Syllabus Statistiek voor Biologen

Hans Maassen

27 januari 2012

Inhoudsopgave

Hoofdstuk 1

Inleiding

De biologie is de studie van de levende wereld. Dit is een breed onderwerp, dat zich uistrekt van moleculen via cellen en organismen tot hele ecosystemen. In al deze gebieden probeert de bioloog steeds de volgende vragen te beantwoorden:

- **Wat** zijn de verschijnselen? Wat doet zich voor?
- **Hoe** steekt het in elkaar? Wat veroorzaakt wat? Kunnen we de werkelijkheid modelleren?

Dit zijn de algemene vragen van elke natuurwetenschap. Maar meer nog dan in de natuurkunde en de scheikunde komt men in de biologie steeds weer hetzelfde soort probleem tegen, dat alle waarneming bemoeilijkt, en alle uitspraken onzeker en voorlopig maakt. Of het nu gaat om het metabolisme van een cel, of om de vraag hoe een ecosysteem zijn grote aantal soorten kan ondersteunen: steeds weer stuit de bioloog op: **variabiliteit**. Deze neemt allerlei vormen aan:

1.1 Meetonnauwkeurigheid.

De eerste bron van variabiliteit kom je in alle natuurwetenschappen tegen, en wordt feitelijk door de onderzoeker zelf geïntroduceerd: de meetfout. Ook wanneer je herhaaldelijk precies dezelfde grootheid meet, dan nog zul je steeds een iets ander resultaat verkrijgen. Neem bijvoorbeeld de meting van de concentratie van een chemische stof in een oplossing, Om die uit tevoeren moet je 10 ml van de oplossing uit een Erlenmeyer pipetteren, en daar een gegeven hoeveelheid van een reagens aan toevoegen. Met een spectrofotometer bepaal je dan de kleur van het eindproduct, en daarmee de concentratie van de chemische stof in het oorspronkelijke vat.
Bij elk van deze stappen introduceer je een kleine onnauwkeurigheid, de feitelijke waarde zit net een beetje onder of boven de bedoelde waarde. Sommige van die foutjes zullen elkaar versterken, sommige heffen elkaar gedeeltelijk op. Soms bevat

de pipet een beetje meer dan 10 ml, soms een beetje minder. Soms is de hoeveelheid reagens wat groter, soms wat kleiner. Misschien laat je de kleur de ene keer wat meer tijd om zich te ontwikkelen, de andere keer wat minder. De instelling van spectrofotometer kan een beetje variëren. Soms heb je geluk, en heffen de fouten elkaar precies op. Maar daar heb je niets aan, omdat je dat niet weet.

1.2 Variabiliteit in fenotypen en genotypen

Ook als we met oneindige nauwkeurigheid zouden meten zouden we nog te maken krijgen met de variatie die er in natuur bestaat. Bij gelijk fenotype kan door verschillende inwerking van de omgeving grote variatie ontstaan in kleur en vorm van bijvoorbeeld de bladeren van een plant. Anderzijds verschillen ook de fenotypen binnen een populatie vaak aanzienlijk. Deze verschillen kunnen verkleind of vergroot worden door de inwerking van de omgeving. Dus wanneer je geïnteresseerd bent in "de" bladlengte van planten op een bepaalde plek, zul je daar een mouw aan moeten passen.

1.3 Variatie in ruimte en tijd

Ook op terreinen die betrekkelijk homogeen zijn, zoals een bos of een heideveld, zie je toch overal variatie: hier staan berken in een kluitje bij elkaar, daar zie je een bosje beuken, sommige in kleine groepjes, sommige geïsoleerd. Op sommige plaatsen groeien paddestoelen. De zaden, sporen en larven van de verschillende organismen zijn soms passief verspreid door de wind, soms actief verplaatst door vogels en bijen. Soms viel de appel niet ver van de boom. Bij het bepalen van de dichtheid van een soort in een gebied moet hiermee rekening worden gehouden.

1.4 Waarom statistiek voor beschrijving?

Voor het zuiver beschrijvende deel van het vak is al statistiek nodig. Omdat we altijd maar een deel van de werkelijkheid waarnemen, door steekproeven of andere toevallige keuzen, terwijl we in het geheel geïnteresseerd zijn, zijn we gedwongen te gokken. Kansrekening helpt daarbij.

1.5 Het aantonen van causaal verband

Door waarneming alleen kan geen causaal verband worden aangetoond. Het blijft altijd mogelijk dat twee factoren, die steeds blijken samen te gaan, toch geen invloed op elkaar uitoefenen. Dit is mogelijk doordat zij bijvoorbeeld door een derde worden aangestuurd. Een ander instrument van de wetenschap is wèl toe in staat oorzakelijk

verband althans zeer aannemelijk te meken: het experiment. Als wij willen weten of er een invloed bestaat van A op B, kunnen wij A zelf in de hand nemen. En als dan blijkt dat B al onze grillen volgt, mogen we aannemen dat A invloed heeft op B. Ook hierbij is kansrekening nodig, om te kwantificeren hoe aannemelijk die conclusie nou precies is.

Hoofdstuk 2

Theorievorming en toetsing

Als voorbeeld kijken we naar het verspreidingspatroon van de pissebed (*Sphaeroma rugicauda*, ook wel keldermot of ringkreeft). Dit bekende insect wordt bijvoorbeeld aangetroffen in bladafval, onder stenen, en onder de schors van dode bomen. Wat opvalt is dat hun verdeling over de ruimte heel ongelijk is. Zo zullen sommige stukken boomschors bijna geen pissebedden herbergen, terwijl andere krioelen van de beestjes. Hoe kunnen we dit verklaren?

2.1 Vraagstelling

Eén mogelijkheid is, dat er niets te verklaren valt! Misschien is het puur toeval dat er op de ene plek wat meer beestjes voorkomen dan op de andere. Om deze mogelijkheid uit te sluiten zou eigenlijk al een statistische test nodig zijn. Maar daar weten we nu nog te weinig van; laten we daarom voor het moment aannemen dat de ongelijkheid in de verdeling veel te groot is om zich per toeval voor te doen.

Wat is er zoal aan verklaringen mogelijk?

Wel, misschien zijn de pissebedden ooit begonnen met een mooi egale verdeling, en zijn ze toen op sommige plaatsen, door ongunstige omstandigheden, uitgestorven en op andere plaatsen juist vermeerderd. Of misschien verzamelen ze zich op plaatsen die bijzonder gunstig voor hen zijn, bijvoorbeeld door aanwezig voedsel, vochtigheid, of lichtval. Ook is het mogelijk dat de aanwezigheid van pissebedden soortgenoten aantrekt, zodat zij uiteindelijk steeds in kluitjes voorkomen, op overigens vrij willekeurige plekken.

Stel dat we, na wikken en wegen, tot de hypothese komen dat de vochtigheidsgraad van een plek de bepalende factor is. Dit noemen we: onze **onderzoekshypothese**, en we willen nagaan of deze juist is. Dat wil zeggen: we willen weten of de vochtigheid werkelijk de **oorzaak** is van opeenhopingen van pissbedden: of deze beestjes inderdaad de voorkeur geven aan vochtige omgevingen boven droge plekken.

2.2 Experiment

Zoals in het vorige hoofdstuk betoogd, is voor het aantonen van zo'n oorzakelijk verband een experiment nodig. Een goed idee lijkt de proefopstelling in Fig. 2.1. Een pissebed wordt in een gang geplaatst die zich uitsplitst in twee zijgangen. We leggen in de ene gang een in vocht gedrenkt lapje stof, in de andere een droog lapje. Verder construeren we de gang zó dat de pissebed niet kan omkeren, en, tenzij hij helemaal stil gaat zitten, gedwongen is een keuze te maken tussen beide zijgangen. We herhalen de proef met verschillende beestjes, en noteren welke kant ze op gaan.

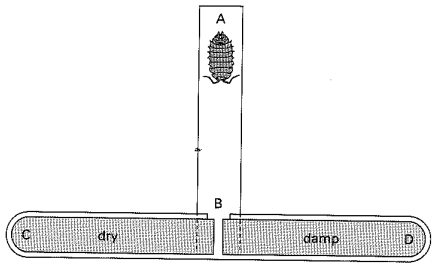


Figure 2.1 A simple choice chamber.

Stel nu dat blijkt dat van de twintig beestjes die wij in onze opstelling hebben losgelaten, er 14 de rechter (vochtige) gang in slaan, 5 de linkergang, terwijl er 1 eigenlijk geen keuze maakt: hij blijft voor of vlakbij de splitsing stilzitten, en we blazen het experiment af. Wat zij we door deze uitslag wijzer geworden?

Op het eerste gezicht suggereert dit resultaat een bevestiging van ons vermoeden: de pissebedden zijn in ruime meerderheid de vochtige gang ingeslagen.
Maar laten we niet te hard van stapel lopen, en de mogelijkheden eens op een rijtje zetten.

2.2.1 Mogelijkheden

We kunnen de volgende mogelijkheden onderscheiden:

(a) Pissebedden zoeken vochtige plekken;

(b) Pissebedden hebben geen voorkeur voor vocht, maar er is een ander verschil tussen de zijgangen waardoor zij liever de rechtergang kiezen;

(c) Pissebedden gaan liever rechtsaf dan linksaf;

(d) Het verschil komt door onze behandeling. Wij pakken de beestjes bijvoorbeeld steeds op dezelfde manier uit hun verblijf op, en we beschadigen steeds aan dezelfde kant een pootje;

(e) De beestjes hebben helemaal geen voorkeur, maar wij hebben ze wat langer de kans gegeven als ze rechtsaf leken te slaan;

(f) De pissebedden hebben geen voorkeur, maar toevallig sloegen er meer rechtsaf dan linksaf.

Mogelijkheid (f) is de meest serieuze concurrent van onze onderzoekshypothese, mogelijkheid (a). We komen daar uitgebreid op terug. Maar laten we eerst afrekenen met de mogelijkheden (b), (c), (d) en (e). Deze kunnen we nagenoeg uitschakelen door een zorgvuldige experimentele opzet.

2.2.2 Opzet van de proef

Mogelijkheid (b) kan worden bestreden door ervoor te zorgen dat de vochtigheid echt het enige onderscheid is tussen de beide gangen: we moeten de opstelling mooi horizontaal plaatsen, met gelijke verlichting van alle kanten, afgeschermd van het aardmagnetisch veld, van lawaaibronnen, etcetera.
Hier duikt een probleem op. Dit kan een dure onderneming worden! Maar een belangrijker obstakel is wel het volgende: hoeveel kosten of moeite we er ook in steken, hoe kunnen we ooit weten dat we *alle* factoren die de voorkeur van de pissebedden kunnen beïnvloeden, hebben uitgeschakeld? Hoe kunnen we weten wat belangrijk is in het leven van een pissebed?

De volgende truc, die in experimenten vaak wordt toegepast, lost dit dilemma op, en helpt direct ook mogelijkheden (c) en (d) de wereld uit: *randomiseren*. We laten telkens opnieuw door het lot bepalen, welke zijgang we bevochtigen, en welke we droog laten. Als de beestjes dan tòch nog in meerderheid de vochtige kant kiezen, althans meer dan redelijkerwijs aan het toeval kan worden toegeschreven, dan zijn de andere verschillen tussen de gangen ((b), (c) en (d)) uitgeschakeld.

Tenslotte is er mogelijkheid (e): ons vooroordeel tijdens de proef. We moeten natuurlijk eerlijk en onbevooroordeeld te werk gaan. Het beste zou het zijn als de experimentator helemaal niet zou *weten* welke kant in het huidige experiment de vochtkant is. Maar als dat te lastig is, laten we dan in elk geval zorgen voor een precies protocol: hoe lang moet het beestje stil zitten, voordat we het proefje opgeven? Hoe ver moet het een gang in wandelen voordat we besluiten dat het die gang *gekozen* heeft?
Deze zaken moeten *van tevoren* netjes in een protocol worden vastgelegd.

2.3 Statistische analyse

Tenslotte komen we op ons hoofdonderwerp: de statistiek. We hebben 20 proefjes gedaan, netjes gerandomiseerd en volgens protocol. En inderdaad: in 14 gevallen

koos de pissebed de vochtige kant, in 5 gevallen de droge kant, en in 1 geval vond geen keuze plaats.

Hebben we nu iets aangetoond (mogelijkheid (a)), of kan een kritische tegenstander dit resultaat afdoen als toeval (mogelijkheid (f))?

2.3.1 De significantie-drempel

Opnieuw staan we voor een fundamenteel dilemma: Als je per sé wilt, kun je namelijk *alles* afdoen als toeval! In principe is er een positieve kans dat de pissebedden zonder aarzeling allemaal rechtsaf marcheren, ookal zien ze geen enkel verschil tussen de gangen. Die kans is misschien wel één op miljoen, maar onze kritische tegenstander zal kunnen opperen dat dit toch niet uit te sluiten valt. Als hij keihard bewijs blijft eisen, komen we niet verder.

