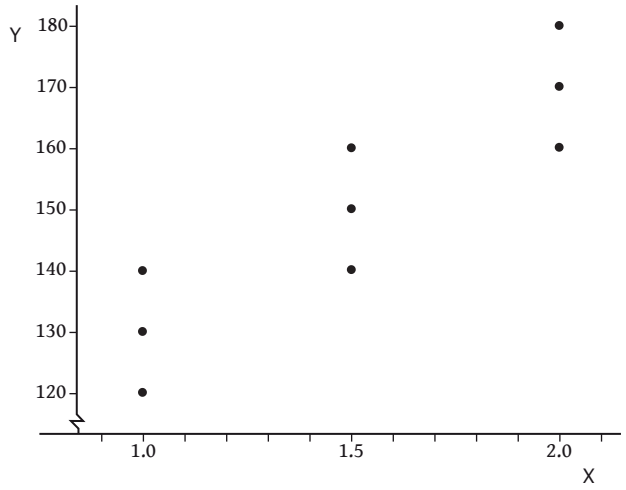
$\hat{Y}_i = \beta_0 + \beta_1 \cdot X_i$ voor de populatie zijn b_0 en b_1 de schatters (*estimates*) voor de parameters β_0 en β_1, het populatie-*intercept* en de populatie-*slope*.

Het symbool 'Y met een dakje' is dus niet hetzelfde als de geobserveerde waarde van y (y_i). De geobserveerde waarde van y is die waarde qua lengte, die we echt gemeten (geobserveerd) hebben. De (hoogte van de) regressielijn is dus wél hetzelfde als $\hat{y}_i$ en staat voor die waarde die je *zou* geven als beste voorspelling als je wél informatie over x (leeftijd) tot je beschikking hebt. Laat alsjeblieft duidelijk zijn dat zowel bij kaarsen als bij aapjes, je nieuwe voorspelling nog steeds niet altijd perfect hoeft te zijn. Ik wil hier mee dus zeggen dat de geobserveerde waarde, heel vaak, toch even iets anders is dan je voorspelde waarde.

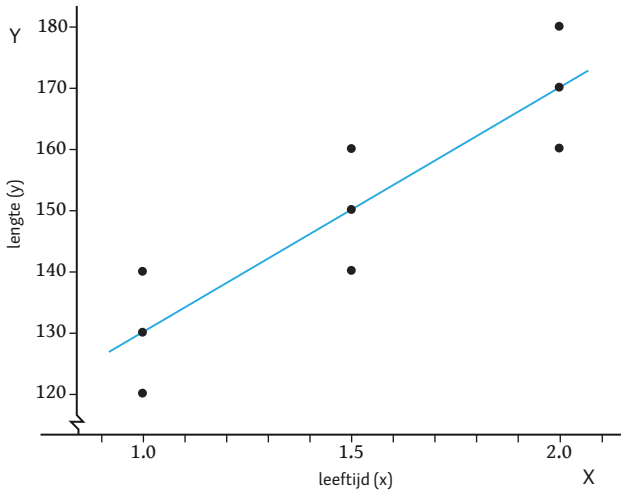
Om het woord 'regressie' beter te begrijpen, zou je ook kunnen zeggen dat je de geobserveerde waarden van y *terugbrengt* (*re-gressie*) naar de (x) waarden. Je kijkt dan dus of op basis van de x-waarden, of de y-waarden (beter) verklaard kunnen worden. Laten we eerst maar kijken of we de ligging (positie) van de lijn kunnen bepalen aan de hand van de puntenwolk (zie figuur 5B).

In figuur 5B is het hopelijk *over*-duidelijk wat de ligging of positie van de lijn moet zijn. Maar bij 'echte' data is dat natuurlijk moeilijker te zien, maar wel *ongeveer* te schatten. De ligging van een regressie-lijn kan natuurlijk ook berekent worden door de formules voor b_0 en b_1 in te vullen (of een programma als *R* te gebruiken). Die komen zo, maar eerst de bepaling aan de hand van onze grafiek (figuur 5B). Als we hier een (zo goed mogelijk passende) lijn door de puntenwolk willen trekken, zal die – bij elke waarde van x – precies door de middelste puntjes moeten worden getrokken (figuur 5C). Voor precies deze positie, liggen de 9 punten gemiddeld gezien, het dichtst bij de regressielijn (verticaal gezien).

figuur 5B

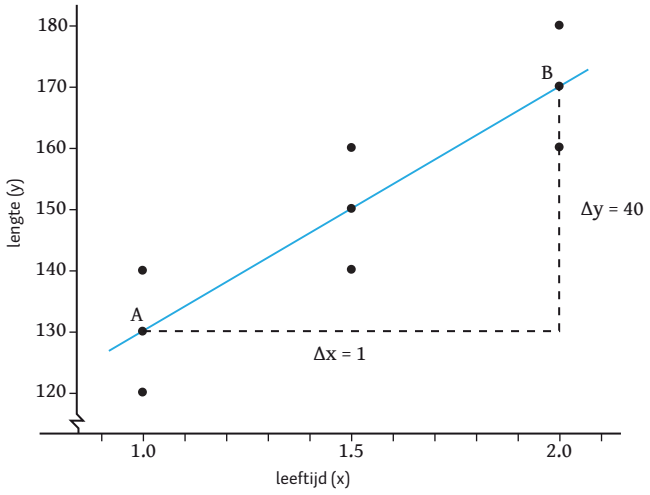


figuur 5C



Een regressie-lijn (of model) heeft twee parameters (of statistieken in geval van steekproef) die ons precies vertellen hoe we de lijn zouden moeten tekenen (mochten we zelf een plaatje ervan maken). Eerst *schatten* we de waarde van de *slope* door twee punten te zoeken die op de lijn liggen. Neem bijvoorbeeld de twee punten A en B met bijbehorende coördinaten $A(x_a = 1, \hat{y}_a = 130)$ en $B(x_b = 2, \hat{y}_b = 170)$(in figuur 5D). *Hoeveel* eenheden (in jaren) moet je naar rechts (horizontaal), en *hoeveel* eenheden (in cm) moet je vervolgens omhoog (verticaal), om van punt A naar punt B te wandelen?

figuur 5D



Je moet dus 1 eenheid (in jaren) naar rechts wandelen en 40 eenheden (in cm) omhoog, om van punt A naar punt B te wandelen. Of kortweg, omdat de lijn recht is, stijgt de lijn 40 cm per jaar.

Eigenlijk gewoon de groeisnelheid van aapjes (in cm per jaar) dus. Uiteraard zullen aapjes niet eeuwig (even snel) blijven groeien, maar deze groei-trend (40 cm per jaar) geldt in ieder geval voor onze aapjes, voor hun waarden van x. Om vervolgens de start-positie of het intercept het lijntje te bepalen, zou je bijvoorbeeld vanuit punt A (met $x_a = 1$) horizontaal naar de y-as kunnen wandelen, je zult dus één jaar naar links moeten wandelen, maar hoeveel moet je nu omlaag om weer op ons lijntje terecht te komen? Juist omdat de lijn recht is kunnen we dus ook terug redeneren. Een eenheid naar *links* wandelen vanuit punt A, betekent dus *nu* 40 eenheden *omlaag*. Omdat punt A een hoogte had van $\hat{y}_a = 130$, komen we dus uit op de y-as op een hoogte van 90 cm, het snijpunt met de y-as. In formule vorm wordt dit dus:

$\hat{lengte}_i = 90 + 40 \cdot leeftijd_i$ of in termen van Y-dakje en X;

$\hat{Y} = 90 + 40 \cdot X_i$ Je hebt dus nu een *heuse* regressie-vergelijking!

De berekening voor b_0 en b_1 aan de hand van formules en statistieken die we al berekend hadden:

Voor de *slope*:

$b_1 = \frac{s_{xy}}{s_x^2}$ waarbij de covariantie een waarde heeft van $s_{xy} = 7.50$ en de variantie een waarde van $s_x^2 = .1875$

invullen geeft:

$b_1 = 7.5 \,/\, .1875 = 40$

Deze waarde voor de *slope* hadden wij ook gevonden aan de hand van de grafiek.

Voor het *intercept*:

$b_0 = \bar{y} - b_1 \cdot \bar{x}$ met $\bar{y} = 150$ en $\bar{x} = 1.50$ geeft:

$b_0 = 150 - 40 \cdot 1.50 = 90$ klopt wéér, fijn!

5§2 Voorspelkracht

Nu we weten hoe de ligging is van de regressie-lijn voor onze aapjes, willen we natuurlijk ook weten, *hoe goed* ons nieuw voorspel-modelletje (de regressie-vergelijking) werkt. Hoe krachtig is die nieuwe manier van voorspellen? Hoe groot zou de gemiddelde gok-fout zijn als (gegeven dat) we dit nieuwe model gebruiken? We willen dus weten *hoe krachtig* een model is qua voorspelling. Het *mooie* (in ieder geval voor mij) van dit hoofdstuk is dat de kern (of praktische essentie) van statistiek, eindelijk duidelijk gaat worden. Nieuwe begrippen die je hier leert, zijn toepasbaar (generaliseerbaar) op *elk* model. Of je nu spreekt over een gedachte (idee) als model (een verwachting), een trein-modelletje van je pappa, een foto-model van TV laatst (dat pappa stiekem leuker vindt), een regressie-model, een confirmatief factor model, of werkelijk wat dan ook, het maakt mij niet uit, want elk idee of model, komt in *meer* of *mindere* mate overeen, met datgeen wat we willen bereiken of waar we naar op zoek zijn (of waar we *echt* van dromen, (ffwoowh)). En waar zijn we naar op zoek? Een prettig leven voor jezelf en anderen, mag ik hopen, maar in de wetenschap, naar een zo goed mogelijke beschrijving van wat er (in en) om ons heen gebeurt. Al die aspecten van de werkelijkheid (datgeen wat we zien of observeren) kunnen voor een *gedeelte* verklaard worden, maar ook, voor een gedeelte *niet*. In dit hoofdstuk wordt duidelijk *in hoeverre* we de werkelijkheid (met een beter model dan het nul model) kunnen verklaren, maar ook *in hoeverre,* of voor welke gedeelte we de werkelijkheid niet kunnen verklaren. We keren snel terug naar ons voorspel-spelletje, maar *nu* aan de hand van ons nieuwe model, onze regressie-vergelijking.

TABEL 5A

i	Y_i	X_i
1	120	1.0
2	130	1.0
3	140	1.0
4	140	1.5
5	150	1.5
6	160	1.5
7	160	2.0
8	170	2.0
9	180	2.0

$n = 9$

We beginnen vanaf het begin. Kijk naar tabel 5A, je mag alle informatie (data-punten en statistieken) gebruiken die je ziet. Alle negen aapjes staan op de gang en er komt er slechts één binnen wandelen, natuurlijk vertel ik je niet welke (ik weet het stiekem wel).

Vraag 1 – Wat is je beste gok qua lengte als er een aapje binnen komt wandelen?

Je roept nu natuurlijk heel hard: '150 cm, want dat is het grote gemiddelde, ook wel het nul-model, als ik iets anders had geroepen (140 cm of 20) dan zou de gemiddelde gok-fout groter worden en dat wil ik niet!' (Dit was het goede antwoord, heel goed!)

Even later... Heel toevallig komt aapje nummer 9 binnen wandelen. Oeh, je hebt in dit geval een verkeerde voorspelling gemaakt. Hoe groot was de gok-fout of individuele afwijking voor case 9?

$Y_i - \bar{Y}$ Individuele afwijking (naar het grote gemiddelde) voor case *i*.

$Y_9 - \bar{Y} = 180 - 150 = 30$ Individuele afwijking voor case 9.