Nee, we zullen met onze kritische tegenstander een akkoordje moeten sluiten: hoe klein moet de kans worden die *datgene wat we zien gebeuren*, zou maken, *als er geen enkele voorkeur zou zijn*, opdat wij van de tegenstander gelijk krijgen? Hoe gek moet het worden voordat onze tegenstander zich gewonnen geeft?

Na enig onderhandelen komen de tegenstander en wij tot een vergelijk: een kans van α vindt hij klein genoeg. Heel gebruikelijk is: $\alpha = 5\%$.

Dit is natuurlijk een practische overweging. Als de tegenstander bijvoorbeeld een kans van 1 op miljoen blijft eisen, krijgen we wel heel betrouwbare experimenten, maar dan moeten we maanden, zo niet jaren proefjes doen. Omgekeerd, als de tegenstander erg gauw tevreden is, dan zijn de proefjes gauw klaar, maar het resultaat is niet te vertrouwen.

Deze zogenaamde *significantiedrempel* α varieert van het ene wetenschapsgebied tot het andere, en hangt ook af van het belang van het onderzoek en de beschikbare middelen. Zelf vind ik $\alpha = 5\%$ aan de hoge kant: je accepteert dan als het ware dat één op de twintig wetenschappelijke resultaten niet deugt. Nogal veel. Toch zullen we deze waarde van α meestal hanteren, omdat dit in de biologie praktijk is.

2.3.2 De nulhypothese

Om het pleit te beslechten moeten we nu dus een kans gaan uitrekenen. Namelijk: de kans dat de gedane waarneming *of een nog gekkere* zich zou voordoen als er geen enkele verschil in voorkeur tussen vochtige an droge gangen zou bestaan.

Deze kans noemen we: de p-waarde. Hij wordt vaak aangduid met de letter P (van "*probability*").

De voorwaarde waaronder we de kans uitrekenen, namelijk dat het verschil in voorkeur tussen vochtig en droog nul is, noemen we: **de nulhypothese**. Deze moet aan de volgende twee eisen voldoen:

1. De nulhypothese staat voor de *afwezigheid* van een effect of afwijking, die we eigenlijk juist graag willen *aantonen*.

2. Onder de nulhypothese zijn we in staat, P uit te rekenen.

De onderzoekshypothese (in ons geval: pissebedden houden van vocht) heet in dit verband **de alternatieve hypothese**. De kans P zullen we met α gaan vergelijken. Valt hij kleiner uit dan α, dan wordt de nulhypothese onaannemelijk, en de kritische tegenstander geeft zich gewonnen. We *verwerpen* de nulhypothese en nemen de alternatieve hypothese aan.

Valt hij daarentegen groter uit dan α, dan ... weten we nog niks! In ons experiment hebben de pissebedden niet genoeg voorkeur laten zien om onze strenge kriticus te overtuigen Maar het kan natuurlijk best zo zijn, dat ze er een (misschien lichte) voorkeur voor vochtige plekken op na houden. Nader onderzoek is geboden. Maar hoe dan ook: we kunnen niet concluderen dat de pissebedden *geen* voorkeur hebben: de nulhypothese wordt **niet bewezen**!

2.3.3 De kansberekening

Tijdens de berekening nemen we de nulhypothese aan: pissbedden trekken zich niets aan van vochtigheid of droogte. Dat wil zeggen: ze gaan net zo gemakkelijk de droge gang in als de vochtige gang.

Wat is de kans dat er niettemin van de 19 pissebedden die een keuze hebben gemaakt, er 14, of zelfs nog meer, de vochtige gang in lopen?
Om dit uit te rekenen, kijken we naar *alle* mogelijke uitslagen van het experiment met de 19 pissebedden. Deze uitslagen kunnen we weergeven met rijtjes die er zó uitzien:

$$VVDVDDDVDVVDVDVVDDV$$

Het zijn rijtjes van 19 letters, bestaande uit alleen de letters V (van 'Vochtig') en D (van 'Droog'). Hoeveel van die rijtjes zijn er? Wel, voor rijtjes van twee letters zijn er 4 mogelijkheden:

$$VV, VD, DV \text{ en } DD\ .$$

Rijtjes van drie letters zijn er tweemaal zoveel:

$$VVV, VVD, VDV, VDD, DVV, DVD, DDV \text{ en } DDD\ .$$

(Bij elk van de rijtjes van twee letters kun je er twee van drie letters maken door er een V of een D achter te zetten.) Enzovoort: er zijn $2^4 = 16$ rijtjes van vier letters, $2^5 = 32$ rijtjes van vijf letters, en uiteindelijk $2^{19} = 524288$ rijtjes van 19 letters, ongeveer een half miljoen.
Deze zijn onder de nulhypothese allemaal even waarschijnlijk. De kans op één ervan is dus steeds 1 op 524288. In het bijzonder is de kans dat de pissebedden allemaal de vochtige gang ingaan, en *geen enkele* de droge gang, dat is de kans op het rijtje $VVVVVVVVVVVVVVVVVVV$, gelijk aan 1/524288. (Laten we deze kans π_0 noemen.)

Goed, dus als de pissebedden alle negentien de vochtige gang in waren gegaan, was onze proef zeker geslaagd geweest, want $\pi_0 = 1/524288 \approx 0{,}000001907\ldots$ ligt ruimschoots onder onze overeengekomen drempelwaarde $\alpha = 0{,}05$. Maar helaas is dit niet het geval: het zijn er maar 14 geweest. We moeten dus nog de kansen bijtellen op 18, 17, 16, 15 en 14 inslagen in de vochtige gang, dat wil zeggen: de kans op een rijtje met 18, 17, 16, 15 of 14 letters V en respectievelijk 1, 2, 3, 4 of 5 letters D.
Hoeveel rijtjes zijn er met één D erin? Welnu, deze kan op 19 verschillende plaatsen staan, dus dit geeft 19 mogelijkheden:

$$DVVVVVVVVVVVVVVVVV,\ VDVVVVVVVVVVVVVVVVV,$$
$$\ldots\ldots\ldots,$$
$$VVVVVVVVVVVVVVVVVDV,\ VVVVVVVVVVVVVVVVVVD\ .$$

De kans op 1 D-afslag is daarom

$$\pi_1 = \frac{19}{524288} \approx 0{,}000036239\ .$$

Laten we nu de rijtjes tellen met 17 keer V en 2 keer D. Met andere woorden: op hoeveel manieren kunnen we twee D's kwijt op 19 posities? Stel je voor: je hebt twee letters D in de hand, (we noemen ze D_1 en D_2), en je wilt deze op twee van de negentien posities zetten. Voor de letter D_1 heb je 19 mogelijkheden. Bij elke daarvan heb je voor de letter D_2 er nog 18 over. Voor het plaatsen van het tweetal heb je dus $19 \times 18 = 342$ mogelijkheden. De rest van de posities vul je aan met V's. Heb je nu correct het aantal rijtjes met 2 D's en 17 V's geteld? Nee! Elke mogelijkheid is dubbel geteld: één keer met D_1 voorop, en één keer met D_2 voorop. Het correcte aantal rijtjes is daarom

$$\frac{342}{2} = 171\ .$$

We noemen dit: $\binom{19}{2}$, oftewel "negentien boven twee". Hieronder verstaan we het aantal manieren om uit een verzameling van 19 elementen er 2 te kiezen. Ga bij jezelf na dat dit precies hetzelfde is als wat we hierboven hebben geteld.
De kans op precies 2 D-afslagen is dus, nog steeds onder de nulhypothese:

$$\pi_2 = \frac{\binom{19}{2}}{2^{19}} = \frac{171}{524288} \approx 0{,}0003262\ .$$

Nu berekenen we de kans π_3 op precies 3 D-afslagen: Hiervoor moeten we $\binom{19}{3}$ bepalen, het aantal manieren om 3 van de 19 posities aan te wijzen. We gaan op dezelfde manier te werk: we nemen drie D's in de hand, zeg D_1, D_2 en D_3. Er zijn nu $19 \times 18 \times 17$ mogelijke manieren om deze drie letters op drie verschillende posities neer te zetten, en de rest aan te vullen met V's. We krijgen dan rijtjes van het type

$$VVVD_3VVVD_1VVVVVVD_2VVVV\ .$$

Maar staan deze rijtjes voor allemaal verschillende proefuitslagen? Nee, natuurlijk! De nummertjes die onder de D-en hangen, en die aangeven wanneer wij die letter daar hebben neergezet, zijn helemaal niet relevant! Dus hoe vaak is iedere mogelijkheid nu geteld? Dat hangt ervan af, op hoeveel manieren we de cijfertjes 1, 2 en 3 op een rijtje kunnen zetten. Dit zijn er zes:

$$123,\ 132,\ 213,\ 312,\ 231,\ 321\ .$$

(Ik kan de 1 op drie plaatsen neerzetten, bij elk van deze mogelijkheden kan de 2 op twee plaatsen staan, en de 3 voegt zich naar de rest: het aantal manieren is dus $3 \times 2 \times 1 = 6$.)
Door dit aantal moeten we delen. Er komt:

$$\binom{19}{3} = \frac{19 \times 18 \times 17}{3 \times 2 \times 1} = 969.$$

En dus:

$$\pi_3 = \frac{\binom{19}{3}}{2^{19}} = \frac{969}{524288} \approx 0{,}001848\ .$$

Op dezelfde manier berekenen we de kansen π_4 en π_5 op 4 respectievelijk 5 D-afslagen:

$$\begin{aligned}
\pi_4 &= \frac{\binom{19}{4}}{2^{19}} = \frac{19 \times 18 \times 17 \times 16}{4 \times 3 \times 2 \times 1} \cdot \frac{1}{524288} = \frac{3876}{524288} \approx 0{,}007393\ . \\
\pi_5 &= \frac{\binom{19}{5}}{2^{19}} = \frac{19 \times 18 \times 17 \times 16 \times 15}{5 \times 4 \times 3 \times 2 \times 1} \cdot \frac{1}{524288} = \frac{11628}{524288} \approx 0{,}022179\ .
\end{aligned}$$

De kans op hoogstens 5 D-afslagen (de p-waarde) is dus onder de nulhypothese:

$$\begin{aligned}
P &= \pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 = \frac{\binom{19}{0} + \binom{19}{1} + \binom{19}{2} + \binom{19}{3} + \binom{19}{4} + \binom{19}{5}}{2^{19}} \\
&= \frac{1 + 19 + 171 + 969 + 3876 + 11628}{524288} = \frac{16664}{524288} \approx 0{,}03178\ .
\end{aligned}$$

2.3.4 Toetsing

Deze kans ligt nog net beneden onze overeengekomen 5%-grens. Dat wil zeggen: onder de nulhypothese heeft het resultaat van de proef (namelijk minstens 14 keuzes voor de vochtige gang) een voldoende lage kans om de nulhypothese te verwerpen. Conclusie: pissebedden hebben een voorkeur voor de vochtige gang.
Er is een oorzakelijk verband tussen de vochtigheid van een plek en het aldaar vóórkomen van pissebedden.

Hiermee is in principe de kous af. We maken nog een slotopmerking over **eenzijdig** respectievelijk **tweezijdig** toetsen. Van het begin af aan zijn we ervan uitgegaan

dat pissebedden, als ze al een voorkeur hebben, er een hebben voor de *vochtige* gang. We zouden zeer verbaasd staan als ze in het experiment plotseling in meerderheid de *droge* gang zouden kiezen. Zo'n uitslag zou dus een negatief resultaat zijn, koren op de molen van onze kritische opponent.
Maar heel vaak ligt de situatie anders, en hebben we vooraf geen idee welke kant de voorkeur opgaat die we aan het meten zijn. De **alternatieve hypothese** is dan niet: 'pissebedden hebben een voorkeur voor vochtig', maar: 'pissebedden hebben een voorkeur'. Een grote meerderheid voor de droge gang zou dan dus net zo goed een reden zijn om de nulhypothese te verwerpen ten gunste van de alternatieve hypothese. Dit heet: *tweezijdig toetsen.*
Maarr. . . . . : dan zou de eis voor verwerping luiden: $P \leq \alpha$, waarbij

$$\begin{aligned} P &= 2^{-19}\left(\binom{19}{0}+\binom{19}{1}+\binom{19}{2}+\binom{19}{3}+\binom{19}{4}+\binom{19}{5}\right) \\ &\quad +2^{-19}\left(\binom{19}{14}+\binom{19}{15}+\binom{19}{16}+\binom{19}{17}+\binom{19}{18}+\binom{19}{19}\right) \\ &\approx 2 \times 0{,}03178 = 0{,}06356\ , \end{aligned}$$

en dan zou de nulhypothese *niet* worden verworpen bij $\alpha = 0{,}05$!
We zien dat tweezijdig toetsen hogere significantie-eisen stelt dan éézijdig toetsen. In de praktijk is tweezijdige toetsing aan te bevelen, tenzij je echt heel goede biologische redenen hebt om aan te nemen dat het effect maar één kant uit kan gaan.

2.4 De binomiale verdeling

We laten nu het voorbeeld van de pissebedden achter ons. Uit het voorbeeld trekken we de wiskundige conclusie dat, als alle rijtjes van, zeg n nullen en enen, evenveel kans hebben, de kans π_k op k enen en $n-k$ nullen gelijk is aan:

$$\pi_k = 2^{-k}\binom{n}{k}\ .$$

We noemen dit de *kansverdeling* van het aantal enen. Voor het geval $n = 19$ vind je deze verdeling in Tabel 2.1.

k	$\binom{19}{k}$	$\pi_k = 2^{-19}\binom{19}{k}$
0	1	0,0000019
1	19	0,0000362
2	171	0,0003262
3	969	0,0018482
4	3876	0,0073929
5	11628	0,0221787
6	27132	0,0517502
7	50388	0,0961075
8	75582	0,1441612
9	92378	0,1761971
10	92378	0,1761971
11	75582	0,1441612
12	50388	0,0961075
13	27132	0,0517502
14	11628	0,0221787
15	3876	0,0073929
16	969	0,0018482
17	171	0,0003262
18	19	0,0000362
19	1	0,0000019

Tabel 2.1: Binomiale verdeling met $n = 19$ en $p = \frac{1}{2}$

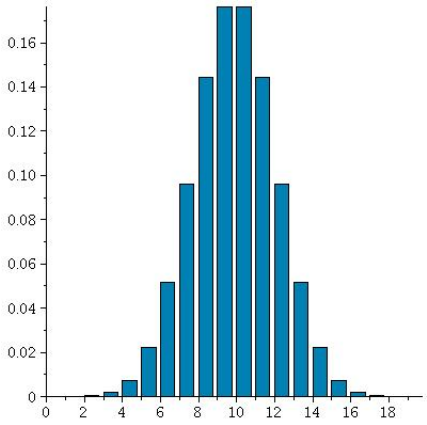
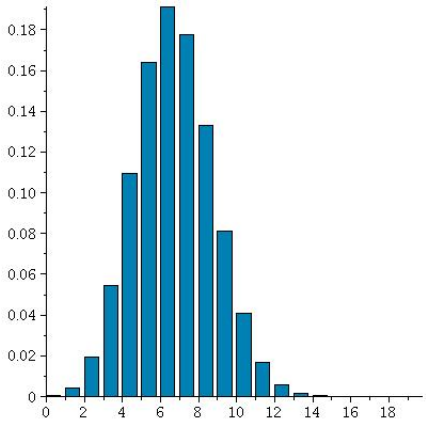


Fig. 2.1: Binomiale verdeling met $n = 19$ en $p = \frac{1}{2}$



Binomiale verdeling met $n = 19$ en $p = \frac{1}{3}$

2.4.1 Ongelijke kansen

Bij het berekenen van deze kansen zijn we er steeds van uitgegaan dat de kansen op 'succes' of 'mislukking' (in ons geval een afslag naar een vochtige of juist naar een droge gang) gelijk zijn, dat wil zeggen: allebei gelijk aan $\frac{1}{2}$. Wat als de kans op succes gelijk is aan $\frac{1}{3}$? Of aan $\frac{1}{4}$? Of meer algemeen: wat wordt de kansverdeling $\pi_0, \pi_1, \ldots, \pi_n$ als de 'succeskans' p is, en de kans op een mislukking $q := 1 - p$?

Welnu, de kans π_0 op 0 successen is dan q^n. De kans π_1 op 1 succes is gelijk aan het aantal mogelijke rijtjes bestaande uit één 1 en $n-1$ 0-en, vermenigvuldigd met de kans op één dergelijk rijtje. Het aantal rijtjes is n, en de kans op zo'n rijtje is pq^{n-1}. Kortom:

$$\pi_1 = npq^{n-1} .$$

En zo gaan we door: de kans op k enen en $n-k$ nullen is

$$\pi_k = \binom{n}{k} p^k q^{n-k} . \tag{2.1}$$

De kansen $\pi_0, \pi_1, \ldots, \pi_n$ vertellen je bij elk mogelijk aantal enen wat de kans daarop is. We noemen zo'n rijtje kansen dan ook de *kansverdeling* van het aantal enen. Zie Figuur 2.1. Meer hierover in §5.2.

2.5 Toetsing met de binomiale verdeling

We geven nu nog een 'praktijkprobleem' waar toetsing aan te pas komt, maar nu met een binomiale verdeling die $p \neq \frac{1}{2}$ heeft.

Zieke bomen. In een groot bos is een boomziekte uitgebroken. De boswachter beweert dat een derde deel van de bomen is aangetast. Wij zijn daar echter niet zo zeker van, en willen de bewering van de boswachter *toetsen*.
Probleem hierbij is, dat het heel moeilijk te zien is of een boom aan de ziekte lijdt: daarvoor moet een stuk uit de bast voor analyse meegenomen worden naar het laboratorium. We gaan daarom niet alle zieke bomen in het bos tellen, maar nemen een steekproef van derdig bomen: uit elk daarvan snijden we een stukje bast weg, en analyseren het. De uitslag: k van deze dertig bomen blijken ziek. (Hierbij is bijvoorbeeld $k = 16$, $k = 12$, of $k = 4$.) Voor welke waarden van k stellen wij de boswachter in het ongelijk? En wanneer geven wij hem 'the benefit of the doubt'? (We hanteren een significantieniveau van 5%.)
Dit probleem lijkt nogal op dat van de pissebedden. Er is een *nulhypothese* die we kunnen gebruiken om kansen te berekenen op steekproefuitslagen: dit is de bewering van de boswachter.
Als we ervan uit gaan dat hij gelijk heeft, dan wordt de kans op k zieke bomen in de steekproef gegeven door (2.1) met $n = 30$ en $p = \frac{1}{3}$. Eerst berekenen we a en b

zó dat nog nèt geldt:

$$\sum_{k=0}^{a} \pi_k \leq 2{,}5\% \quad \text{en} \quad \sum_{k=b}^{n} \pi_k \leq 2{,}5\% \; . \tag{2.2}$$

Vervolgens zullen we de hypothese dat $p = \frac{1}{3}$ verwerpen als $k \leq a$ of $k \geq b$.

2.6 Berekening met de grafische rekenmachine

Het met de hand uitrekenen van a en b is lastig. We besluiten gebruik te maken van de grafische rekenmachine, bijvoorbeeld de `TI-84 plus` van Texas Instruments. (Als je een andere GR hebt, gaat de berekening op ongeveer dezelfde manier. Kijk eventueel in je gebruiksaanwijzing wat de overeenkomstige toetsen zijn.)
De `TI-84 plus` kan op twee manieren de binomiale verdeling bepalen: hij kent de commando's `binompdf` en `binomcdf`.

2.6.1 `binompdf`

Deze afkorting staat voor *binomial probability density function*, hetgeen betekent: *binomiale kansdichtheidsfunctie*. Het commando berekent de kans π_k uit formule (2.1), dat is de kans op k successen bij n pogingen, steeds met slaagkans p. Dit doe je zó:
Druk op `2nd DISTR`. Op het scherm verschijnt een rijtje kansverdelingen. De binomiale dichtheidsfunctie staat onder nummer `A`. We krijgen hem door te drukken op `ALPHA A`. Inderdaad verschijnt nu op het scherm: `binompdf(`.
We voeren onze argumenten in: eerst de waarde van n, dan die van p, en tenslotte een waarde voor k, zeg $k = 10$. We sluiten af met een haakje. Er komt te staan:

```
binompdf(30,.3333333333,10)
```

Druk nu op `ENTER`, en de uitkomst verschijnt:

```
.5847596
```

Een behoorlijk grote kans! Alle andere π_k zijn natuurlijk kleiner, want de som is 1. Dit is wel te begrijpen: 10 is immers $30 \times \frac{1}{3}$, het te verwachten aantal zieke bomen in een steekproef van 30, als de boswachter gelijk heeft.

2.6.2 `binomcdf`

Dit betekent: *binomial cumulative distribution function*, dus *binomiale cumulatieve verdelingsfunctie*. Dat cumulatieve is precies wat wij willen! In formule (2.2) staat immers de som van alle binomiale kansen van $\mathbb{P}[K = 0]$ tot en met $\mathbb{P}[K = a]$. (Onder $\mathbb{P}[\cdots]$ verstaan wij: 'de kans dat $\cdots$'.) De binomiale kansen bereken je als volgt:
Druk op `2nd DISTR ALPHA B`. Ditmaal verschijnt op het scherm: `binomcdf(`.
Nu voeren we in: eerst n, dan p, dan onze gok voor a; zeg $a = 5$.

Er komt:

```
.0354541718
```

Dit is nog een beetje te veel, we willen immers dat de kans kleiner wordt dan 0,025. We proberen daarom ook $a = 4$. Resultaat voor $\mathbb{P}[K \leq 4]$:

```
.0122297238
```

Nu is de kans inderdaad kleiner dan 2,5%, dus onze waarde voor a wordt $a = 4$. Op dezelfde manier willen we b bepalen zó dat de kans dat $K \geq b$ kleiner is dan 2,5%, maar de kans dat $K \geq b - 1$ groter dan 2,5%.

Een complicatie hierbij is dat onze GR alleen cumulatieve kansen *van onder af* wil geven. We kunnen dit omzeilen door gebruik te maken van het feit dat de kans dat $K \geq b$ gelijk is aan $1 -$ (de kans dat $K < b$), en dit weer aan $1 -$ (de kans dat $K \leq b-1$). We moeten dus het punt opzoeken waar de cumulatieve kans de waarde 0,975 passeert. Enig zoeken levert:

$$\mathbb{P}[K \leq 15] = 0{,}9812\ldots \qquad \text{en} \qquad \mathbb{P}[K \leq 14] = 0{,}9565\ldots$$

Dat houdt in dat de kans dat $K \geq 16$ nog juist onder 0,975 ligt. Anders gezegd:

$$\mathbb{P}[K \geq 15] > 0{,}025 > \mathbb{P}[K \geq 16]\ .$$

Conclusie: De schatting van de boswachter wordt *verworpen* als we in onze steekproef 16 of meer zieke bomen treffen, of als dit er juist maar 4 zijn of minder. Kortom, we verwerpen als

$$|k - 10| \geq 6\ .$$

Hoofdstuk 3

Beschrijvende Statistiek

De proeven die in de biologie worden gedaan zijn natuurlijk niet allemaal van het simpele type als hierboven bij de pissebedden en de zieke bomen: een rij uitslagen met steeds maar twee mogelijkheden.

3.1 Typen waarnemingen:

Laten we eens inventariseren wat voor soort waarnemingen (data) er zoal mogelijk zijn.

1. nominaal: alleen een stel klassen wordt onderscheiden (bijv. vochtig/droog, maar ook: kleur, soort, ras, geslacht);

2. ordinaal: klassen met een rangorde (bijv. geluidssterkte, sterkte van aardbevingen volgens Richter, ontwikkelingsstadium van een plant);

3. intervalschaal: meting met een vaste eenheid, maar zonder dat er een zinvol nulpunt aan te geven is (bijv. tijd in jaartallen, temperatuur in graden Celsius of Fahrenheit);

4. ratioschaal: er is een vaste eenheid en een duidelijk nulpunt (temperatuur in graden Kelvin, lengte, gewicht, concentratie).

Op een ratioschaal zijn bijvoorbeeld de begrippen 'verhouding' en 'logaritme' zinvol. Deze begrippen hebben echter geen betekenis op een intervalschaal. Wel kun je daar nog van het begrip "gemiddelde" spreken. Op een ordinale schaal is ook dit weggevallen, maar daar heb je nog wel het begrip "mediaan". Dit begrip heeft tenslotte ook geen zin meer bij nominale variabelen.

Al deze soorten data (behalve natuurlijk de nominale) vallen weer uiteen in *discrete* en *continue* data.

- Voorbeelden van discrete data: legselgrootte, aantal bloemen per plant, tentamencijfer;
- Continue data: lengte, gewicht, overlevingspercentage, temperatuur.

3.2 Het weergeven van de gegevens

De eerste stap in de statistiek is het ordenen van data tot een overzichtelijk geheel. Bij de beschrijving van continue data staan we voor de keuze, met welke nauwkeurigheid we de gegevens presenteren. De volgende vuistregel biedt daarbij enig houvast.