Vraag 2 – Alle aapjes staan weer op de gang en ik vraag je nog een keer te gokken, maar nu krijg je extra informatie: Ik vertel je dat het aapje dat binnenkomt, een leeftijd heeft van twee jaar (x = 2). Ik heb je dus (diagnostische) informatie gegeven over een tweede variabele (leeftijd) waarvan we eerder al hadden gezien dat deze een positieve samenhang vertoont met lengte (r_{xy}= .89), wat betekent dat jongere aapjes over het algemeen wat kleiner zijn dan oudere aapjes. Sterker nog, we hebben zelfs een regressie-vergelijking gebouwd, die ons helpt bij het – beter – voorspellen:

$$\hat{Y}_i = 90 + 40 \cdot X_i$$

Omdat je dus de waarde qua leeftijd van mij hebt gekregen, kan je die (diagnostische) informatie (x = 2) gebruiken en invullen als x-waarde in de regressie-vergelijking.

$$\hat{Y}_{x=2} = 90 + 40 \cdot 2.0 = 170$$

Nu je weet wat de voorspelde waarde is voor alle aapjes, die precies 2 jaar oud zijn, roep je dus nu als antwoord (weer heel hard): '170 cm, want het aapje is een half jaar ouder dan gemiddeld en zal dus ook wel wat langer zijn. Sterker nog, ik voorspel dat hij 20 cm boven het grote gemiddelde zit.'

Ik heel blij, want je hebt het antwoord alweer goed! Doordat je nu gebruik hebt gemaakt van het nieuwe model (dat rekening houd met leeftijd, onze regressie-vergelijking), heb je je meest basale voorspelling (150) met 20 cm bijgesteld. Door regressie zeggen we nu 170 in plaats van 150 en dat is 20 punten meer. En besef dus dat je deze verschuiving in voorspelling (20 cm) *altijd*

(systematisch) zou doen als een aapje een half jaar ouder is dan het gemiddelde ($\bar{x}$ = 1.50).

Even later... *Heel toevallig* komt aapje nummer 9 weer binnen wandelen (hij is immers één van de twee-jarigen). Laten we snel kijken hoeveel jouw laatste voorspelling (gebaseerd op de leeftijd) nu nog afwijkt van de werkelijkheid. Aapje nummer 9 heeft een geobserveerde waarde van Y_i = 180 terwijl jouw voorspelling, $\hat{Y}_9 = \hat{Y}_{x=2}$ = 170 was. *Nu* ligt de geobserveerde waarde *nog maar* 10 cm boven onze nieuwe (en hopelijk betere) verwachting! Conclusie:

Ten opzichte van het nul-model, wijkt aapje nummer 9, 30 cm af, omdat:

$$Y_9 - \bar{Y} = 180 - 150 = 30$$

Door het regressie-model hebben we de voorspelling met 20 cm aangepast, omdat de nieuwe voorspelling, 170 was:

$$\hat{Y}_9 - \bar{Y} = 170 - 150 = 20$$

Omdat we het regressie-model hebben gebruikt, maken we nu een kleinere gok-fout van nog maar 10 cm:

$$Y_9 - \hat{Y}_9 = 180 - 170 = 10$$

Conclusie nog een keer, maar dan in andere woorden: We hebben de *totale* afwijking van aapje nummer 9 (30 cm) gesplit (of opgedeeld) in twee stukjes. Het eerste stukje (20 cm) is het gedeelte van de totale afwijking dat door ons regressie-model verklaard is. Het tweede stukje is het gedeelte dat we niet konden verklaren, het laatste restje dus.
We hebben nu alleen naar deze opdeling gekeken voor aapje nummer 9. En het moge inmiddels duidelijk zijn dat voor *dit* aapje het regressie-model een betere voorspelling gaf dan het nul model.

Ik laat dit verschil qua betere voorspelling, specifiek voor aapje nummer 9 en algemeen, even statistisch, technisch of wiskundig zien:

Heel algemeen:

$DATA = FIT + RESIDU$ Je komt deze uitspraak in vele vormen tegen, bijvoorbeeld:

Observed = Expected + Error Maar we bedoelen dus echt hetzelfde, wordt zo duidelijk...

Aan de hand van het nul-model (M_0) en voor aapje nummer 9:

$$Data = M_0 + Error$$

$$180 = 150 + 30$$

Aapje nummer 9 heeft een lengte van 180, terwijl het nul-model 150 voorspelde, waar hij dus nog 30 cm naast zat, de gokfout of error die je maakt in het geval van het nul-model.

$$Y_9 = \bar{Y} + e_9$$

Zijn score (Y_9) is dus opgebouwd uit twee termen, het grote gemiddelde en de error (gokfout) die je in zijn geval maakt.

Als je dus nog niet weet over welk aapje je praat, hou je het *lekker* algemeen (voor je steekproef):

$$Y_i = \bar{Y} + e_i$$

Maar als je over de populatie spreekt moet het natuurlijk aan de hand van parameters:

$$Y_i = \mu_Y + \varepsilon_i$$

Nu aan de hand van het *regressie-model* (M_1) en voor aapje nummer 9:

$$Data = M_1 + Error$$

$$180 = 170 + 10$$

Op basis van het regressie-model zou je zeggen dat aapje nummer 9 – die twee jaar is – dat hij 170 cm zou moeten zijn, en dan hou je nog maar een error over van 10 cm.

$$Y_9 = \hat{Y}_9 + e_9$$

Zijn score (Y_9) is dus opgebouwd uit twee termen, de voorspelde waarde ($\hat{Y}_9$) door M_1 en de error die je in zijn geval maakt. Waar kwam y-dakje ook al weer vandaan?

$$Y_9 = (90 + 40 \cdot X_9) + e_9 \quad \text{of:}$$

$$y_9 = 90 + 40 \cdot x_9 + e_9$$

Als je dus nog niet weet over welk aapje je praat, hou je het *lekker* algemeen:

$$y_i = \hat{y}_i + e_i \quad \text{of:}$$

$$y_i = (90 + 40 \cdot X_i) + e_i \quad \text{Aangezien de haakjes hier voor 'Jan Doedel' staan:}$$

$$y_i = 90 + 40 \cdot X_i + e_i$$

Als je *zelfs* nog niet weet wat de waarden van het *intercept* (b_0) en de *slope* (b_1) zijn in je steekproef:

$$y_i = b_0 + b_1 \cdot X_i + e_i$$

En natuurlijk willen we weten hoe het zit met het *intercept* en de *slope* voor de hele populatie aapjes, dus aan de hand van parameters:

$$y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

Dit noem je het regressie-model voor de populatie, waar we dus precies laten zien, hoe wij een geobserveerde score *zouden* opdelen (als we de waarden voor het *intercept* en de *slope* zouden weten).

Hoe zou de opdeling qua voorspelling (fit) en error (residu) bij de andere aapjes zitten? We gaan het bekijken aan de hand van een tabel. Om tabel 5B aan te maken, zul je dus het spelletje moeten spelen voor *ieder* aapje uit onze steekproef (of heel goed je formules moeten kennen).

TABEL 5B

i	Y_i	X_i	$\hat{Y}_i$	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$	$\hat{Y}_i - \bar{Y}$	$(\hat{Y}_i - \bar{Y})^2$	$Y_i - \hat{Y}_i = e_i$	$(Y_i - \hat{Y}_i)^2 = e_i^2$
1	120	1.0	130	-30	900	-20	400	-10	100
2	130	1.0	130	-20	400	-20	400	0	0
3	140	1.0	130	-10	100	-20	400	10	100
4	140	1.5	150	-10	100	0	0	-10	100
5	150	1.5	150	0	0	0	0	0	0
6	160	1.5	150	10	100	0	0	10	100
7	160	2.0	170	10	100	20	400	-10	100
8	170	2.0	170	20	400	20	400	0	0
9	180	2.0	170	30	900	20	400	10	100
				$\sum$=0	$\sum$= 3000	$\sum$=0	$\sum$= 2400	$\sum$=0	$\sum$= 600

Mijn gedachte nu:

Oh wat een frustratie, natuurlijk voor jou, maar vooral voor mij, want ik wil zo graag naar de essentie en we zijn er bijna. Maar uit ervaring weet ik dat het zo moeilijk is om je aandacht bij *suffe* dingen te houden. Natuurlijk ben ik inmiddels dus bang dat ik je kwijt ben, juist omdat we in een snelle wereld leven waar aandacht *alleen* breed is... Hou vol, nog eventjes, want de kern komt écht bijna.... 'Please, please, please, let me get what I want this time' (The Smiths) 'Lord knows it would be the first time...' We leven immers in een wereld waar je bijna niet meer hoeft na te denken, want alles wordt voor je gedaan.... 'Beam me up Scotty, even if I don't know what *beaming* really is'. Laten we wel wezen, weet jij nog hoe melk gemaakt wordt of hoe je een band moet plakken? Of weet jij écht hoe jouw computer werkt? Meestal niet van belang, omdat *het* – gewoon – werkt, of – voor jou – wordt gedaan, dus waarom zou je nadenken en moeite doen?

Waarom ben je iets wetenschappelijks gaan studeren?

Inzicht Kicks Ass!

Anyway, tabel 5B is de keiharde werkelijkheid, waar we eerst doorheen moeten.

In tabel 5B heb ik een aantal extra kolommen aangemaakt. Een aantal kolommen zou je al moeten herkennen. Kolom 5 en 6 geven respectievelijk de individuele afwijkingen ten opzichte van het grote gemiddelde en de gekwadrateerde afwijkingen, die we ook nodig hadden om de variantie en de standaard afwijking voor lengte (de afhankelijke variabele) uit te rekenen. De gekwadrateerde afwijkingen (tot het grote gemiddelde) tellen samen op tot 3000. In kolom 4 vind je de voorspellingen (qua lengte, Y-dakje dus) op basis van de regressie-vergelijking. Dus op basis van de leeftijd van die aapjes, die je invult in de regressie-vergelijking om hun voorspelde waarde uit te rekenen. In kolom 7 vind je de *systematische verschuiving* in voorspelling *door* het regressie-model (t.o.v. het nul model), de optelling is natuurlijk 0. Daarom kolom 8, maar daar vind je de gekwadrateerde *verschuivingen*, die tellen samen op tot 2400. In de ener laatste kolom vind je de error (e_i), wat de gok-foutjes zijn *als* je dus regressie gebruikt als voorspelling, die tellen (natuurlijk) ook op tot nul. Daarom de laatste kolom met gekwadrateerde errors, die samen optellen tot 600.

Sum of Squares (due to) Total, **SST**

De sommatie van de gekwadrateerde individuele afwijkingen (3000) naar het grote gemiddelde noemen we ook wel *Sum of Squares* (due to) *Total*, of kortweg SST:

$$SST = \sum_{i=1}^{i=n} (Y_i - \bar{Y})^2$$

Sum of Squares (due to) Regression, *SSM*

De sommatie van de gekwadrateerde systematische verschuivingen door het regessie-model (2400) noemen we ook wel *Sum of Squares* (due to) *Regression* (M_1) of kortweg SSREG of SSM:

$$SSM = \sum_{i=1}^{i=n} (\hat{Y}_i - \bar{Y})^2$$

Sum of Squares (due to) Error, *SSE*

De sommatie van de gekwadrateerde afwijkingen van een observatie naar de voorspelling van het regressie-model, of de gekwadrateerde errors (600), noemen we ook wel *Sum of Squares* (due to) *Error*, of kortweg SSE:

$$SSE = \sum_{i=1}^{i=n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{i=n} e_i^2$$

Nu hebben we dus *drie soorten* sommaties van kwadraten uitgerekend. SST, SSM en SSE. Laten we die stelling van daarnet er nog eens bijpakken:

$$DATA = FIT + RESIDU$$

Kan die *al omvattende* stelling niet toevallig ook worden uitgedrukt in termen van de drie verschillende vormen van *de sum of squares*?

Ja!

$$SSTotal = SSModel + SSError$$

Of nog schoner (zonder woorden):

$$\sum_{i=1}^{i=n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{i=n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{i=n} (Y_i - \hat{Y}_i)^2$$

Wel zo fijn voor het marsmannetje. Maar in het bijzondere geval van onze aapjes bedoelen we natuurlijk gewoon:

$$3000 = 2400 + 600$$

Wat je nu ziet is een opdeling van de *totale* variatie (SST) in lengte, in een gedeelte verklaard (SSM) en een gedeelte onverklaard (SSE). SST is ook een maat voor (totale) variatie, net zoals de variantie en de standaard afwijking. Het is alleen niet de gemiddelde oppervlakte van een vierkantje (zoals de variantie dat wel is), maar puur de optelling van al die blauwe vierkantjes, dus hun totale oppervlakte, Bij ons dus een oppervlakte van 3000 cm^2. Denk maar aan een tuintje vol van (totale) variatie (verschillende plantjes, bloemen, onkruid, aarde) met een oppervlakte van 3000 vierkante centimeter. In dat tuintje is een gedeelte strak georganiseerd (rijtjes met prachtige bloementjes, SSM) en een gedeelte is wildgroei (random onkruid, SSE). Welk *gedeelte* van dit tuintje (totale variatie) is strak georganiseerd? Aangezien het strakke gedeelte bij elkaar een oppervlakte had van 2400 cm^2, beslaat dat dus vier-vijfde deel van het hele tuintje, want:

$$\frac{SSM}{SST}$$

Gedeelte systematische variatie ten opzichte van totale variatie, bij ons dus:

$$\frac{2400}{3000} = \frac{4}{5} = .80$$

Proportioneel gezien, neemt het systematische of het verklaarde gedeelte dus een proportie van .80 (of 80 procent) in beslag.