De afrondingsregel van Ehrenberg. Bepaal vanaf welke decimale positie de gegevens "variabel" worden, d.w.z. vanaf welke positie het gehele bereik van 0 tot 9 in de data vertegenwoordigd is. Rond de getallen af totdat je *twee* variabele posities overhoudt.

Voorbeeldje: je hebt de vijf metingen

$$181.633,\ 182.796,\ 189.132,\ 189.239,\ 191.151\ .$$

Geef hiervan bij presentaties alleen de eerste vier cijfers weer:

$$181.6,\ 182.8,\ 189.1,\ 189.2,\ 191.2$$

De eerste twee posities zijn niet "variabel", de derde wel.

Voorbeeld: vliegende herten. Tabel 3.1 hieronder is erg onoverzichtelijk: de data zijn niet gerangschikt en de vele decimalen ontnemen het zicht op het geheel. Het gaat om lichaamslengte (in cm) van 100 vliegende herten. In Tabel 3.2 staan

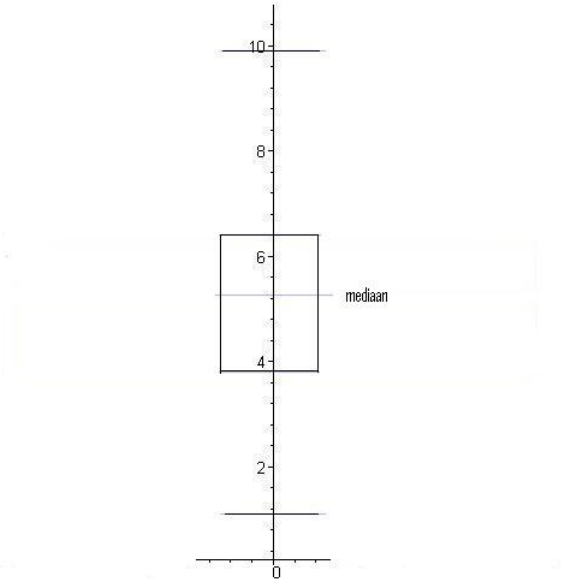
4.816	3.832	4.792	1.842	6.441	7.410	1.900	2.683	8.382	5.833
6.324	5.602	5.478	7.071	4.605	3.348	4.134	5.826	5.772	4.561
2.738	1.353	2.884	4.730	4.357	8.624	7.907	3.720	1.069	2.530
5.815	4.197	5.974	5.041	6.591	2.655	6.551	5.276	7.094	5.524
2.584	3.851	6.386	5.776	5.408	6.933	1.711	9.851	6.332	6.156
6.418	5.208	7.211	7.175	8.633	7.751	6.103	6.882	3.992	6.358
3.324	9.199	4.066	1.959	7.717	2.047	2.966	3.843	4.087	5.300
8.303	6.259	6.845	2.511	3.958	4.157	3.728	5.290	7.182	3.115
3.987	3.486	4.372	4.547	3.399	7.779	7.013	4.738	4.335	5.035
5.499	5.837	7.245	6.355	5.668	3.026	4.611	8.935	3.837	1.423

Tabel 3.1: Lengtes van vliegende herten

1.1	2.6	3.4	4.0	4.6	5.3	5.8	6.3	6.9	7.7
1.4	2.7	3.5	4.1	4.6	5.3	5.8	6.3	6.9	7.8
1.4	2.7	3.7	4.1	4.6	5.3	5.8	6.4	7.0	7.8
1.7	2.7	3.7	4.1	4.7	5.4	5.8	6.4	7.1	7.9
1.8	2.9	3.8	4.2	4.7	5.5	5.8	6.4	7.1	8.3
1.9	3.0	3.8	4.2	4.8	5.5	6.0	6.4	7.2	8.4
2.0	3.0	3.8	4.3	4.8	5.5	6.1	6.4	7.2	8.6
2.0	3.1	3.9	4.4	5.0	5.6	6.2	6.6	7.2	8.6
2.5	3.3	4.0	4.4	5.0	5.7	6.2	6.6	7.2	8.9
2.5	3.3	4.0	4.5	5.2	5.8	6.3	6.8	7.4	9.9

Tabel 3.2: Lengtes van vliegende herten, afgerond en geordend

dezelfde gegevens opnieuw, maar nu geordend, en afgerond volgens de regel van Ehrenberg: Dat is al een hele verbetering. Maar vervolgens kunnen deze data grafisch worden gepresenteerd als een staafdiagram (histogram, zie hieronder) of een box-whiskerplot. Om tot deze weergave te komen gaan we uit van de tweede tabel. We zien dat de lengtes variëren van 1.1 tot 9.9. Het verschil 9.9-1.1=8.8 heet het *bereik* (de "range") van de gegevens. De *mediaan* is de (middelste) waarde, zó gekozen dat de helft van de data links ervan, en de helft rechts ervan ligt. Het *eerste kwartiel* is een waarde zó, dat een kwart van de waarnemingen links ligt, en de rest rechts; het *derde kwartiel* heeft juist een kwart links liggen, en de rest rechts. Het interval tussen het eerste en het derde kwartiel heet het *interkwartielbereik*.



Mediaan: 5.25, interkwartielbereik: $[3.8, 6.4]$

Frequentietabel. We maken allereerst een *indeling van de gegevens in klassen.* Het aantal klassen nemen we ongeveer gelijk aan $\sqrt{n}$ als n het aantal gegevens is.

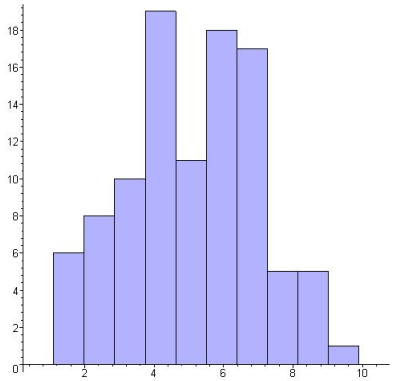


Fig. 3.2: Histogram lengtes vliegende herten

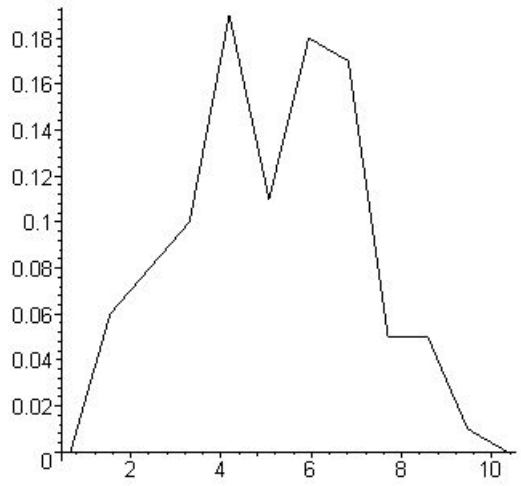
Hier hebben we 100 gegevens, dus hanteren we 10 klassen. De breedte van de klassen nemen we gelijk, dus de range gedeeld door het aantal klassen, i.e. 0.88. De eerste klasse is dus het interval $[1.1, 1.98]$, de tweede $[1.98, 2.86]$ enz.

Vervolgens maken we een frequentietabel, waarin we aangeven hoeveel van de gegevens tot ieder van de klassen behoren:

klasse	1	2	3	4	5	6	7	8	9	10
frequentie	6	8	10	19	11	18	17	5	5	1

Histogram. Deze frequentietabel kunnen we vervolgens grafisch weergeven in een staafdiagram of histogram. Het eenvoudigste histogram wordt geconstrueerd vanuit de frequentietabel met staven op het betreffende interval en hoogte de bij de betreffende klasse behorende frequentie. De totale oppervlakte van het histogram is dan het aantal gegevens maal de klassenbreedte. Door de hoogte te delen door deze oppervlakte krijgen we het *relatieve frequentie histogram* met totale oppervlakte 1.

De relatieve frequentieveelhoek Deze ontstaat als volgt. Plot punten (x_i, y_i) voor elke klasse, waar x_i het midden van de klasse is en y_i de relatieve frequentie van de klasse. Verbind vervolgens deze punten met lijnstukjes.



Relatieve frequentieveelhoek voor lengtes vliegende herten

We maken de veelhoek af door frequentie nul aan te nemen in de "intervallen" net vóór het eerste interval en net na het laatste. De totale oppervlakte onder de relatieve frequentieveelhoek is 1, want deze oppervlakte is dezelfde als die onder het relatieve frequentie histogram.

Hier volgt de definitie van enkele begrippen en hun waarde voor de gegevens van vliegende herten:

Steekproefgemiddelde: neem de som van alle waarden en deel door het aantal waarden. (Zie paragraaf 3.4, formule **??**.)

Steekproef-standaarddeviatie: (Zie paragraaf 3.4, formule 3.8.)

Mediaan: Bij een oneven aantal gegevens $x_1 \leq x_2 \leq \cdots \leq x_{2m-1}$: neem de middelste waarde x_m. Bij een even aantal gegevens $x_1 \leq x_2 \leq \cdots \leq x_{2m}$: neem $\frac{1}{2}(x_m + x_{m+1})$. In ons voorbeeld zijn er 100 gegevens; $x_{50} = 5.2$ en $x_{51} = 5.3$. De mediaan 5.25 ligt hier precies midden in.

Interkwartielbereik: het interval met als grenzen de 25 percentielwaarde (waar 25% van de waarden onder liggen) en de 75 percentielwaarde (waar 75% van alle meetwaarden onder liggen). In ons voorbeeld: $[3.8, 6.4]$.

Box-Whisker plot: De grenzen van de box corresponderen met het interkwartielbereik. De uitsteeksels hebben een lengte van maximaal 1.5 maal het interkwartielbereik, maar steken niet buiten het bereik van de data uit. De waarden die nog buiten deze uitsteeksels vallen heten *uitbijters*. De data van vliegende herten hebben geen uitbijters, want er zijn geen waarden groter dan $6.4 + 1.5 \cdot (6.4 - 3.8) = 10.3$ of kleiner dan $3.8 - 1.5 \cdot (6.4 - 3.8) = -0.1$

3.3 Continue verdelingen

De vliegende herten uit het voorbeeld zijn gevangen in een groot bos. Ze kunnen daarom worden beschouwd als een *steekproef* uit de *populatie* bestaande uit alle vliegende herten in dat bos.

Deze populatie is heel groot, en ons onbekend. We beschrijven haar daarom met een geïdealiseerd *wiskundig model*: een continue verdeling. Als denkbeeldige populatie nemen we een deel van de reële getallen, en we stellen ons voor dat de kans op een trekking wordt gegeven door een *kansdichtheidsfunctie* $f(x)$. Dit is een functie die overal positief of nul is, en waarvan de totale integraal gelijk is aan 1. De kans op een trekking in het interval $[a, b]$ wordt gegeven door

$$\int_a^b f(x)\,dx\ .$$

Populatie-gemiddelde en -standaarddeviatie. Het populatie-gemiddelde is gedefinieerd als:

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x)\,dx\ ; \tag{3.1}$$

De populatie-variantie is:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx\ . \tag{3.2}$$

En de populatie-standaarddeviatie:

$$\sigma = \sqrt{\sigma^2}\ . \tag{3.3}$$

3.4 Steekproeven uit continue verdelingen

Uit onze populatie trekken we een steekproef van n individuen en meetwaarden $x_1, x_2, \ldots, x_n$. Om een idee te krijgen van de populatie-grootheden μ en σ berekenen we de volgende *steekproef-grootheden* of *statistieken*:
Het steekproef-gemiddelde

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k\ ; \tag{3.4}$$

de steekproef-variantie

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n} (x_k - \overline{x})^2\ , \tag{3.5}$$

en de steekproef-standaarddeviatie

$$s = \sqrt{s^2}\ . \tag{3.6}$$

3.4.1 Notatie en lettergebruik

1. De variantie is een *kwadratische grootheid*, in tegenstelling tot de standaarddeviatie.

2. Populatieparameters worden aangegeven met Griekse letters, steekproefvariabelen ('statistieken') met Romeinse letters.

3. Vóórdat we de steekproef nemen, is de het steekproefgemiddelde natuurlijk nog onbekend. We geven zo'n *onbekende door het toeval bepaalde variabele* of *stochastische variabele* of kortweg *stochast* aan met een Romeinse hoofdletters. Zo is bijvoorbeeld het steekproefgemiddelde $\overline{x}$ meestal verschillend van μ. Maar *gemiddeld over alle steekproeven* is $\overline{X}$ wèl gelijk aan μ. Zie hierover Hoofdstuk 4.