Ik kwam op een eiland waar ik een werkelijkheid tegen kwam, die mijn pet, in eerste instantie te boven ging. Ik zag een boel aan (totale) variatie qua lengtes, bij die schattige aapjes op dat eiland. In eerste instantie kon ik die variatie niet thuis brengen en moest ik accepteren dat de gemiddelde gok-fout niet lager kon dan 19.36 cm. Eigenlijk onacceptabel. Het vermoeden rees, dat die variatie, echt niet *zomaar*, of puur random was. Sterker nog, ik kreeg zelfs het idee, dat deze variatie op lengte wel eens kon samenhangen met een andere variabele, namelijk leeftijd. Ook op leeftijd was er namelijk spreiding en het bleek dat jonge aapjes over het algemeen wat kleiner zijn dan oudere aapjes (r_{xy} = .89). Vanwege deze samenhang, heb ik besloten om hun lengtes aan de hand van een regressie-vergelijking te voorspellen ($\hat{Y}_i$ = 90 + 40·X_i). Dit lukte zo goed, dat ik zelfs in staat was om maar liefst tachtig procent van de totale variatie op lengte te verklaren, enkel en alleen op basis van de variatie op leeftijd!

Grand Final. [on page 100, fully unexpected]

(Proportion of) Variance Accounted For, *VAF*

We hebben nu de belangrijkste maat binnen de statistiek (wetenschap) gevonden. Wanneer je een model bouwt, dus een idee over de werkelijkheid hebt, zijn we vooral geïnteresseerd in de mate waarin dat model overeenkomt met die werkelijkheid, de geobserveerde wereld. De maat die we hebben gevonden, noemen we de proportie verklaarde variantie of ook wel de (*Proportion of*) *Variance Accounted For* (VAF). Het lullige (hehe) is dat je deze waarde ook vindt, door – simpelweg – de correlatie (r_{xy}) te kwadrateren. Maar waar het mij om ging is dat je nu gezien hebt, *hoe* we naar de wereld kijken en hoe we verklaarde en onverklaarde zaken van elkaar scheiden en daar maten aan toe kennen. Resumé, met de juiste symbolen:

$$R^2_{yx} = \frac{SSM}{SST}$$

R-kwadraat noemen we dus ook wel de *proportie verklaarde variantie* (de VAF). Ik gebruik hier een hoofdletter omdat ik het over een *model* heb. Maar dit is hetzelfde als:

$$r^2_{yx} = r^2_{xy} = \frac{SSM}{SST}$$

Voor onze aapjes:

$$VAF = \frac{2400}{3000} = .80$$

80 procent van de totale variatie in lengte scores kan verklaard worden op basis van (variatie in) leeftijd, we houden dus 20 procent onverklaarde variatie over:

$$1 - R^2_{yx} = 1 - \frac{SSM}{SST} = \frac{SSE}{SST}$$

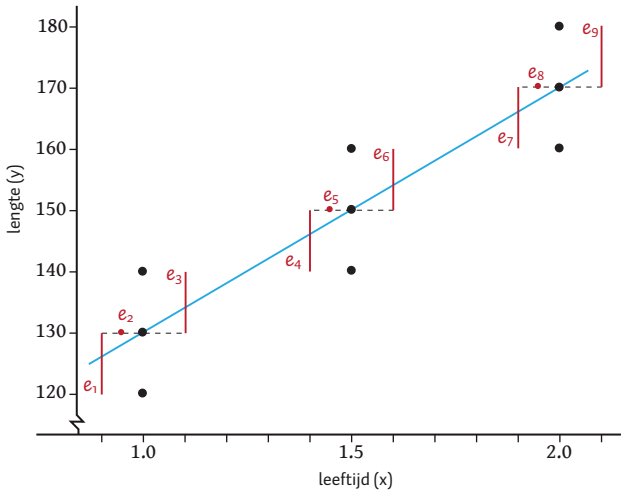
Dit is de proportie onverklaarde variantie, het gedeelte van de totale variatie in lengtes scores, dat we niet konden verklaren, voor onze aapjes:

$$1 - .80 = .20$$

Gemiddelde error

Op basis van het nul-model hebben we te maken met een gemiddelde gok-fout van 19.36 cm, de standaard afwijking. Maar hoe zit het met de gemiddelde gok-fout die we overhouden als we het regressie-model gebruiken. SSE is een maat voor de hoeveelheid error en we hebben de individuele errors gebruikt om SSE te berekenen. Een individuele error (e_i) is ook wel de (verticale) afstand van een observatie naar de regressie-lijn. Dit zijn dus de gokfoutjes die je overhoud als je de regressie-lijn gebruikt om een voorspelling te doen.

figuur 5E



De regressie-lijn hebben we zodanig getrokken dat al die individuele errors (error-termen), gemiddeld gezien, zo klein mogelijk zijn. Dat was ook het doel, een model bouwen, zodanig dat we beter kunnen voorspellen. Wat is de gemiddelde waarde van al die negen error-termen? Ik vraag hier dus naar de standaard afwijking voor de errors. Natuurlijk gaat dat via de variantie voor de error (s_e^2):

$$s_e^2 = \frac{SSE}{n-p-1}$$

De formule voor variantie van de error. waarbij *p* voor het aantal predictoren staan, bij ons alleen de onafhankelijke variabele leeftijd, *p* is 1.
Anders geformuleerd:

$$s_e^2 = \frac{\sum_{i=1}^{i=n} (Y_i - \hat{Y}_i)^2}{n-p-1} \qquad \text{of:}$$

$$s_e^2 = \frac{\sum_{i=1}^{i=n} e_i^2}{n-p-1} \qquad \text{invullen geeft:}$$

$$s_e^2 = \frac{600}{9-1-1} = 600/7 = 85.71429 \qquad \text{Dus de error variantie heeft een waarde van 85.71}$$

Wij willen natuurlijk de standaardafwijking voor de error:

$$s_e = \sqrt{s_e^2} = \sqrt{85.71429} = 9.258201$$

De standaardafwijking van de error heeft dus een waarde van 9.26 (cm)

Nu weten we dus eindelijk hoe goed we kunnen voorspellen aan de hand van het regressie-model in termen van gemiddelde gokfout. We hebben dus nu nog maar te maken met een gemiddelde gokfout van 9.26. Dit is dus ook meteen de gemiddelde afstand van een observatie naar de regressie lijn, de gemiddelde error waarde. We hebben dus de gemiddelde gokfout van 19.36 teruggebracht naar een waarde van 9.26, ruim 10 punten kleiner dus. En dit allemaal aan de hand van een predictor, de variabele leeftijd.

Dus twintig procent van de variatie in lengtes kon niet verklaard worden. Misschien als we een *nog* complexer model bouwen dat we dan nog minder error (SSE) overhouden. Misschien zou een verklarend model op basis van leeftijd én vitamientjes wel meer dan tachtig procent kunnen verklaren! Multipele Regressie-Analyse is het onderwerp van het volgende hoofdstuk. Maar

eerst nog de generalisatie van ons (gevonden) enkelvoudig regressie-model naar de populatie. We mogen dan een leuk effect hebben ontdekt, maar zijn onze resultaten ook significant?

5§3 Betrouwbaarheids-intervallen en Significantietoetsen voor Regressie-Parameters.

We hebben nu het belangrijkste begrip, de proportie verklaarde variantie (VAF) van een model behandeld. Bij ons geeft de VAF dus aan welk deel van de totale variatie in lengte-scores, verklaard kan worden op basis van variatie in leeftijd-scores. Aan de hand van leeftijd, konden we 80 procent van de variatie in lengte verklaren. Deze waarde geldt in ieder geval voor onze steekproef, maar we gaan nu na in hoeverre we onze regressie-vergelijking kunnen generaliseren naar de populatie. In hoofdstuk 4 hebben we gekeken naar de generalisatie van steekproef-gegevens (statistieken) naar populatie-gegevens (parameters) aan de hand van betrouwbaarheidsintervallen en significantie-toetsen. Gaan we nu weer doen, maar dan dus voor onze regressie-vergelijking. Alleen de *slope* is een maat voor verband niet het *intercept*.

Betrouwbaarheidsinterval (BI) voor de populatie-slope.

Om een betrouwbaarheidsinterval (BI) te bouwen (berekenen), heb je *nooit* een H_0 verwachting nodig. Ik laat het in stappen zien:

Stap 1 In onze enige echte steekproef hebben we een slope gevonden met een waar van $b_1 = 40$ gebaseerd op een steekproef van $n = 9$. Stel dat je nu nog veel meer steekproeven zou nemen. Vervolgens reken je voor al die steekproeven de waarde van hun slope uit, zeg duizend keer (liefst nog vaker natuurlijk). Elke keer als je een steekproef neemt, is de verwachting dat je een waarde rond de 40 zal vinden (want dat vond je ook in je enige echte steekproef). Dus al die duizend *slopes* liggen qua waarde rond de 40, de één wat dichter bij dan de andere. Door steekproef-fluctuatie zal de waarde dus rond 40 variëren. Om te weten met hoeveel variatie de slope zal variëren, kun je de standaard error (voor de *slope*) uitrekenen.

$$SE_{b_1} = \frac{s_e}{\sqrt{(n-1)\cdot s_x^2}}$$

Met de volgende gegevens $s_e^2 = 85.7143$ $s_x^2 = .1875$ en $n = 9$ kunnen we de formule invullen:

$$SE_{b_1} = \frac{85.7143}{\sqrt{8*.1875}} = 9.2582 / \sqrt{1.5} = 7.56$$

De waarde van de standaard error voor de slope is dus 7.56 (cm). Elke keer als je een steekproef trekt, is je beste gok voor de *slope* 40, maar al die hypothetische steekproef-*slopes* zullen daar gemiddeld 7.56 cm naast zitten. De standaard error voor de *slope*, stelt niks anders voor dan gemiddelde gok-fout voor de *slope*, als je dus het spelletje speelt en telkens 40 gokt, omdat dat je verwachting is.

Stap 2 Om de *lower* en *upper bound* voor de ware slope (dus β_1 voor de populatie) uit te rekenen, moeten we nog twee dingen weten. We moeten eerst bepalen welk betrouwbaarheidsniveau (*confidence level CL*) we willen en daarna kunnen we in de t-tabel opzoeken hoeveel standaard errors (*t*-waarde) we naar links, en naar rechts moeten (de *margin of error*) om bij de grenzen te komen. Standaard gebruiken we het 95 procent betrouwbaarheidsniveau of *confidence level* (*CL*)(tenzij anders vermeld). In ons geval moeten we de t-waarde opzoeken voor 7 vrijheidsgraden ($df = n - p - 1 = 7$, waarbij p voor het aantal predictoren staat, bij ons alleen de variabele leeftijd, dus 1). Als je de t-waarde opzoekt, zul je in dit geval uitkomen op $t^*_7 = 2.365$. Dit betekent dat we 2.365 standaard errors naar links, en naar rechts moeten (vanuit onze beste puntschatting $b_1 = 40$) om bij de *lower* en *upper bound* te komen. In formule-vorm:

$$Lower\ Bound = b_1 - t^*_{df} \cdot SE_{b_1}$$

met $df = n - p - 1$ (het aantal vrijheids graden voor de error)

$$Upper\ Bound = b_1 + t^*_{df} \cdot SE_{b_1}$$

invullen met onze gegevens:

$$Lower\ Bound = 40 - 2.365 \cdot 7.56 = 22.12$$

$$Upper\ Bound = 40 + 2.365 \cdot 7.56 = 57.88$$

Stap 3 Nu kunnen we tot een conclusie overgaan. We hebben één steekproef getrokken en vonden een waarde van b_1 = 40 en we kunnen nu stellen dat bij steekproef-herhaling, we een waarde voor de slope tussen de 22.12 en de 57.88 kunnen verwachten met 95 procent zekerheid. En op basis hiervan kunnen we nu stellen dat de ware slope β_1 voor de populatie een waarde heeft tussen de 22.12 en de 57.88, met 95 procent zekerheid. Deze laatste interpretatie, net zoals bij de interpretatie voor een BI voor het gemiddelde, mag eigenlijk niet zo strikt gedaan worden, maar in de praktijk zie je het bijna niet anders.

Besef dat we *niet* een H_0 verwachting of idee nodig hadden om dit 95 % betrouwbaarheidsinterval [22.12, 57.88] te construeren. Maar het handige van zo'n betrouwbaarheidsinterval, is dat op basis van dit interval, we toch een uitspraak kunnen doen over de waarschijnlijkheid van de H_0. Onder de H_0, verwachten we dat er geen verband is en dat je dus geen lineaire regressie kan toepassen, om beter te voorspellen. Dit betekent ook wel dat voor elke waarde van x (leeftijd), de gemiddelde waarde op y (lengte), altijd hetzelfde is (bij ons dus altijd 150 cm en dat het niet uit maakt of een aapje jonger of ouder is). In dit geval kan je dus ook zeggen dat de regressielijn dan horizontaal zou moeten lopen, want dan zou de voorspelling qua y, toch voor elke waarde van x, hetzelfde zijn. Die lijn zou dus een slope van 0 (moeten) hebben, want voor elke toename van 1 punt op x, gaat de lijn *niet* omhoog of omlaag. Technisch gezien:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Als de H_0 waar *zou* zijn, en je *zou* dan een steekproef nemen en de *slope* uitrekenen, welke waarde zou je dan verwachten? Als er geen verband is, zou je natuurlijk (idealiter of ongeveer) b_1 = 0 moeten vinden. Kom je die H_0 verwachting tegen als je van je *lower bound* naar *upper bound* wandelt? Nee! En wat je *niet* tegen komt op je betrouwbaarheidsinterval, is onwaarschijnlijk! Omdat de waarde 0 niet op ons interval ligt, kunnen we dus de H_0 verwerpen! We kunnen dus aannemen dat de werkelijke waarde voor de *slope* écht van nul afwijkt en dat er dus wel een positief verband is, de lijn gaat echt omhoog, ook in de populatie (of de *slope precies* een waarde van 40 heeft, weten we natuurlijk niet).

Significantietoets voor de populatie *slope*.

Je kan ook direct de *slope* (b_1) toetsen, door een t-toets voor de slope uit te voeren. In dit geval bereken je eerst de toets-statistiek t en vervolgens zoek je de bijbehorende p-waarde (significantie) op in t-tabel (statistiek programma's doen dit normaal voor je). Aan de hand van je *p*-waarde kun je een beslissing maken over de waarschijnlijkheid (of houdbaarheid) van de H_0. Laten we het maar meteen doen.

Stap 1 Stel H_0 en H_1 op:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Stap 2 Bereken de standaard error voor b_1

$SE_{b_1} = 7.56$ Dit hadden wij al gedaan !

Stap 3 Bereken bijbehorende toets-statistiek *t en* zoek de tweezijdige overschrijdingskans op in de *t*-tabel.

$$t = \frac{b_1}{SE_{b_1}}$$

Hier staats de t-waarde dus voor het *aantal* keer dat de standaard error tussen de nul-verwachting ($b_1 = 0$) en de berekende waarde ($b_1 = 40$) past. Invullen geeft:

$$t = \frac{40}{7.56} = 5.29 \qquad \text{met } df = n - p - 1 = 7$$

Opzoeken in t-tabel geeft:

$$.0005 < p < .0010 \qquad \text{éénzijdige overschrijdingskans.}$$

En voor tweezijdig, verdubbel je beide waarden:

$$.0010 < p < .0020 \qquad \text{tweezijdige overschrijdingskans.}$$

Stap 4 Nu kunnen we weer tot een conclusie overgaan: de tweezijdige *p*-waarde ligt tussen de twee waarden .001 en de .002 en dat is behoorlijk klein! Qua interpretatie kunnen we dus zeggen dat de kans om een extremere waarde (absoluut gezien) te vinden, dan wij hebben gevonden ($b_1 = 40$) extreem klein is (gegeven dat de H_0 waar zou zijn). Dus we *schrikken* een beetje van onze vondst gegeven de H_0. Of korter (het gaat natuurlijk om de beslissing): Is de gevonden *p*-waarde kleiner dan een $\alpha = .05$? Ja! Dus de H_0 moet verworpen worden en de alternatieve hypothese is hiermee aannemelijk. De populatie-slope wijkt wel af van nul en de regressielijn kruipt dus, lineair omhoog. *Precies* dezelfde conclusie als bij het betrouwbaarheidsinterval dat we hiervoor gedaan hebben. Dit is altijd zo. Een twéézijdige significantietoets met een α van .05, komt qua conclusie, volledig overeen met een BI met een *confidence level* van .95 ($1 - \alpha = .95$) . Als je conservatiever wilt zijn en dus strenger wilt toetsen, bijvoorbeeld met een α van .01, zou je dus ook een BI kunnen gebruiken met een *CL* van .99.

Toetsing van het Intercept

Je ziet dat ik het hier helemaal niet heb over het intercept. Nogmaals het intercept zegt niks over de helling van de regressie-lijn en dus ook niks over het verband. Het intercept staat enkel en alleen voor de voorspelde waarde van y voor die cases (aapjes) die 0 scoren op x, precies nul jaar zijn dus. Bij ons was de waarde voor het intercept, $b_0 = 90$. Met andere woorden dit suggereert dat aapjes die net geboren zijn ongeveer 90 cm zouden moeten zijn (lijkt me ietwat groot). En bedenk, dat in onze steekproef, we niet eens aapjes hadden die nul jaar waren. Uitspraken doen over de voorspelde lengte, buiten de range van geobserveerde waarden van x, lijkt me een beetje arrogant en dom. Je zou eerst moeten onderzoeken of het verband tusen x en y daar ook geld. (volgens mij groeien aapjes als ze net geboren zijn als kool, en daarna gaat het wat langzamer). Afgezien hiervan kan je wel toetsen of het intercept significant afwijkt van nul, maar waarom zou je? Wil je bewijzen dat aapjes echt langer zijn dan 0 cm als ze geboren zijn? Statistiek programma's geven in de *output*, bijbehorende significanties en vaak ook betrouwbaarheidsintervallen, maar we doen er dus weinig mee in de praktijk.

Toetsing van het regressiemodel aan de hand van een ANOVA tabel.

In het volgende hoofdstuk behandel ik nog een manier om een regressiemodel te toetsen, aan de hand van een ANOVA tabel. ANOVA staat voor Analysis of Variance, en kijkt specifiek of het verklaarde gedeelte van de variatie in Y scores significant van nul afwijkt, aan de hand van de kwadratensommen die we eerder hebben uitgerekend. Dit is vooral handig als je meerdere predictoren in je regressiemodel hebt. Omdat je dan dan *eerst* het model als geheel wilt toetsen (ANOVA) en vervolgens verder wilt kijken, per predictor, of de slope afwijkt van nul. We hebben er nu maar 1 dus volstaat wat we tot zover gedaan hebben. ANOVA dus later!

Rapporteren.

Om te onderzoeken of de lengte van aapjes voorspeld kan worden op basis van leeftijd is een lineaire regressie-analyse uitgevoerd, met lengte als afhankelijke variabele en leeftijd als predictor.
Leeftijd voorspelde significant de lengte scores (b_1 = 40), $t(7) = 5.29$, $p = .001$. We kunnen dus concluderen dat voor twee aapjes die één jaar schelen, de oudste ongeveer 40 cm groter is.

5§4 Assumpties over de Populatie.

Mag je altijd je p-waarde geloven en op basis daarvan een beslissing nemen? Nee, nee, nee, dat mag niet zomaar! voordat je je p-waarde gebruikt om een conclusie te trekken moet je eigenlijk altijd eerst nog een aantal zaken checken.

Stel je voor dat ik bij jou een kopje thee wil drinken. Dan ga ik naar jou huis en bel aan (aanbellen is een toets uitvoeren). Stel dat er niemand open doet, mag ik dan meteen concluderen dat je niet thuis bent? Mischien doen we dat wel in de praktijk, maar zou het niet handig zijn om eerst even te kijken (checken) of je niet toevallig de muziek op tien hebt staan? Het zou kunnen en in dat geval zou je de bel misschien niet horen. Stel dat ik geen muziek uit je huis hoor komen, maar wel het licht zie branden (en je hebt nog steeds niet open gedaan), kan ik bijvoorbeeld concluderen dat je me niet aardig vindt, maar misschien heb je wel gewoon een koptelefoon op. Het zou toch wat zijn als ik met een onterechte conclusie weer naar huis ga... Als ik bijvoorbeeld twijfel over de muziek, zou ik ook hard op je ramen kunnen bonken, zodat je me dan wel hoort. Wat ik hier mee wil zeggen, zijn eigenlijk twee dingen.
Wanneer je een significantie toets uitvoert en interpreteert, zul je *eerst* moeten checken of aan bepaalde (rand-) voorwaarden voldaan is, of je assumpties (of aannames) *over* de *populatie*, waaruit jij een steekproef hebt getrokken, dus kloppen of aannemelijk zijn. Mocht dit niet het geval zijn, kan je niet zomaar je resultaten (schattingen van parameters, toets-statistieken en *p*-waarden) serieus nemen en had je misschien een ander soort toets of analyse moeten doen, die wel opgewassen (robuust) is tegen rariteiten in de populatie. Dit zijn *robuustere* toetsen (die bestaan, de zogenaamde non-parametrische toetsen of *bootstrap*-methodes, heel leuk trouwens, maar daar ga ik nu niet op in).

De 4 Aannames die je officieel moet checken voordat je je regressie-analyse interpreteert:

1. Onafhankelijk van Observaties

Dit betekent dat alle mogelijke cases in de populatie (maar dus ook in je steekproef) elkaar niet mogen beïnvloeden (of samenhangen). Deze aanname gaat meer over hoe je je steekproef hebt getrokken (qua onderzoekopzet) en is niet echt een statistische aanname. Als een aapje een broertje is van een ander aapje binnen onze steekproef dan zou het goed kunnen zijn dat als de één heel groot is (en dus boven de regressie-lijn zit qua observatie), de ander er dus *ook* boven zit (omdat ze genetisch gerelateerd zijn). Je zou in dit geval kunnen zeggen dat hun error-termen (de afstanden van een observatie naar de regressielijn) gecorreleerd zijn (omdat ze vanwege hun genen dus allebei een positieve error hebben (boven de lijn zitten). Het belangrijkste punt is hier dat wanneer je een steekproef neemt je goed oplet of respondenten elkaar niet beïnvloeden. Bijvoorbeeld als je respondenten een vragenlijst invullen, dat ze niet van elkaar afkijken, want dan zijn de antwoorden (dus ook de errors) aan elkaar gerelateerd. Gewoon goed opletten tijdens je onderzoek en kijken of de metingen, het afnemen van je vragenlijsten en dergelijke, dus eerlijk gebeuren. Deze aanname van onafhankelijke observaties wordt eigenlijk regelmatig geschonden, maar weinig onderzoekers die zich er echt om bekommeren. Zoals in het geval naar studies naar leerprestaties van leerlingen van verschillende klassen binnen één school, waar de leerlingen binnen één klas zich soms systematisch anders gedragen dan de leerlingen binnen één andere klas. Leerlingen binnen een klas kunnen elkaar beïnvloeden. Of hun leraar is meer gemotiveerd en daarom slaat de lesmethode beter aan dan bij andere leraren. Natuurlijk zijn

er statistische analyses die dit soort problemen (afhankelijkheid van observaties) opvangen, zoals *Multi-Level Regression Analysis*. Maar dat zijn onderwerpen voor een research-master. Mijn advies, blijf opletten, maar wees niet te streng.