4. Merk op dat we bij het berekenen van de steekproef-variantie hebben gedeeld door $n-1$, niet door n. De reden is dat, gemiddeld over *alle mogelijke* steekproeven, de som $\sum_{i=1}^{n}(X_i - \overline{X})^2$ niet $n\sigma^2$ oplevert, maar $(n-1)\sigma^2$! Er gaat één vrijheidsgraad verloren doordat het gemiddelde $\overline{X}$ niet constant is, maar met de steekproef meebeweegt. Door te delen door $n-1$ brengen we de verwachting van de stochast S^2 precies op σ^2. Zie ook hierover Hoofdstuk 4

3.4.2 Gemiddelde en spreiding met de GR

Om een rij getallen in de GR (specifiek: de TI-84 Plus) in te voeren, typ je eerst een {, dan de getallen, gescheiden door komma's, en tenslotte } `STO> 2nd LIST`. Kies nu een naam voor de rij uit, je hebt de keuze tussen `L1`, `L2`, ... `L7`. Stel, je kiest: `L1`. Vanaf dat moment is je getallenrij bereikbaar onder de naam L_1, die je met de toetsen `2nd L1` kunt oproepen.
Het gemiddelde van de rij vind je met `2nd LIST` $\leftarrow$ `3`. Er verschijnt nu:

```
mean(     ,
```

waarna je afsluit met `2nd L1 ) ENTER`, en je uitkomst $\overline{x}$ wordt getoond. De standaarddeviatie s wordt verkregen door `2nd LIST` $\leftarrow$ `7` in plaats van `2nd LIST` $\leftarrow$ `3` in te geven, zodat je

```
StDev(     ,
```

te zien krijgt. Vul weer L_1 in, en s wordt getoond. Een uitkomst die getoond wordt, kun je bewaren met `STO> 2nd ALPHA A ENTER`. (Een andere letter is ook mogelijk.) Je kunt dan het resultaat terugroepen onder de letter A (dat wil zeggen, door in te typen: `ALPHA A ENTER`.

3.4.3 Toepassing op de vliegende herten

Als we deze bewerkingen toepassen op de tabel van de lengten van vliegende herten, Tabel 3.1, vinden we:

- gemiddelde
$$\overline{x} = \frac{1}{100} \cdot (4.816 + \ldots + 1.423) = 5.100 \tag{3.7}$$
- variantie
$$s^2 = \frac{1}{99} \cdot \left[(4.816 - 5.100)^2 + \ldots + (1.432 - 5.100)^2\right] = 3.690 \; ;$$
- standaarddeviatie
$$s = \sqrt{3.690} = 1.921 \; . \tag{3.8}$$

3.4.4 Schatting:

- μ: dit is de 'typische lengte' van een vliegend hert.
  $\overline{x}$: schatter van μ op basis van de steekproefgegevens.
- σ: de spreiding in de populatie rond het gemiddelde μ;
  s: schatter van σ op basis van steekproefgegevens.

Dit is het onderwerp van Hoofstuk 4.

3.4.5 Afrondingsregels voor afgeleide variabelen

- bij tussenberekeningen *alle* beschikbare cijfers gebruiken
- voor *presentatie* van het resultaat niet alle cijfers weergeven:
  - bij $n < 25$: cijfers volgens Ehrenberg
  - bij $25 \leq n < 100$: één cijfer meer dan Ehrenberg
  - bij $n \geq 100$: twee cijfers meer dan Ehrenberg

Hoofdstuk 4

Schattingen en de centrale limietstelling

Als we onze waarnemingen netjes hebben geordend, willen we er ook iets mee doen. We willen conclusies trekken over de wereld. In Hoofdstuk 2 hebben we een hypothese getoetst over pissebedden: zij bleken een voorkeur aan de dag te leggen voor vochtige plekken. In Hoofdstuk 3 hebben we van een steekproef van lengten van vliegende herten het gemiddelde en de standaarddeviatie bepaald. Wat kunnen we daaruit concluderen?

4.1 Schattingen

Welnu, de steekproef is getrokken uit een *populatie*: zeg de verzameling van *alle* vliegende herten uit een bepaalde streek. Eigenlijk zijn het gemiddelde $\overline{x}$ uit (3.7) en de standaarddeviatie s uit (3.8) helemaal niet zo interessant. Veel meer van belang is de gemiddelde lengte van een vliegend hert in de natuur, en de spreiding die daarin zit, dus het *populatiegemiddelde* μ uit (3.1) en de populatie-standaarddeviatie σ uit (3.2). Deze willen we uit onze steekproefgrootheden $\overline{x}$ en s kunnen *schatten*.

Een ander voorbeeld is dat van de zieke bomen in het bos uit § 2.5. Dit keer is er echter geen boswachter met een uitgesproken opninie, maar willen we gewoon weten welk percentage p van de bomen besmet is: we willen p *schatten* uit de gegevens k en n.

4.1.1 Het schatten van een kans

In een bos staan Z zieke bomen en G gezonde bomen. Het totale aantal bomen is dus $B = Z + G$. We willen weten wat Z/B is, de fractie zieke bomen. Dit is hetzelfde als de kans op een zieke boom als ik een willekeurige boom uitkies:

$$p = \frac{Z}{B} \, .$$

We schatten deze fractie door een steekproef van, zeg, n bomen te nemen, en te tellen hoeveel zieke bome erbij zijn. Als dit aantal k is, dan *schatten* we onze fractie p met

$$\widehat{p} := \frac{k}{n} . \tag{4.1}$$

Het is duidelijk dat deze schatting niet precies goed hoeft te zijn, want we hebben (lang) niet alle bomen onderzocht. Hoe goed is onze schatting? Wat kunnen we zeggen over het verschil tussen $\widehat{p}$ en p?
Om iets zinnigs te kunnen zeggen, doen we *een gedachten-experiment*. We stellen: Wat zou er gebeuren als we niet **één steekproef van n bomen** zouden nemen, maar **alle mogelijke steekproeven van n bomen**?
We zouden dan heel veel schattingen van de kans p krijgen:

$$\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \ldots, \widehat{p}_N . \tag{4.2}$$

Als we **de variantie in al deze denkbeeldige schattingen** zouden weten, zouden we ook een idee hebben, hoe goed onze ene eigen schatting is. Deze is immers gebaseerd op één willekeurige keuze uit alle mogelijke steekproeven, en waarom zou die veel beter of slechter zijn dan de 'modale' schatting in de lijst (4.2)?

Nadeel is wel dat $\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \ldots, \widehat{p}_N$ een *ontzettend* lange lijst is! Er zijn namelijk $N = B^n$ steekproeven mogelijk. Maar gelukkig hoeven we niet de hele lijst in onze GR in te typen. Dat zou een levenswerk zijn. Met nadenken komen we er ook. Daar gaat-ie:
Het aantal mogelijke steekproeven met alleen gezonde bomen is G^n. Dus in de lijst (4.2) staan G^n nullen. Dat ruimt op. Met dezelfde methode als uit Hoofdstuk 2 vinden we het aantal keren dat de schatting $\widehat{p} = k/n$ in de lijst staat:

$$\binom{n}{k} Z^k G^{n-k} .$$

Hier rinkelt een bel. We zien dat de schattingen voldoen aan de binomiale verdeling! De kans op een steekproef die een schatting k/n oplevert is immers

$$\begin{aligned} \frac{1}{N} \quad &\times \quad \left(\text{aantal steekproeven met schatting } \frac{k}{n} \right) \\ &= \quad \frac{1}{B^n} \times \binom{n}{k} Z^k G^{n-k} = \binom{n}{k} p^k (1-p)^{n-k} . \end{aligned}$$

We moeten dus zien uit te rekenen, wat de spreiding is van de binomiale verdeling. Voor wie dit al eens gezien heeft: de variantie van de binomiale verdeling is

$$\sigma^2 = np(1-p) . \tag{4.3}$$

En omdat de variantie een kwadratische grootheid is, is de variantie in $\widehat{p} = k/n$ dus gelijk aan

$$\sigma^2_{\widehat{p}} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \ .$$

Dus

$$\sigma_{\widehat{p}} = \frac{c}{\sqrt{n}} \ . \tag{4.4}$$

De constante c is hier $\sqrt{p(1-p)}$, de standaarddeviatie van een $\{0,1\}$-waardige stochast. Maar afgezien daarvan is (4.4) een heel algemeen verschijnsel:

De $\sqrt{n}$-wet: Het verschil tussen de schatter en het echte gemiddelde is, in de zin van de wortel uit het gemiddelde kwadraat, omgekeerd evenredig met de wortel uit de steekproefgrootte.

4.1.2 Bewijs van de $\sqrt{n}$-wet

Laten we deze belangrijke wet bewijzen door de variantie echt uit te rekenen van de denkbeeldige rij data (4.2). Het bewijs is pittig, maar als je het begrijpt, werpt het wel enig licht op het succes van steekproeven. Als je het niet begrijpt, kun je altijd nog de $\sqrt{n}$-wet (4.4) op basis van (4.3) gewoon geloven.
Eerste definiëren we een functie op de bomen die kijkt of ze ziek zijn:

$$\chi(b) := \begin{cases} 1 & \text{als de boom } b \text{ ziek is} \\ 0 & \text{anders .} \end{cases}$$

De variantie σ^2_k van het aantal zieke bomen in de steekproef is

$$\begin{aligned}
\sigma^2_k &= \frac{1}{N} \sum_{\text{steekproeven}} (\text{aantal zieke bomen} - np)^2 \\
&= \frac{1}{N} \sum_{\text{steekproeven}} \left(\sum_{i=1}^{n} (\chi(b_i) - p) \right)^2 \\
&= \frac{1}{N} \sum_{\text{steekproeven}} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} (\chi(b_i) - p)(\chi(b_j) - p) \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \cdot \frac{1}{N} \sum_{\text{steekproeven}} (\chi(b_i) - p)(\chi(b_j) - p)
\end{aligned}$$

Als $i \neq j$, dan is hier de som over alle steekproeven 0, omdat $Z = pB$:

$$\begin{aligned}
\sum_{\text{steekproeven}} \big(\chi(b_i) - p\big)(\chi(b_j) - p) &= \sum_{\text{steekproeven}} \big(\chi(b_i)\chi(b_j) - p\chi(b_i) - p\chi(p_j) + p^2\big) \\
&= B^{n-2}Z^2 - 2pB^{n-1}Z + B^n p^2 = 0 \ .
\end{aligned}$$

Dus is

$$\begin{aligned}
\sigma_k^2 &= \sum_{i=1}^{n} \cdot \frac{1}{N} \sum_{\text{steekproeven}} \big(\chi(b_i) - p\big)^2 \\
&= \sum_{i=1}^{n} \cdot \frac{1}{N} \sum_{\text{steekproeven}} \big(\chi(b_i) - 2p\chi(b_i) + p^2\big) \\
&= \frac{n}{N}\big(B^{n-1}Z - 2pB^{n-1}Z + p^2\big) \\
&= n \cdot \frac{B^n}{N}(p - 2p^2 + p^2) \\
&= np(1-p) \; .
\end{aligned}$$

Daarmee hebben we (4.3) bewezen.

4.1.3 Het schatten van een gemiddelde

Als we een trekking $x_1, x_2, \ldots, x_n$ uit een populatie hebben gedaan, dan ligt het voor de hand als schatter van het populatiegemiddelde μ het steekproefgemiddelde $\overline{x}$ te nemen. Maar ook hier is het duidelijk dat μ best ook iets groter of kleiner kan zijn dan $\overline{x}$. Hoe betrouwbaar is onze schatting van μ?
Ditmaal zonder bewijs herhalen we de $\sqrt{n}$-wet:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \; . \tag{4.5}$$

Deze *fout in het steekproefgemiddelde* $\sigma_{\overline{X}}$ wordt ook wel de *standaardfout* genoemd. In het Engels: *standard error* of *s.e.*.
Merk op dat we hier ineens zijn overgegaan op de hoofdletternotatie $\overline{X}$. Dit doen we bij voorkeur als de steekproef nog niet is genoemen, en het steekproefgemiddelde dus nog onbekend of **stochastisch** is. Alleen dan kun je immers zinvol van een *spreiding* spreken.

4.2 De centrale limietstelling

Behalve op de $\sqrt{n}$-wet is de statistiek gebaseerd op nog een tweede wonder: Terwijl bij toenemende steekproefgrootte de kansverdeling van het gemiddelde $\overline{X}$ steeds smaller wordt, neemt zij bovendien steeds meer een vaste vorm aan, een vorm die altijd dezelfde is, en die niet afhangt van de vorm van de kansverdeling die in de populatie heerst: de klokvorm of **normale verdeling**. Dit wonder heet de *centrale limietstelling*.