2. Lineariteit

Er zijn talloze vormen van verband, wij kijken specifiek naar rechtlijnige (lineaire) verbanden in dit hoofdstuk. Het is dan altijd zaak om even aan de hand van een scatterplot (of via ingewikkeldere manieren) te kijken hoe de puntenwolk eruit ziet. Natuurlijk komt het maar zelden voor dat een (populatie) puntenwolk echt prachtig rechtlijnig is. Soms lijkt een wolk meer op een HEMA-worst en dan is een rechte-lijn, als regressie, geen afspiegeling van wat er echt gebeurt en dus onjuist (al zou het zomaar beter kunnen zijn dan het nul-model en kun je dus toch centjes verdienen met je lineaire regressie-model). Maar er zijn ook regressie-modellen die wel rekening houden met kromlijnige (*curve linear*) verbanden tussen variabelen. Soms, als je eerst één of meer variabelen transformeert (verandert door *bijvoorbeeld* eerst de wortel te nemen van elke score) en dan de getransformeerde variabel(en) uitzet in een grafiek, ziet het verband er soms wel weer rechtlijnig uit. Dit is vaak de meest makkelijke oplossing, maar ook hier zijn natuurlijk weer andere oplossingingen voor door kromlijnge regressie-analyses toe te passen op je data.

3. Homoscedasticiteit

We willen heel graag dat de voorspelling qua y-waarde, de afhankijkelijke variabele dus, voor iedere waarde van x (dus over de hele range van de predictor, die je geobserveerd hebt) ongeveer even goed gaat. Stel dat bij jonge aapjes er minder spreiding qua lengte is dan bij oudere aapjes. Dit zou dan betekenen dat je de lengtes van jonge aapjes preciezer zou kunnen voorspellen dan bij oudere aapjes. Bij jongere aapjes zou in dit geval de error, gemiddeld kleiner zijn en dat is niet eerlijk! Graag willen we dus dat de spreiding rond de regressie-lijn dus voor iedere waarde van de predictor, dus ongeveer gelijk is. Bij schending van homoscedasticiteit (dus bij heteroscedasticiteit), zou *ik* het voor nu alleen even melden in mijn rapportage, maar me er niet te druk over maken. Mocht er sprake zijn van ongelijke spreiding rond de regressielijn, kan dit ook duiden op mogelijke andere (interactie) effecten die ook een rol kunnen spelen in de verklaring van Y. Dus misschien is het zaak om een multiple-regressie uit te voeren, dus een voorspel-model op basis van meerdere predictoren.

4. Normaal verdeelde residuen (errors or error-termen)

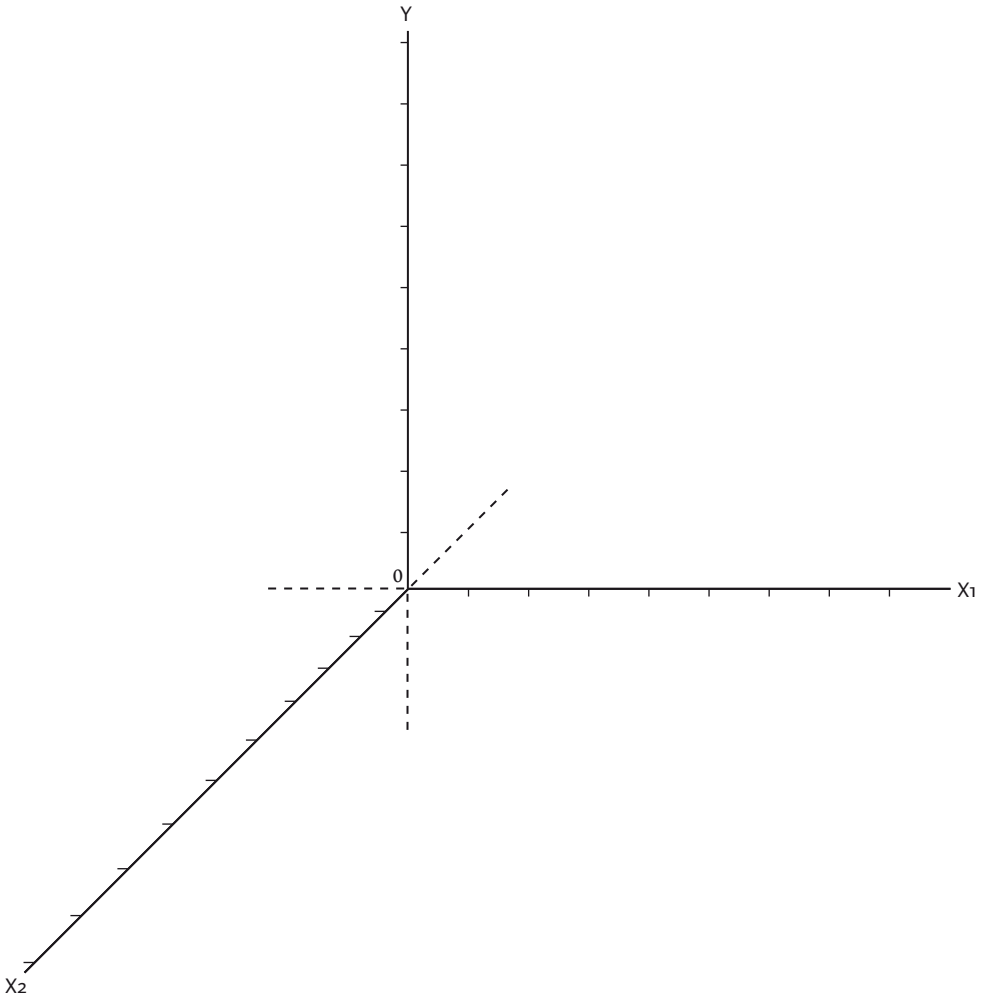
De errors of residuen in de populatie (ε_i) zouden normaal verdeeld moeten zijn (met een gemiddelde van nul en een standaardafijking σ_e). Waar dit op neer komt is dat voor elke waarde van de predictor, er dichtbij de regressielijn wat meer observaties zitten dan als je je wat meer van de regressielijn af beweegt (maar dan dus wel verticaal gezien). Er zijn verschillende manieren om dit te checken, bijvoorbeeld door een histogram te maken van alle residuen in je steekproef. Als de vorm van de verdeling van de residuen niet te veel afwijkt van normaliteit zijn we blij en zeggen we dat het er goed uit ziet. In de populatie zullen de residuen dan ook wel normaal verdeeld zijn. Mocht het heel veel afwijken van normaliteit (een normaal verdeling), bijvoorbeeld bij een hele scheve verdeling, dan zou je bijvoorbeeld weer aan een transformatie van een (aantal) variabele(n) kunnen denken. Maar voor nu is dat buiten de scope van deze handleiding!

Meervoudige Lineaire Regressie-Analyse.

6§0 Meer is beter.

Nu we enkelvoudige lineaire regressie-analyse achter de rug hebben, heb je de basis overwonnen. Echt heel veel zaken in de statistiek zijn eigenlijk alleen maar generalisaties van wat je in hoofdstuk 5 geleerd hebt. In dit hoofdstuk gaan we proberen om een nog beter model te bouwen door één of meer predictoren toe te voegen aan ons model. We krijgen dus een complexer model dan een enkelvoudig regressie-model. Complexere modellen, dus met meer predictoren, voorspellen over het algemeen beter dan eenvoudigere modellen. Althans, dit is waar, onder twee voorwaarden. Als eerste moet het eenvoudigere model niet al honderd procent van de te verklaren variabele verklaren, want dan is er geen verbetering meer mogelijk. En als tweede geldt dit alleen altijd voor je steekproef. Dus complexere modellen kunnen bijna altijd de afhankelijke variabelen beter verklaren binnen je steekproef, maar of die verbetering in voorspelkracht ook generaliseerbaar is naar de populatie is nog maar de vraag. De verbetering in voorspelling is alleen generaliseerbaar naar de populatie als de toename in proportie verklaarde variantie, ook daadwerkelijk significant is. De variatie in lengte scores van onze aapjes konden wij voor 80 procent verklaren aan de hand van verschillen in hun leeftijden. Dus M_1 (het enkelvoudige regressie-model) verklaarde 80 procent *meer* dan M_0, het nulmodel. Het nulmodel (het grote gemiddelde qua lengte) verklaart natuurlijk helemaal niets van de variatie, nul procent dus. Wij gaan M_2 bouwen, een multipel regressie-model met twee predictoren, namelijk leeftijd en het aantal bananen dat een aap per dag eet. Ik noem deze tweede variabele kortweg 'banaan'. Voor *onze* aapjes zal dit model dus wel beter zijn dan het model dat alleen gebaseerd is op de variabele leeftijd. Maar of het dus ook *significant beter* voorspelt, valt dus nog te bezien en zal getoetst moeten worden. Gaan we doen!

figuur 6A

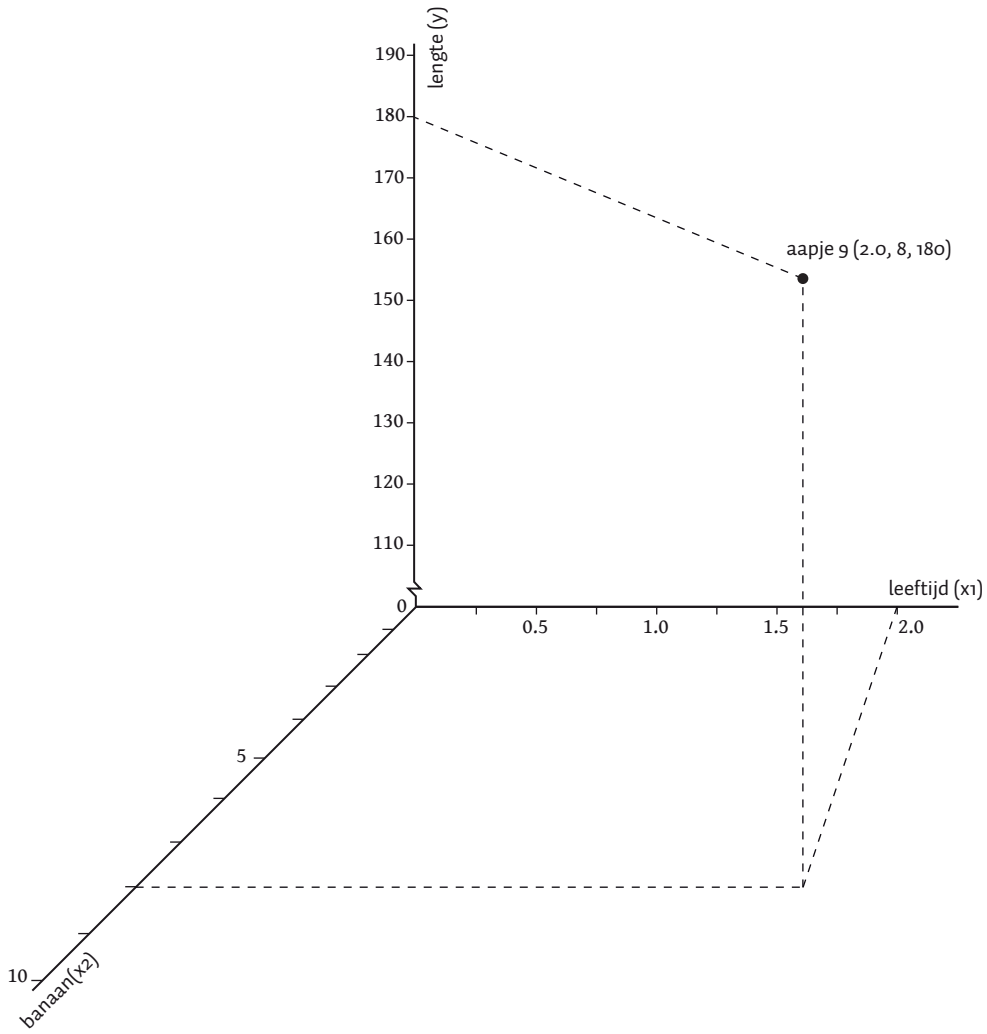


6§1 Van een twee-dimensionale wereld naar een drie-dimensionale wereld.