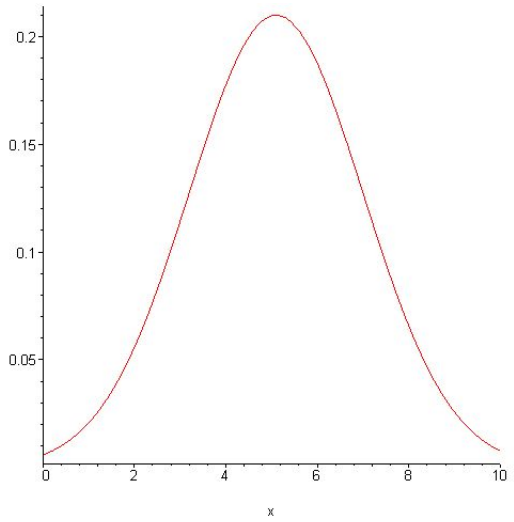
4.2.1 De normale verdeling

In paragraaf 3.3 hebben we gezien hoe een continue verdeling kan worden beschreven met een *kansdichtheidsfunctie*. Dit is een niet-negatieve functie op f op $\mathbb{R}$ waarvoor $\int_{-\infty}^{\infty} f(x)\ dx = 1$. Dit betekent dat de totale oppervlakte tussen de grafiek van f en de x-as oppervlakte 1 bedraagt. Bij zo'n kansdichtheidsfunctie kun je voor elk interval een kans definiëren: de kans dat een grootheid waarden tussen reële getallen a en b (met $a < b$) aanneemt is de integraal $\int_a^b f(x)\ dx$.

Voor een klein interval is die kans klein, en naarmate het interval groter wordt (naar beide kanten) nadert die kans meer tot 1. Een eenvoudig voorbeeld hiervan is de *uniforme dichtheidsfunctie* op een interval $[a, b]$: we nemen $f(x) = \frac{1}{b-a}$ als $a \leq x \leq b$ en $f(x) = 0$ als $x < a$ of $x > b$.

Het voorbeeld waar het ons nu om gaat is de *normale dichtheidsfunctie*

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$



normale verdeling met $\mu = 5.1$ en $\sigma = 1.9$

Een grootheid X die deze functie als kansdichtheid heeft, noemen we *normaal verdeeld*. Zo'n stochastische variabele heeft dan vanzelf gemiddelde μ en variantie σ^2. Een normale verdeling met $\mu = 0$ en $\sigma = 1$ noemen we *standaard-normaal*. Elke normale verdeling is te schalen naar een standaard-normale verdeling: als X normaal verdeeld is met gemiddelde μ en variantie σ, dan is de grootheid

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

standaard-normaal verdeeld. Er is dus voor de oppervlaktes onder de grafiek van *alle* normale verdelingsfuncties maar één tabel nodig. (Zie Tabel 4.5 aan het einde van dit hoofdstuk).

In de praktijk zijn veel grootheden bij benadering normaal verdeeld, zoals lichaamslengten, en steekproefgemiddelden. Dit laatste (en misschien ook wel het eerste) komt door de genoemde *centrale limietstelling*.

4.2.2 De centrale limietstelling in de praktijk

Het staafdiagram van de binomiale verdeling $\text{Bin}(n, p)$ gaat voor grote n steeds meer lijken op de grafiek van de normale verdeling.

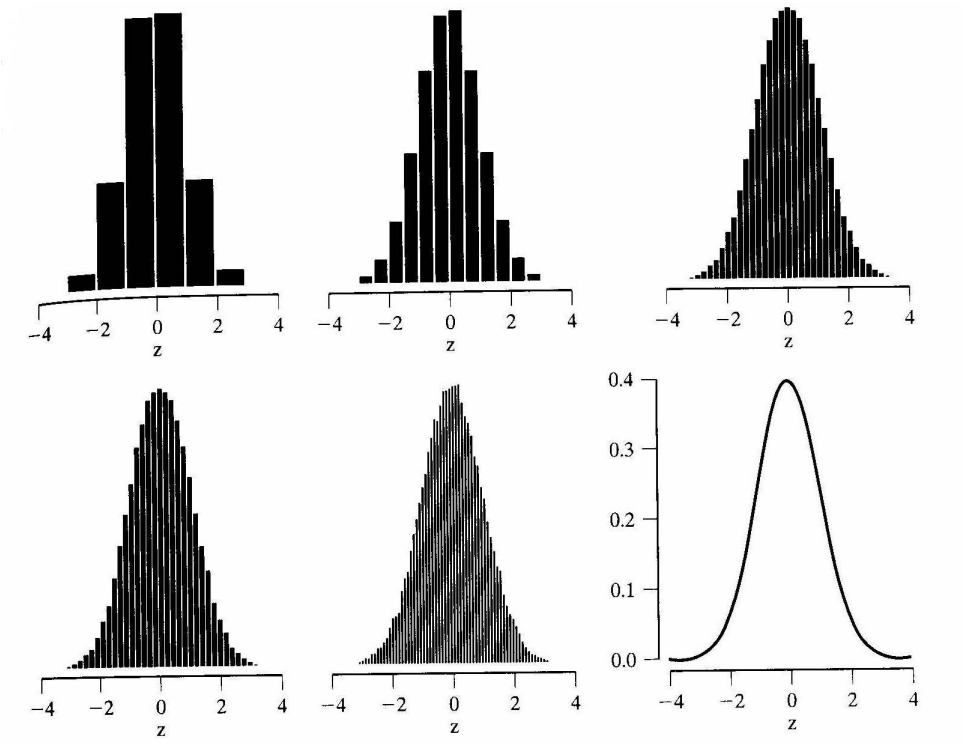


FIGURE 4.2: The progression of a histogram to a continuous distribution as the sample size increases.

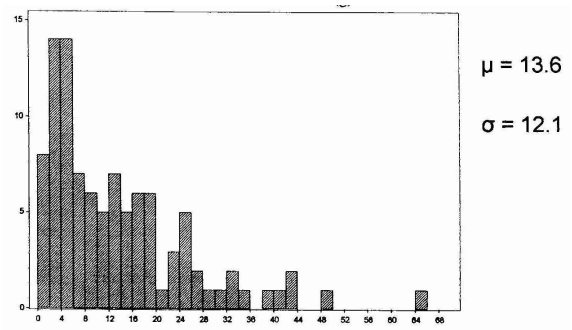
Maar er zijn nog veel meer omstandigheden waaronder de normale verdeling uit zichzelf de kop opsteekt:

Een computerexperiment We bekijken weer het voorbeeld uit § 3.2: het gewicht van 100 Vliegende Herten: Tabel 3.1. Maar nu doen we even net of dit de *gehele populatie* is! We hebben alle vliegende herten in het bos gevangen.
Dan zijn de populatieparameters:

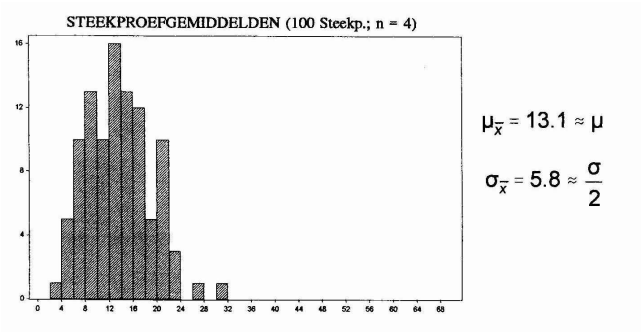
$$\mu = 13.58, \quad \sigma^2 = 146.66, \quad \text{dus} \quad \sigma = 12.11 \ .$$

Aan het histogram in paragraaf 3.2 zien we dat de verdeling sterk asymmetrisch is.

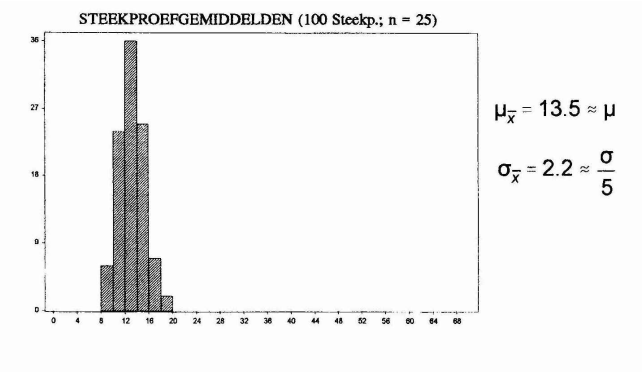
We nemen nu een aantal steekproeven van verschillende grootten uit deze populatie: 100 steekproeven met $n = 1$, 100 steekproeven met $n = 4$ en 100 steekproeven met $n = 25$.
Hier is de eerste uitslag. Deze ziet er ongeveer zo uit als de verdeling zelf (Fig. 3.2).



en hier is de uitslag van de tweede serie steekproeven:



en hier de derde:



Conclusies

- De steekproefgemiddelden $\overline{X}$ zijn bij benadering normaal verdeeld rond het populatiegemiddelde μ. (NB: dit is het geval *ondanks de asymmetrische populatieverdeling!*)

- De variantie van de verdeling van $\overline{X}$ is σ^2/n en dus evenredig met de variantie σ^2 van de onderliggende populatie en omgekeerd evenredig met de steekproefgrootte.

- De standaarddeviatie ($= \sqrt{\sigma^2/n}$) van deze verdeling is een maat voor hoever de steekproefgemiddelden van het populatiegemiddelde verwijderd zijn. Dit is de $\sqrt{n}$-wet.

Herinnering:

$$\text{Standard error} = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \, .$$

De uitslag van dit computer-experiment is een voorbeeld van de **centrale limietstelling**, die hier expliciet formuleren:

Stelling 1. (Centrale limietstelling voor bekende σ:) *Als we alle mogelijke steekproeven ter grootte n trekken uit een populatie met gemiddelde μ en variantie σ^2 dan zijn de steekproefgemiddelden $\overline{X}$, mits n groot genoeg is,*

- *bij benadering* normaal *verdeeld*

- *met gemiddelde $\mu_{\overline{X}} = \mu$ en variantie $\sigma^2_{\overline{X}} = \sigma^2/n$*

4.2.3 Hoe groot moet n zijn?

Wat betekent in deze stelling eigenlijk 'bij benadering normaal verdeeld'? Wel, dat mogen we zelf kiezen. Alleen, hoe strengere eisen we stellen, des te groter moet n zijn om eraan te voldoen. De volgende vuistregels geven aan hoe hiermee in de praktijk mee om te gaan.

1. Als de onderliggende populatie *normaal* verdeeld is, dan is $\overline{X}$ altijd — onafhankelijk van de steekproefgrootte — normaal verdeeld met gemiddelde μ en standaarddeviatie $\sigma/\sqrt{n}$.

2. Als de verdeling van de onderliggende populatie niet normaal is, maar wel *symmetrisch*, dan is $\overline{X}$ al bij relatief kleine steekproeven ($n \leq 10$) 'in goede benadering' normaal verdeeld met gemiddelde μ en standaarddeviatie $\sigma/\sqrt{n}$.

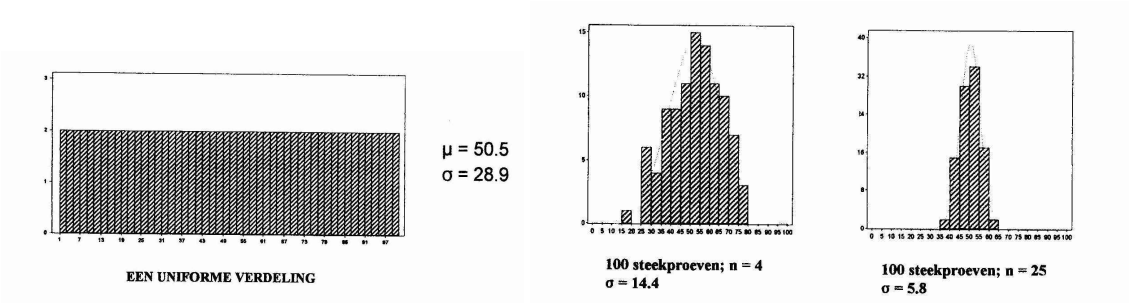
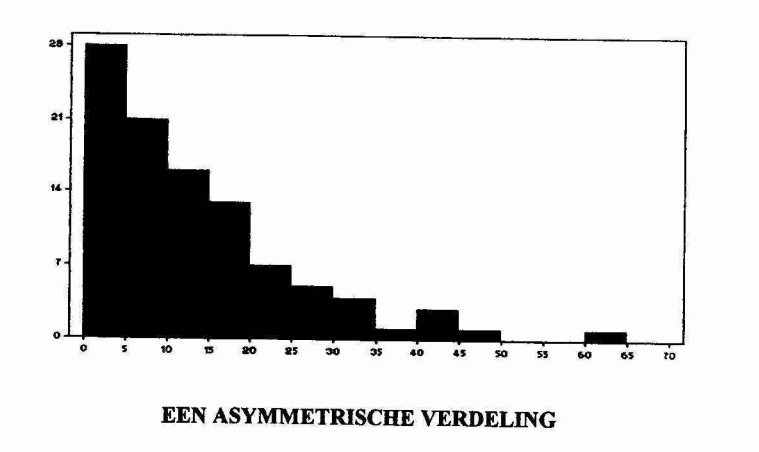
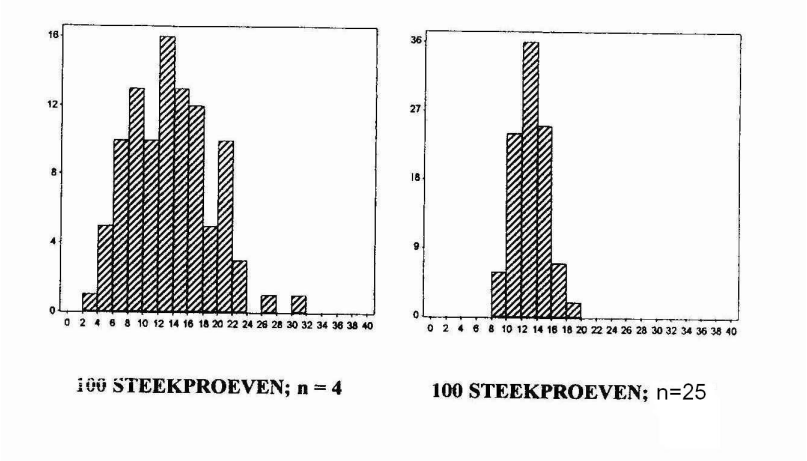


Fig. 4.2.3: Steekproeven uit de uniforme verdeling

3. Als de verdeling van de onderliggende populatie *sterk asymmetrisch* is, dan is $\overline{X}$ pas bij grotere steekproeven ($n \geq 30$, soms pas vanaf $n \geq 100$) bij benadering normaal verdeeld met gemiddelde μ en standaarddeviatie $\sigma/\sqrt{n}$.