Tot zover hebben we alleen te maken gehad met twee variabelen, dus twee dimensies. Als we grafisch het verband tussen leeftijd en lengte willen laten zien, hebben we alleen een x en een y as nodig om een puntenwolk te plotten. Om een x en een y as te tekenen, heb je alleen een plat vlak nodig. Het enige wat je in dit geval hoeft te doen, is te beginnen met een willekeurige rechte lijn. Het maakt eigenlijk niet uit waar, of hoe je die tekent, zolang die maar recht is. Dit is je eerste as. Vervolgens teken je een tweede lijn die precies een hoek maakt van 90 graden met de eerste as. Als twee lijnen een hoek van 90 graden met elkaar maken, zeggen we ook wel dat ze orthogonaal staan ten opzichte van elkaar. Als je vervolgens een schaalverdeling kiest (de meeteenheid per variabele), kun je je twee-assig stelsel afmaken en elk punt van je puntenwolk – qua positie – definiëren. Dit doe je door voor elk punt (behorend tot een aapje) de twee bijbehorende coördinaten te geven, dus zijn leeftijd en lengte. Zodra er een derde dimensie bijkomt, moeten we dus eigenlijk van het papier af want we willen dat de derde as ook een hoek maakt van 90 graden met de twee andere assen. Als je aan een kubus denkt ben je dus al een heel eind. Een kubus heeft drie dimensies (lengte, breedte (of diepte) en de hoogte). Een drie-assig stelsel zonder de eenheden per variabele (x_1, x_2, y)zie je in figuur 6A. Je zou dus eigenlijk moeten denken dat de nieuwe as (de x_2 as) uit het papier zou moeten komen. Gelukkig kennen de meesten van ons de 3D-films van de bioskoop en weten we allemaal hoe het voelt als de *suggestie* wordt gewekt, dat er meer is dan een plat beeld. Suggestie doet leven.

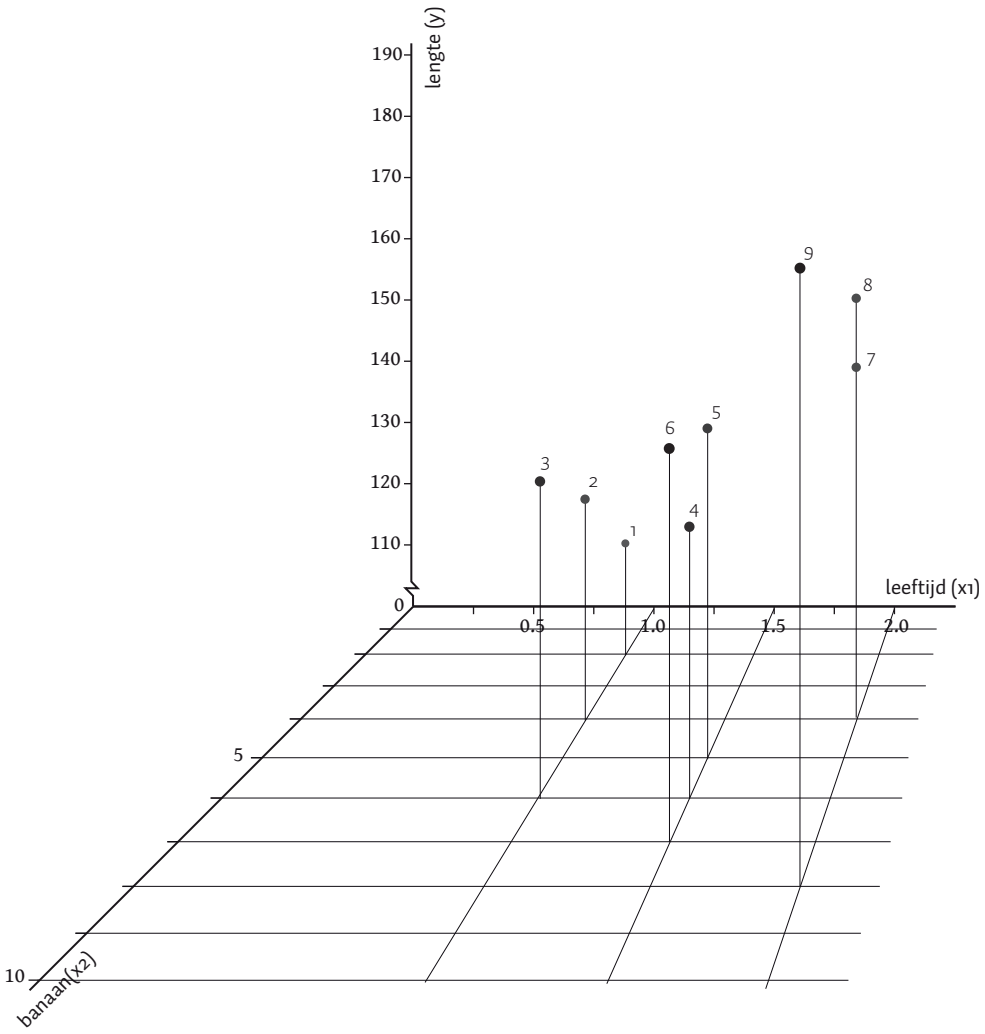
Aan de hand van ons drie assig stelsel, kunnen we dus weer elk punt een positie geven in onze drie-dimensionale wereld als we weten wat een aapje, op de drie variabelen scoort. De predictor 'leeftijd' kennen we toe aan de x_1-as, de tweede predictor, 'banaan', aan de x_2-as, en de afhankelijke variabele 'lengte' aan de y-as. In figuur 6B heb ik ook vast het punt voor aapje nummer 9 gezet, (2.0, 8, 180).

figuur 6B



In figuur 6C heb ik alle bijbehorende punten gezet, een echte drie-dimensionale puntenwolk dus (de *suggestie* is echt, want het is nog steeds op een plat vlak getekend!).

figuur 6C



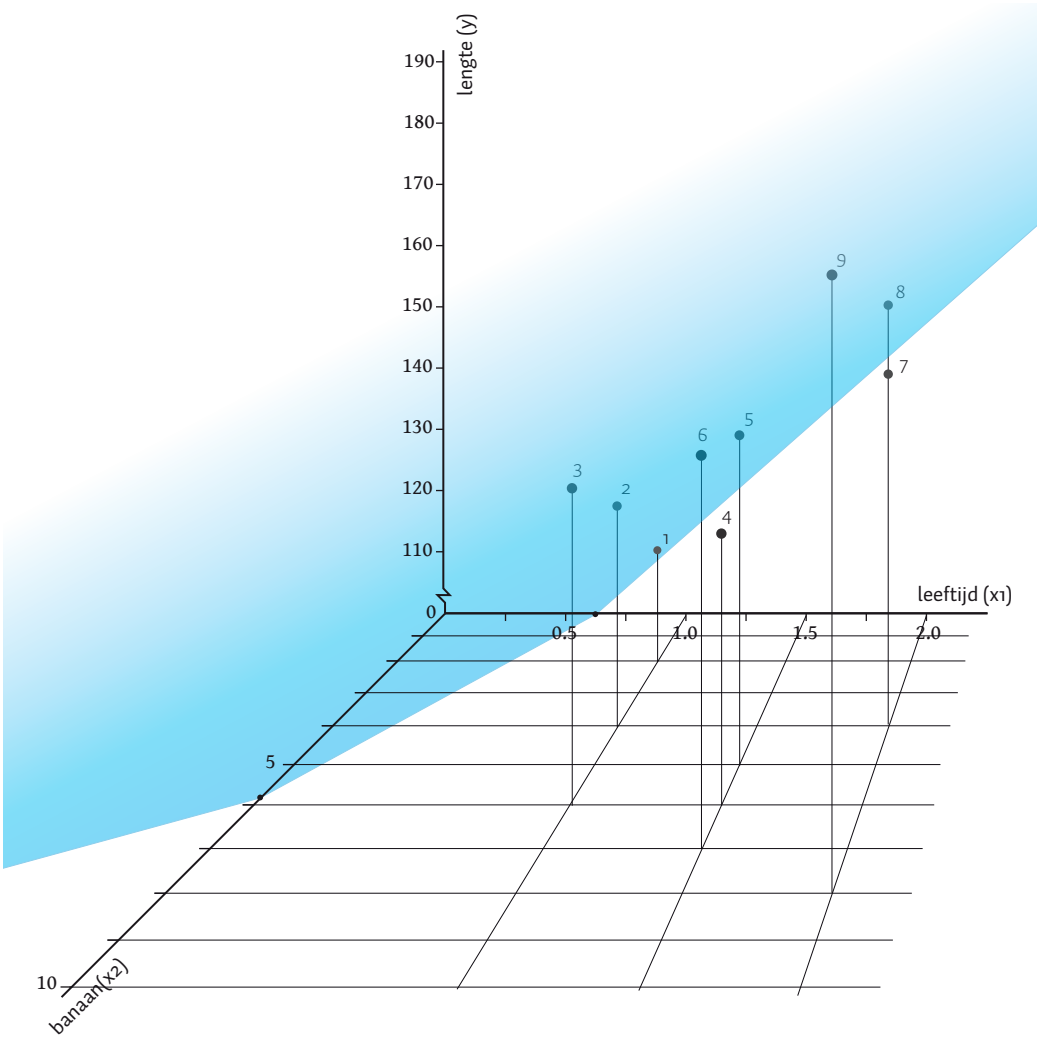
Omdat we nu met meerdere voorspellers te maken hebben, ziet het regressie-model er niet meer uit als een rechte lijn door een platte puntenwolk (de regressielijn). In het geval van twee predictoren hebben we te maken met een regressie-vlak, zie figuur 6D.

Dit vlak moet zo dicht mogelijk bij elk puntje liggen, of omgedraaid: alle observaties moeten, gemiddeld gezien, zo dicht mogelijk bij het regressie-vlak liggen. Net zoals bij enkelvoudige regressie-analyse, gaat het er om dat de verticale afstand van een observatie naar het regressie-vlak – een residu of error dus – gemiddeld gezien, zo klein mogelijk is, zie figuur 6E (pagina 112).

Mocht je nog meer predictoren toevoegen aan een model, dan wordt het wel heel moeilijk om daar nog een grafische voorstelling van te maken. In 4 (of meer) dimensionale ruimtes kunnen we prima berekeningen uitvoeren en dus ook modelleren of voorspellen, maar om er nog een grafische voorstelling van te maken, gaat de meesten boven hun pet en is ook niet nodig. Het gaat immers puur om de voorspelling en dat gaat gewoon aan de hand van een regressievergelijking! Omdat we nu extra voorspellers gebruiken, is de vergelijking alleen wat langer. In het bijzondere geval van twee predictoren kunnen we de voorspelde waarde van lengte ($\hat{y}_i$) als volgt uitdrukken:

$$\hat{Y}_i = b_0 + b_1 \cdot X_{i1} + b_2 \cdot X_{i2}$$

figuur 6D

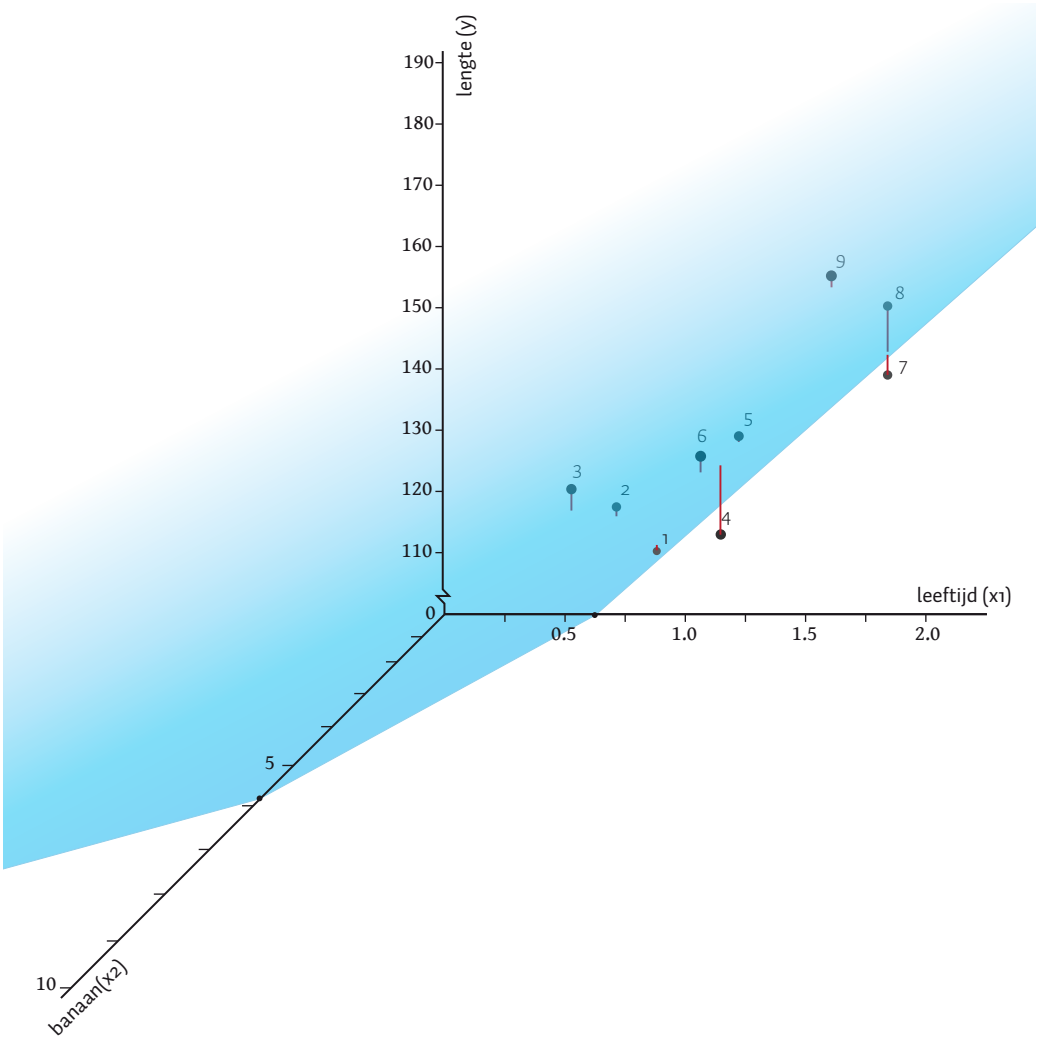


Y-dakje is weer de voorspelde waarde en is dus gelijk aan wat er rechts van het is-gelijk-teken staat. Maar wat staat er eigenlijk aan de rechterkant? Is het een som of een produkt? Het is een Som van drie termen, want we tellen drie dingen op, namelijk het intercept en iets met x_1 en x_2. Het is niet zomaar een som, het is namelijk een *gewogen* som. Een gewogen som kennen jullie allemaal! Want iedereen van jullie kan zijn eindcijfer voor een vak berekenen, ook als jullie verteld wordt dat het tentamen cijfer twee keer zo zwaar meetelt als de inleveropdracht bijvoorbeeld. Ook hier bereken je een gewogen som of een gewogen gemiddelde (is wiskundig gezien precies het zelfde). Aan de 'regressie-gewichten' b_1 en b_2, kun je zien (als je hun numerieke waarden kent) hoe zwaar x_1 en x_2 meetellen voor de voorspelling van y. En mocht je met meer dan twee predictoren te maken hebben, kun je de regressie-vergelijking gewoon uitbreiden:

$$\hat{Y}_i = b_0 + b_1 \cdot X_{i1} + b_2 \cdot X_{i2} + b_3 \cdot X_{i3} \ \ldots\ldots + b_j \cdot X_{ij}$$

Omdat we de letter *i* al hadden weggegeven aan case (of respondent) nummer, nummeren we de predictors, met de letter *j*. En we zeggen ook wel (als je nog niet weet hoeveel predictoren je gaat gebruiken) dat je dus een regressie analyse doet met 'J' predictoren. We geven de slope, behorende bij één bepaalde predictor, het zelfde nummer als dat de predictor heeft gekregen. Omdat wij maar twee predictoren hebben ('leeftijd' en 'banaan') heeft 'J' ook wel de waarde 2. Als je dus een voorspelling wil doen voor case nummer i, dan heb je voor elke predictor (tot en met de J-de, dus voor predictor j=1 en j=2 bij ons) de scores nodig van case i. Die waarden vul je dan in, in de regressievergelijking en dan kun je uitrekenen, wat zijn score op de afhankelijke

figuur 6E



variabele *zou* moeten zijn ($\hat{Y}_i$), de voorspelde waarde. Je zal eerst natuurlijk moeten weten wat de waarden van je regressie-gewichten zijn, anders valt er überhaupt weinig uit te rekenen qua voorspelling.

6§2 De berekening of schatting van regressie-gewichten (… *not*!).

Het berekenen van de regressie-gewichten (het intercept en de *slopes*) gaan we *niet* meer zelf doen, we laten dit geheel over aan de statistiek programma's (natuurlijk kies je dan voor *R* of *RStudio*). Hetzelfde geldt voor de standaard errors voor de regressie-gewichten, je hoeft je om die berekeningen dus niet meer druk te maken. Maar je moet al de uitkomsten (berekende waarden, de schattingen voor de parameters van je model) natuurlijk wel begrijpen! Daarom dus *al* het voorgaande in deze verhandeling. Naarmate we vorderen in de statistiek, zullen we ons dus steeds meer op de output (resultaten en uitkomsten) van zo'n statistiek programma richten. Vervolgens interpreteren we de *output* en op basis daarvan trekken we dan conclusies over de onderzoeksvragen die we proberen op te lossen. Je zult in dit hoofdstuk dus ook steeds *minder* formules zien, omdat ik slechts de resultaten (uitkomsten) zal bespreken.
Eerlijk gezegd, tijdens mijn studiejaren (die ik ontzettend mis), heb ik het vak 'Mathematische Statistiek' gevolgd. Ik vond dat super 'kicke'. Eindelijk werd mij duidelijk dat *echte* statistiek helemaal geen wiskunde A is. Statistiek wordt bijna altijd met wiskunde A geassocieerd, je weet wel, van die eindeloos lange verhaaltjes-sommen, vreselijk. Bij Mathematische Statistiek moest ik – gewoon zoals bij wis B – met een potlood en een gum – laten zien hoe je de oppervlakte onder een curve (lijntje) berekent (en dus niet opzoekt in je z-tabel, of je normalCDF op je 'GR' gebruikt (nee, niet normalPDF), 'integreren' dus. En dat de *slope* niks anders is dan een 'afgeleide functie', ook wel 'dy/dx' voor diegenen die GR-afhankelijk zijn, het lijkt wel de 'Grote vriendelijke Reus tegenwoordig'. Hij mocht dan Grafisch, Vriendelijk en gRoot zijn, maar *inzicht* gaf hij *alleen* aan die mensen, die er 'klaar' voor waren. Kortom, je bent nu *klaar* voor het 'snelle' werk, *omdat*

je de basis overleefd hebt. Dat *hoop* ik natuurlijk, maar dat zal vooral van je aandacht - als belangrijkste predictor - afhangen.
Conclusie: we knallen de data in een programma (het importeren of inlezen van je data-set in je statistiek-programma). Vervolgens vertellen we het programma dat we een regressie-analyse willen doen, welke variabele de afhankelijke variabele is voor ons model en dan welke variabelen we als predictoren willen gebruiken. We doen immers een (*Uni-Variate*) *Multiple Regression Analysis* (MRA). 'Uni-Variaat', omdat we slechts één criterium variabele hebben,'lengte' de afhankelijke variabele).'Multiple', omdat we meer dan één predictor hebben ('leeftijd' en 'banaan', de twee onafhankelijke variabelen). En uiteindelijk 'Regression Analyse', omdat we the variatie in de geobserveerde scores van de variabele Y terug brengen naar $\hat{Y}$. En Y-dakje drukken we dus weer uit in een gewogen som van X-en. Het regressie-model is nog steeds lineair van aard, Y-dakje is een recht vlak in onze 3D grafiek. Dit vlak is zodanig gepositioneerd, dat de verticale afstanden van de observatie naar het regressie-vlak, de residuen, gemiddeld gezien zo klein mogelijk zijn. Het woord 'Variaat' op zich, mag niet onbesproken blijven. De 'Variaat' staat voor $\hat{Y}$ en is dus het regressie-model.
Als je databestand niet te groot is, krijg je binnen één oogblink, de schattingen van je parameters te zien (als je tenminste weet waar je moet zoeken in die output). *Blink* maar, en kijk dan naar tabel 6A.

TABEL 6A **Drie Modellen voor de voorspelling van lengte**

Afhankelijke variable: lengte			*coefficient*	*Std. Error*	*t*	*p*
MODEL 0						
R^2 = .00	intercept only	b_0	150.00	6.46	23.24	< .001
MODEL 1						
R^2 = .80	intercept	b_0	90.00	11.75	7.66	< .001
	leeftijd	b_1	40.00	7.56	5.29	.001
$F(1,7) = 28.00$, $p = .001$						
MODEL 2						
R^2 = .91	intercept	b_0	78.44	9.43	8.31	< .001
	leeftijd	b_1	35.05	5.72	6.12	< .001
	banaan	b_2	3.72	1.35	2.75	.033
$F(2,6) = 30.90$, $p < .001$						

6§2 Interpretatie en evaluatie van een multipele regressie analyse in stappen.

Zoals je ziet, heb ik in tabel 6A ook de twee voorgaande modellen erbij gezet, het nul-model en het model met één predictor (M_1). Wij richten ons nu op het model met twee predictoren (M_2), namelijk het model met 'leeftijd' en 'banaan' als onafhankelijke variabelen. De data waar dit regressie-model op gebaseerd is, vind je in tabel 6B. In de kolom met X_{i2}, vind je nu ook het 'aantal bananen' dat een aapje per dag (gemiddeld) eet, onze tweede predictor dus.

TABEL 6B

i	Y_i	X_{i1}	X_{i2}	$\hat{Y}_i$	$(Y_i - \bar{Y})^2$	$(\hat{Y}_i - \bar{Y})^2$	$Y_i - \hat{Y}_i = e_i$	$(Y_i - \hat{Y}_i)^2 = e_i^2$
1	120	1.0	2	120.92	900	845.76	-0.92	0.84
2	130	1.0	4	128.35	400	468.72	1.65	2.72
3	140	1.0	6	135.78	100	202.15	4.22	17.79
4	140	1.5	6	153.31	100	10.92	-13.31	177.02
5	150	1.5	5	149.59	0	0.17	0.41	0.17
6	160	1.5	7	157.02	100	49.29	2.98	8.87
7	160	2.0	4	163.40	100	179.45	-3.40	11.53
8	170	2.0	4	163.40	400	179.45	6.60	43.61
9	180	2.0	8	178.26	900	798.63	1.74	3.02
					$SST = 3000$	$SSM = 2734.4$	$\sum=0.0$	$SSE = 265.60$

Stap 1 De regressie-vergelijking opstellen en interpreteren.

Voor ons model heeft het intercept (b_0) een waarde van 78.44, de slope voor leeftijd (b_1) nu een waarde van 35.05 en de slope voor banaan (b_2) een waarde van 3.72. In deze tabel staan trouwens *afgeronde* waarden, dus verdere berekeningen met deze waarden, zullen dus niet meer heel precies zijn en net iets anders kunnen zijn dan als we met precieze waarden zouden verder rekenen. Met deze drie (afgeronde) waarden kunnen we de regressie-vergelijking opstellen:

$$\hat{Y}_i = 78.44 + 35.05 \cdot X_{i1} + 3.72 \cdot X_{i2}$$

Het intercept ($b_0 = 78.44$) is de voorspelde waarde voor een aapje dat op beide voorspellers precies een 0 scoort. In onze steekproef hebben we geen aapjes die nul scoren op beide variabelen dus de vraag is wel of deze waarde reëel of relevant is, maar misschien wel de lengte van een aapje als die net geboren is en nog geen bananen heeft gegeten. Beter meet je dat soort aapjes en dan weet je het gewoon, het blijft gevaarlijk om buiten de *range* van jou gemeten x-waarden een voorspelling te doen qua Y.
Het regressie-gewicht (of *slope*) toegekend aan leeftijd ($b_1 = 35.05$), vertelt ons hoeveel twee aapjes van elkaar *zouden* moeten verschillen qua lengte als ze één eenheid verschillen qua leeftijd. Deze interpretatie geldt alleen onder *constant houding* van het aantal bananen, de andere predictor. Anders gezegd, betekenen deze laatste twee zinnen, dat als je twee aapjes neemt die hetzelfde aantal bananen eten, maar een jaar verschillen qua leeftijd, zal het oudere aapje wel 35.05 cm langer zijn dan het jongere aapje.
Het regressie-gewicht toegekend aan de voorspeller 'banaan' ($b_2 = 3.72$) betekent dus dat twee aapjes die de zelfde leeftijd hebben, maar wel 1 banaan verschillen, ongeveer 3.72 cm zullen verschillen qua lengte. Waar bij het aapje dat één banaan meer eet, natuurlijk langer zal zijn vanwege de positieve waarde van b_2.