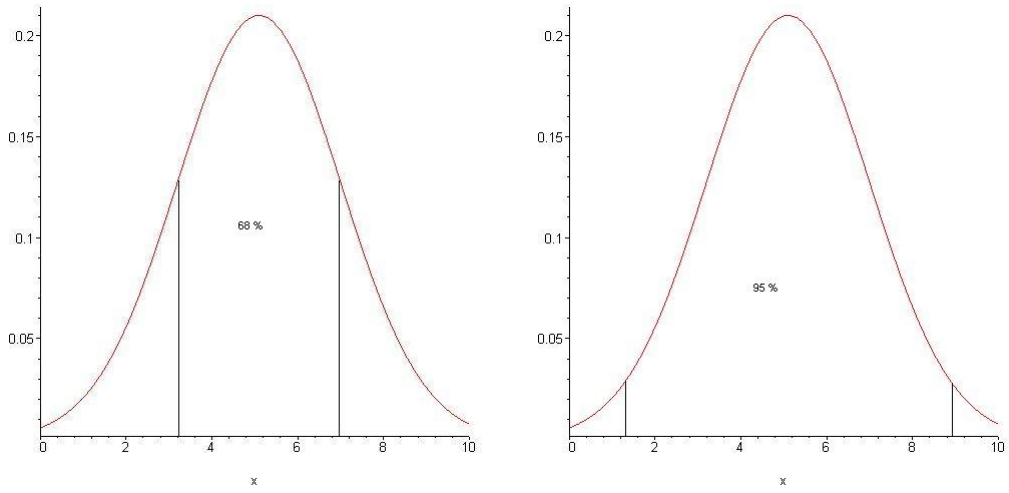


Fig. 4.3: 68% van de gegevens bij een normale verdeling liggen in het interval $[\mu - \sigma, \mu + \sigma]$ en 95% in het interval $[\mu - 2\sigma, \mu + 2\sigma]$

4.3 Theoretische betrouwbaarheidsintervallen

Stel eens even dat je van de populatie **wel** de variantie σ^2 kent, maar **niet** het gemiddelde μ. (Hoe zou dit bijvoorbeeld kunnen? Zie paragraaf 4.4 Met behulp van de centrale limietstelling kunnen we kwantificeren hoe groot de verwachte afwijking is tussen de schatter $\overline{X}$ en de te schatten waarde μ:

$$\overline{X} \text{ is normaal verdeeld rond } \mu \text{ met standaarddeviatie } \sigma/\sqrt{n}$$

Dus: $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is standaard normaal verdeeld

$$\mathbb{P}\left[-1.96 < Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right] = 0.95$$

$$\mathbb{P}\left[-1.96\sigma/\sqrt{n} < \overline{X} - \mu < 1.96\sigma/\sqrt{n}\right] = 0.95$$

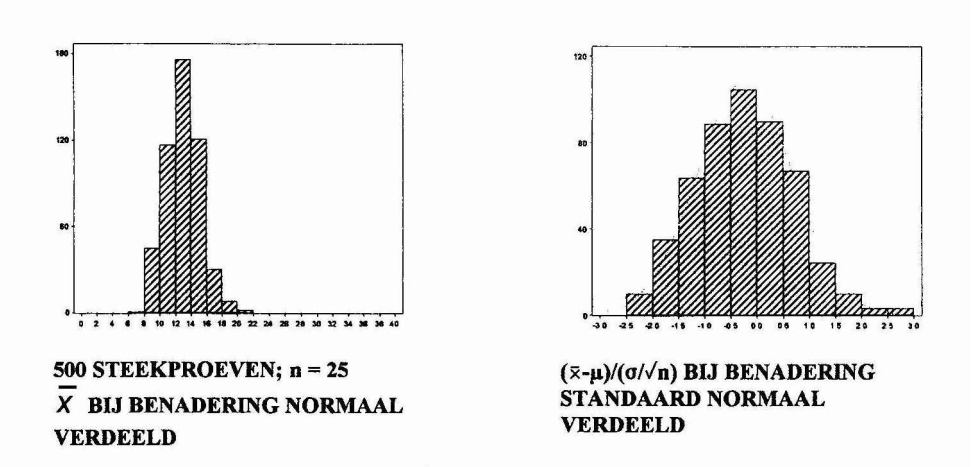
$$\mathbb{P}\left[\overline{X} - 1.96\sigma/\sqrt{n} < \mu < \overline{X} + 1.96\sigma/\sqrt{n}\right] = 0.95$$

Dus met waarschijnlijkheid van 95% ligt de echte waarde μ in het interval

$$[\overline{X} - 1.96\sigma/\sqrt{n}, \overline{X} + 1.96\sigma/\sqrt{n}]$$

Dus een 95% betrouwbaarheidsinterval wordt gegeven door:

$$\mu = \overline{X} \pm 1.96 \cdot \sigma/\sqrt{n}$$



Gewichten vliegende herten: $\mu = 13.58, \sigma/\sqrt{n} = 2.42$

Definitie van z_α: Onder z_α verstaan we het positieve getal waarvoor geldt:

$$\mathbb{P}[Z \geq z_\alpha] = \alpha \ .$$

Dit is de grenswaarde waar volgens de nulhypothese met kans $1-\alpha$ de waarneming onder ligt. Dus ook die **waarboven** we bij een toets met significantie α (éénzijdig) de nulhypothese verwerpen. We merken nog op dat, wegens de symmetrie van de normale verdeling:

$$\mathbb{P}[|Z| \geq z_\alpha] = 2\alpha \ . \tag{4.6}$$

Waarschuwing: In de literatuur wordt vaak gesproken van $z_{95\%}$ in plaats van $z_{2,5\%}$. Wij zullen ook van deze conventie gebruik maken, meestal in de context van betrouwbaarheidsintervallen.
Vuistregel: γ ligt dicht bij 1, en α dicht bij 0. Ze voldoen aan

$$\gamma = 1 - 2\alpha \ .$$

Onder z_γ of z_α wordt vaak *hetzelfde* getal verstaan. namelijk het getal z waarvoor (4.6) geldt, maar ook

$$\mathbb{P}[|Z| \leq z] = \gamma \ .$$

De laatste notatie is meer in gebruik bij betrouwbaarheidsintervallen.

Berekening van z_α met de GR: Roep middels `2nd DISTR 3` de functie `invNorm` op. Typ de waarde van $1-\alpha$ in. Uitkomst: z_α.

Voorbeeld: Isomerase-activiteit De gemiddelde isomerase-activiteit bij 4 muizen was 0.76 met een standaarddeviatie 0.18. Wat kunnen we hieruit concluderen mbt de isomerase-activiteit bij muizen in het algemeen?

- stel: $\sigma = 0.18$

- dan is $1.96 \cdot \sigma/\sqrt{n} = 1.96 \cdot 0.18/\sqrt{4} = 0.176$

- 95%-betrouwbaarheidsinterval: $\mu = 0.76 \pm 0.176$.

Dus: met 95% waarschijnlijkheid ligt de gemiddelde isomerase-activiteit bij muizen tussen 0.584 en 0.936.

Betrouwbaarheidsintervallen voor andere niveaus dan 95% Ook voor andere betrouwbaarheidsniveau's dan $\gamma = 0.95$ kunnen intervallen voor μ bepaald worden; alleen moet dan de waarde $z_{0.95} = 1.96$ vervangen worden:

- 95%-betrouwbaarheidsinterval: $\mu = \overline{X} \pm 1.96\sigma/\sqrt{n}$
- 99%-betrouwbaarheidsinterval: $\mu = \overline{X} \pm 2.58\sigma/\sqrt{n}$
- 99.9%-betrouwbaarheidsinterval: $\mu = \overline{X} \pm 3.29\sigma/\sqrt{n}$

Grotere betrouwbaarheid van de schatting van het gemiddelde gaat dus ten koste van de precisie.

Stel: $\sigma = 2,\ n = 25$ en $\overline{X} = 12.1$
95%: $\mu = 12.1 \pm 1.96 \cdot 2/\sqrt{25} = 12.1 \pm 0.78$
ofwel $\mu \in [11.32, 12.88]$
99%: $\mu = 12.1 \pm 2.58 \cdot 2/\sqrt{25} = 12.1 \pm 1.03$
ofwel $\mu \in [11.07, 13.13]$

4.4 De t-verdeling

Bij de toepassing van de centrale limietstelling komen we het volgende probleem tegen:

Probleem: In de praktijk is de populatie-standaarddeviatie σ niet bekend!

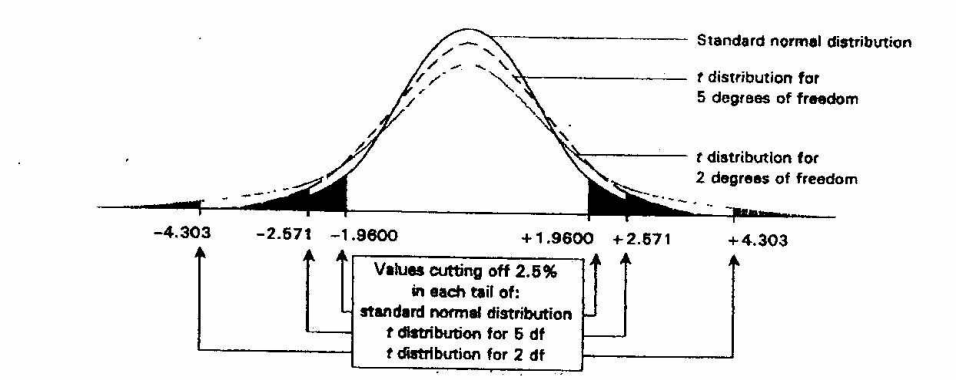
Oplossing voor grote n:

- voor grote steekproeven ($n \geq 100$) geldt met zeer grote kans: $S \approx \sigma$.
- Dan is $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ is bij benadering standaard normaal verdeeld,
- en dus wordt een γ-betrouwbaarheidsinterval gegeven door: $\mu = \overline{x} \pm z_\gamma \cdot s/\sqrt{n}$

Stelling 2. (Centrale limietstelling met σ onbekend): *Stel dat we bij een steekproef met grootte n, uit een normaal verdeelde populatie getrokken, een gemiddelde $\overline{X}$ en een standaarddeviatie s hebben gemeten. Dan geldt:*

- $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ *volgt student's t-verdeling met* $\nu = n-1$ *vrijheidsgraden,*
- *dus een* γ*-betrouwbaarheidsinterval is:* $\mu = \overline{x} \pm t_\gamma \cdot s/\sqrt{n}$

Hierbij is de *t-verdeling* een op de normale verdeling gelijkende klokvorm, met echter een aanzienlijk dikkere staart.