Stap 2 Voorspelde waarden uitrekenen.

Dus nu zien we eindelijk wat de gewogen som van X_1 en X_2 is. Nog wat moeilijker gezegd, $\hat{Y}_i$ is te schrijven als (of is gelijk aan) de 'lineaire combinatie' van X_1 en X_2. Als je nu per aapje de scores op X_1 en X_2 invult, kun je uitrekenen wat zijn lengte *zou* moeten zijn volgens ons nieuwe model. Deze predicted values vind je in de kolom met $\hat{Y}_i$. Ik laat de berekening één keer zien aan de hand van aapje nummer negen.

$$\hat{Y}_9 = 78.44 + 35.05 \cdot X_{91} + 3.72 \cdot X_{92}$$

Aapje nummer 9 heeft op X_1 een waarde van 2.0 en op X_2 een 8, invullen geeft:

$$\hat{Y}_9 = 78.44 + 35.05 \cdot 2.0 + 3.72 \cdot 8 = 178.26$$

Hij zou dus 178.26 cm lang moeten zijn volgens ons model.

Stap 3 Kwadraten sommen uitreken.

Volgens het regressiemodel moet aapje nummer 9 dus 178.26 cm zijn. In werkelijkheid heeft hij een lengte van 180. Het model heeft dit aapje dus nog maar 1.74 cm onderschat. Of anders gezegd, dit aapje zit dus 1.74 cm boven zijn voorspelling en deze waarde is dus gelijk aan zijn error of residu:

$$e_i = Y_i - \hat{Y}_i$$
$$e_9 = Y_9 - \hat{Y}_9$$

$$e_9 = 180 - 178.26 = 1.74$$

De residuen vind je in de ener laatste kolom van tabel 6B en zouden natuurlijk netjes tot nul moeten optellen (als je heel precies zou zijn en niet te veel afrond tussendoor). In de laatste kolom staan de gekwadrateerde residuen, die je nodig hebt om SSE te berekenen (door de gekwadrateerde residuen op te tellen krijg je de *sum of squares due to error*).

Om SSM (*sum of squares due to model*) te berekenen moet je eerst weer weten hoeveel ons model (M_2) de voorspelling heeft veranderd ten opzichte van het nul-model, voor aapje nummer 9:

$$\hat{Y}_i - \bar{Y}$$

De systematische verschuiving in voorspelling, invullen voor aapje negen:

$$\hat{Y}_9 - \bar{Y}$$

$$178.26 - 150 = 28.26$$

Als deze waarde kwadrateert krijg de waarde 798.63 (in de kolom waar je onderaan ook SSM vindt) en de optelling (sommatie) van deze waarden leidt tot SSM.
Ook hier *moet* weer gelden dat de totale variatie in lengte scores (SST) op te delen is in een gedeelte verklaard en een gedeelte onverklaard:

$$SST = SSM + SSE$$

in ons bijzondere geval:

$$SST = 3000 = 2734.40 + 265.60$$

Stap 4 Bereken de proportie verklaarde variantie, de VAF.

We gaan over tot de berekening van '*Multiple* R^2'. We noemen deze R^2 'Multiple', omdat we dus met *meerdere* predictoren te maken hebben. We kijken nu naar een multiple samenhang of correlatie (Multiple R) omdat we kijken naar het (gezamenlijke) effect van enerzijds x_1 en x_2 op anderzijds y. Maar voor de VAF (*variance accounted for*) geldt nog steeds de zelfde formule:

$$R^2 = \frac{SSM}{SST}$$

Voor ons model is dat dus:

$$R^2_{y \cdot x_1 x_2} = \frac{SSM}{SST}$$

Met het subscript voor R^2 geef ik dus aan dat we *y* proberen te voorspellen op basis van x_1 en x_2.

$$R^2_{y \cdot x_1 x_2} = \frac{2734.40}{3000} = .91$$

We kunnen nu zeggen dat ongeveer 91 procent van de totale variatie in lengte-scores, verklaard kan worden door de (gezamenlijke) variatie in de predictoren 'leeftijd' en 'banaan'. Slechts 9 procent blijft onverklaard.

Stap 5 Het checken van aannames voor het uitvoeren en interpreren van je regressie-analyse.

Voor een multipele regressie-analyse zijn er natuurlijk weer een tal van aannames die je officieel eerst dient te checken, net zoals bij enkelvoudige regressie. Zelfs nog *meer* dan bij enkelvoudige regressie, zoals de *afwezigheid* van een *te* hoge correlatie tussen de predictoren. Mocht dit het geval zijn dan spreekt men van 'multicollineariteit'. Maar voor nu, laten we die assumpties even naast ons liggen en slaan we deze stap even over, dus voorlopig *no worries*.

Stap 6 Toetsing van het *gehele* model.

Bij een model met meerdere predictoren toets je altijd het model eerst als *geheel*. Hierbij is de vraag eigenlijk of er *iets* aan de hand is (H_1) of helemaal *niets* (H_0). De nul-hypothese stelt dat van de variatie in Y niets verklaard kan worden door de combinatie van x_1 en x_2, ook wel een horizontaal regressie-vlak (voor elke waarden van de predictoren is het gemiddelde op *y* gelijk). In dit geval is de proportie verklaarde variantie dus 0:

$$H_0 : \rho^2_{y \cdot x_1 x_2} = 0$$ natuurlijk gaan de uitspraken over de populatie.

$$H_1 : \rho^2_{y \cdot x_1 x_2} > 0$$

De alternatieve hypothese ontkent natuurlijk de H_0 en zegt dat de proportie verklaarde variantie groter is dan nul (kleiner dan nul kan niet). Om dit stelsel van hypothese te toetsen berekenen we de bijbehorende toets-statistiek (en *p*-waarde), in dit geval een *F*-waarde. In plaats van een *t*-test doen we nu een *F*-test. Om de *F*-waarde uit te rekenen kun je dat op de volgende manier doen (er zijn heel veel manieren of formules om een *F*-waarde te bereken, maar ik geef je er dus slechts één).

$$F = \frac{R^2}{1 - R^2} \cdot \frac{df_{error}}{df_{model}}$$

Met $df_{model} = p$, waarbij '*p*' voor het *aantal* predictoren staat, dus 2 en $df_{error} = n - p - 1$ dus 9 – 2 – 1 = 6 en aangezien we R^2 al gevonden hadden, kunnen we de boel invullen (ik gebruik nu de onafgeronde waarde voor R^2:

$$F = \frac{.9115}{1 - .9115} \cdot \frac{6}{2} = 30.90$$

We kunnen nu de bijbehorende p-waarde (overschrijdingskans of significantie) opzoeken in de *F*-tabel. Om dit te doen moet je weer rekening houden met de *degrees of freedom* (vrijheidsgraden). Maar nu hebben we te maken met een *koppel* van vrijheidsgraden, namelijk voor het model én voor de error. Respectievelijk dus 2 en 6, we gaan de de p-waarde opzoeken die hoort bij de F-verdeling voor 2 en 6 vrijheidsgraden, 2 voor het model, en 6 voor de error, ook wel $F(2;6) = 30.90$. Normaal rekent het statistiek-programma dat je gebruikt, de p-waarde uit.

Kijk in de *F*-tabel en zoek eerst het aantal vrijheidsgraden op voor de 'teller' (de *numerator*), je gebruikt hier altijd het aantal vrijheidsgraden van het model voor (2 bij ons). Zoek vervolgens naar het aantal vrijheidsgraden op voor de 'noemer' (de *denominator*), hier gebruik je dus altijd de vrijheidsgraden van de error voor (dus 6 bij ons). Vergelijk dan onze F-waarde (30.90) met de gegeven *F*-waarden uit de tabel. De *F*-tabel geeft als hoogste waarde 27.00 met een bijbehorende *p*-waarde (staart oppervlakte) van .001. Onze *F*-waarde is extremer en heeft dus een kleiner staartje dan .001 en we kunnen dus concluderen dat:

$$F(2;6) = 30.90,\ p < .001$$

Bij toetsing van een model mag je *nooit* de p-waarde verdubbelen (onze alternatieve hypothese was immers éénzijdig, een VAF kan alleen maar groter zijn dan nul, niet kleiner). En omdat onze p-waarde veel kleiner is dan $\alpha = .05$, moeten we de H_0 verwerpen en kunnen we concluderen dat R^2 in onze steekproef, significant afweek van nul, en dat in de populatie dat dus ook wel het geval zal zijn (ρ^2). Jippie!

Stap 7 Significantie per predictor, toetsing van de *slopes*.

Nu we weten dat er echt *iets* aan de hand is (een gedeelte van de variatie in *y* wordt verklaard), willen ook weten *hoe* dat komt. Ligt dit aan de variabele 'leeftijd' of aan de twee predictor 'banaan' of aan een combinatie van de twee? Om dit te beoordelen kijken we naar de significanties van de twee regressie-gewichten - of mischien iets meer van deze tijd - naar de betrouwbaarheidsintervallen. Ik kan dus eigenlijk gewoon refereren naar het vorige hoofdstuk, want dat heb ik daar al uitgelegd. Maar ik loop er even snel doorheen.

Om een uitspraak te doen over de significatie per predictor (of hun slope dus significant afwijkt van nul en generaliseerbaar is naar de populatie *slopes*), bereken je eerst de bijbehorende toets-statistiek. We toetsen de *slopes* nog steeds a.d.h.v. de t-verdeling en je gebruikt het aantal vrijheidsgraden van de error $df_{error} = n - p - 1$. In het algemeen, voor de j-de predictor bereken je de t-waarde als volgt:

$$t_{df=n-p-1} = \frac{b_j}{SE_{b_j}}$$ voor de variabel leeftijd wordt dat:

$$t(6) = \frac{b_1}{SE_{b_1}} = 35.046 / 7.724 = 6.12$$

en bijbehorende tweezijdige p-waarde opzoeken leidt tot:

$$t(6) = 6.12,\ p < .001$$

Dus aannemen dat leeftijd wel degelijk een positief effect heeft op lengte. Voor de variabele 'banaan' volgen we dezelfde procedure (ik gebruik weer preciezere getalletjes):

$$t(6) = \frac{b_2}{SE_{b_2}} = 3.716 / 1.352 = 2.75$$

Voor de predictor 'banaan' kunnen we dus ook concluderen dat het effect op lengte significant is ($t(6) = 2.75,\ .02 < p < .04$) en we kunnen dus concluderen dat 'meer bananen' ook echt met een hogere lengte kan worden geassocieerd.

Rapportage (bijvoorbeeld):

Om te onderzoeken of de lengte van een aapje voorspeld of verklaard kan worden op basis van de leeftijd en het aantal bananen dat een aapje per dag eet is een multipele regressie-analyse uitgevoerd. De twee predictoren ('leeftijd' en 'banaan') tezamen bleken 91 procent van de totale variatie in lengte scores te kunnen verklaren. Deze VAF ,was significant (R^2 = .91, $F(2;6)$ = 30.90, p < .001), waardoor we kunnen aannemen dat model generaliseerbaar is naar de populatie aapjes. Verder hadden beide predictoren een positief effect op de lengte. De slope voor leeftijd week significant af van 0 (b_1 = 35.05, $t(6)$ = 6.12, p < .001). Op basis van deze uitkomst kunnen we concluderen dat aapjes die ouder zijn, ook daadwerkelijk met een langere lengte geassocieerd kunnen worden. Ook het effect van de variabele 'banaan' was significant (b_2 = 3.72, $t(6)$ = 2.75, .02 < p < .04) Al met al is het dus handig om beide variabelen te gebruiken wanneer men lengte-scores wil voorspellen vanwege de significante multiple samenhang.

Officieel zijn er nog veel meer 'ditjes en datjes' bij het doen van een multipele regressie-analyse, maar gezien dit een introductie-cursus is, laten we het hierbij. De kleinste aap wenst je een heel fijn Inzicht!