Grafieken van t-verdelingsfunctie bij ν vrijheidsgraden

4.4.1 Definitie van t_γ en t_α:

Deze is net als die van z_γ en z_α, zij het dat er nu een tweede parameter bij komt: het aantal vrijheidsgraden ν. We schrijven dan ook wel $t_{\nu,\alpha}$ en $t_{\nu,\gamma}$. Veronderstelling is steeds dat α dicht bij 0 en γ dicht bij 1 ligt, en dat $\gamma + 2\alpha = 1$. Definities:

$$\mathbb{P}[T \geq t_\alpha] = \alpha \quad \text{en} \quad \mathbb{P}[|T| \leq t_\gamma] = \gamma \; .$$

Voorbeeld: γ-betrouwbaarheidsinterval voor de aktiviteit van het enzym isomerase in het spierweefsel van verschillende diersoorten

$\gamma = 0.95$	n	s	s.e.	t_γ	$\overline{X} \pm t_\gamma \cdot$ s.e.
muis	4	0.18	0.09	3.182	0.76 ± 0.29
kikker	6	0.16	0.06	2.571	1.53 ± 0.17
kreeft	6	0.24	0.10	2.571	1.06 ± 0.25
forel	10	0.60	0.19	2.262	4.22 ± 0.43

Eigenschappen van de t-verdeling:

- Ze hangt af van de steekproefgrootte ($\nu = n - 1$);
- is voor elke ν symmetrisch rond haar gemiddelde 0: $f_\nu(-x) = f_\nu(x)$.
- De variantie is altijd groter dan 1, en neemt af bij toenemende ν;
- De verdeling nadert tot standaard-normale verdeling (z-verdeling) als ν heel groot wordt.
- In paragraaf 4.6 staan de kritische waarden van de t-verdeling getabelleerd voor enkele waarden van de significantie-drempel α.
- t-verdeling op de TI-84 Plus: `2nd DISTR 6` levert: `tcdf(`. Intypen van a, b en νd levert de kans dat $a \leq T \leq b$ voor een t-verdeelde stochast met ν vrijheidsgraden.
- t_α op deze machine: `2nd DISTR 4` levert: `invT(`. Intypen van $1-\alpha$ en ν levert $t_{\nu,\alpha}$. (En trouwens: intypen van α levert $-t_{\nu,\alpha}$.) Merk op dat de normale verdeling hierin bevat is: Voor $\nu \to \infty$ nadert $t_{\nu,\alpha}$ tot z_α.

Voorwaarden voor toepassing van de t-verdeling

- de oorspronkelijke populatie is normaal verdeeld,
- òf de oorspronkelijke verdeling is tenminste symmetrisch en de steekproef is niet te klein ($n \geq 10$),
- òf de steekproef is vrij groot ($n \geq 30$).

Voorbeeld: Een etholooog die met kuikens werkt wil de temperatuur van de broedmachine zodanig afstellen dat de eieren gemiddeld na 504 uur (21 dagen) uitkomen. Hij stelt de machine in, broedt een monster van 25 eieren uit en vindt dat deze gemiddeld na 522 uur uitkomen met een standaarddeviatie van 40 uur. Kan de etholoog op basis van deze uitkomst 90% zeker zijn dat de broedmachine verkeerd is afgesteld?

- o standaardfout: s.e. $= \frac{40}{\sqrt{25}} = 8$
- o aantal vrijheidsgraden: $\nu = 25 - 1 = 24$
- o kritieke t-waarde: $t_{90\%} = 1.711$
- o 90%-betrouwbaarheidsinterval: $522 \pm 1.711 = [508.3, 535.7]$

De streefwaarde 504 uur valt buiten dit interval, dus met 90% zekerheid is de ontwikkeling van de eieren te langzaam, dus de temperatuur van de broedmachine moet omhoog.

4.5 Tabel van de standaard-normale verdeling

Handleiding: Deze tabel geeft de kans dat $Z \leq x$ voor een standaard-normaal verdeelde stochast Z. In de eerste kolom staan de cijfers van x aan weerszijden van de komma. Het tweede cijfer van x achter de komma staat boven de kolom met de uitkomst. Voor de cijfers hiervan moet '0,' worden gezet. 893 betekent dus: 0,893.
Voorbeeld: $\mathbb{P}[Z \leq 1{,}24] = 0{,}893$.
Met de GR: typ `2nd DISTR 2 (-) 2nd EE 99 , 1.24 ) ENTER`. Dit levert de kans dat Z inligt tussen $-10^{99} \approx -\infty$ en 1,24.

x	0	1	2	3	4	5	6	7	8	9
0,0	500	504	508	512	516	520	524	528	532	536
0,1	540	544	548	552	556	560	564	567	571	575
0,2	579	583	587	591	595	599	603	606	610	614
0,3	618	622	626	629	633	637	641	644	648	688
0,4	655	659	663	666	670	674	677	681	684	688
0,5	691	695	698	702	705	709	712	716	719	722
0,6	726	729	732	736	739	742	745	749	752	755
0,7	758	761	764	767	770	773	776	779	782	785
0,8	788	791	794	797	800	802	805	808	811	813
0,9	816	819	821	824	826	829	831	834	836	839
1,0	841	844	846	848	851	853	855	858	860	862
1,1	864	867	869	871	873	875	877	879	881	883
1,2	885	887	889	891	893	894	896	898	900	901
1,3	903	905	907	908	910	911	913	915	916	918
1,4	919	921	922	924	925	926	928	929	931	932
1,5	933	934	936	937	938	939	941	942	943	944
1,6	945	946	947	948	949	951	952	953	954	954
1,7	955	956	957	958	959	960	961	962	962	963
1,8	964	965	966	966	967	968	969	969	970	971
1,9	971	972	973	973	974	974	975	976	976	977
2,0	977	978	978	979	979	980	980	981	981	982
2,1	982	983	983	983	984	984	985	985	985	986
2,2	9861	9864	9868	9871	9875	9878	9881	9884	9887	9890
2,3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2,4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2,5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2,6	9952	9955	9956	9957	9959	9960	9961	9962	9963	9964
2,7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2,8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2,9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986

4.6 Kritieke waarden voor Student t-verdelingen

Voor de gevallen $\alpha = 0.05,\ 0.025,\ 0.01,\ 0.005$ wordt de grenswaarde t_α gegeven met de eigenschap dat $\mathbb{P}[T \geq t_\alpha] = \alpha$ voor een Student-t-verdeelde stochast T met ν vrijheidsgraden. Hiermee kan bij een toets het verwerpingsbereik van de nulhypothese worden vastgesteld. De standaard normale verdeling is de t-verdeling met oneindig veel vrijheidsgraden: $\nu = \infty$.
Met de GR: typ `2nd DISTR 4` en geef de parameterwaarden $1-\alpha$ en ν in, gescheiden door een komma en gevolgd door `) ENTER`. De grenswaarde t_α verschijnt.

ν	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.28	2.845
25	1.708	2.060	2.485	2.787
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
50	1.676	2.009	2.403	2.678
∞	1.645	1.960	2.326	2.576

Hoofdstuk 5

Discrete kansverdelingen

Kansverdelingen zijn voorschriften waaruit men de kans op bepaalde gebeurtenissen kan berekenen. Voor een continue kansverdeling gaat het om een kansdichtheidsfunctie, waarmee we door integratie de kans kunnen uitrekenen dat een uitkomst tussen bepaalde waarden in ligt.
Voor een *discrete* kansverdeling is het veel eenvoudiger: we moeten voor elke waarde van de grootheid gewoon de kans aangeven dat hij wordt aangenomen. Sommeren is gemakkelijker dan integreren!
In dit hoofdstuk behandelen we enkele belangrijke discrete kansverdelingen:

1. Uniforme verdeling (paragraaf 5.1)
2. Binomiale verdeling (paragraaf 5.2)
3. Poisson-verdeling (paragraaf 5.3)

5.1 De discrete uniforme verdeling.

Een biologisch voorbeeld: Wat zou het foerageergedrag van stekelbaarsjes zijn als ze zich niet door predatoren lieten beïnvloeden?

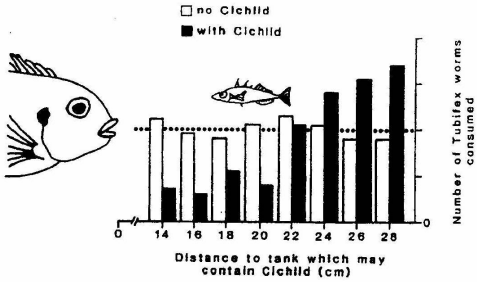


Fig. 12.1 The relationship between the number of *Tubifex* worms eaten by sticklebacks and the distance to a tank which sometimes contains a cichlid predator. (After Milinski 1985.)

Een kansrekening-voorbeeld: een zuivere dobbelsteen.

- Mogelijke uitkomsten $k = 1, \ldots, 6$ uniform verdeeld: $p_k = \frac{1}{6}$.
- Gemiddelde: $\mu = \frac{1}{6} + \frac{2}{6} + \cdots + \frac{6}{6} = \frac{21}{6} = 3.5$.
- Variantie: $\sigma^2 = (1 - 3.5)^2/6 + \cdots + (6 - 3.5)^2/6 = \frac{35}{12} \approx 2.92$.
- Standaarddeviatie: $\sigma \approx \sqrt{\frac{35}{12}} \approx 1.71$.

Algemeen:

- n mogelijke uitkomsten: $y_k,\ k = 1 \ldots n$. (Bijvoorbeeld: $y_k = k$.)
- Kans op uitkomst y_k: $p_k = \mathbb{P}[Y = k] = \frac{1}{n}$.
- Gemiddelde: $\mu = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} y_k$. (In het voorbeeld: $\frac{n+1}{2}$.)
- Variantie: $\sigma^2 = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} (y_k - \mu)^2$. (In het voorbeeld: $\frac{n^2-1}{12}$.)
- standaarddeviatie σ. (In het voorbeeld: $\sqrt{(n^2 - 1)/12}$.)

5.2 De binomiale verdeling

Deze heben we al gezien in § 2.4.1. We geven hier nog een paar voorbeelden.

1. Als 50% van alle nieuwgeboren kinderen jongetjes zijn, in hoeveel procent van alle gezinnen met vijf kinderen zijn dan twee van de kinderen jongetjes en drie meisjes?
2. Als 40% van de vlinders in een populatie blauw zijn en 60% bruin, wat is dan de kans dat van vijf gevangen vlinders twee blauw zijn en drie bruin?
3. Als 40% van alle proefdieren besmet zijn met een virus, wat is dan de kans dat precies twee van de vijf proefdieren in een gegeven experiment besmet zijn?

Algemeen:

- binaire klassificatie: twee mogelijke kategorieën A en B (bijv. man/vrouw, blauw/bruin, geïnfecteerd/niet geïnfecteerd)
- n onafhankelijke trekkingen die elk òf A òf B zijn (hier: $n = 5$)
- elke trekking heeft dezelfde kans p om A te zijn (hier: $p = 0.4$)

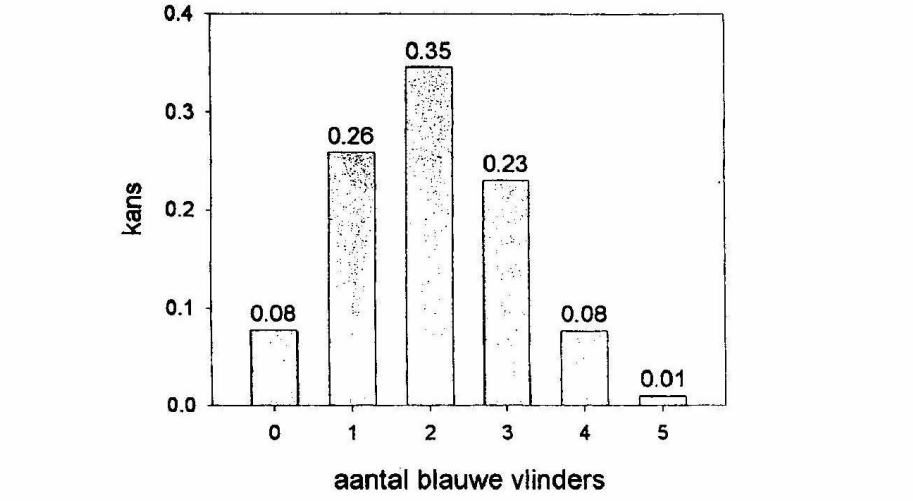
In al deze gevallen wordt de kans dat precies k van de n objecten van type A zijn gegeven door de "binomiale verdeling"

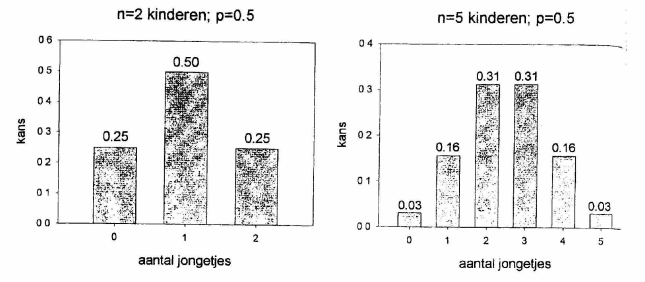
$$\pi_k = \mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Voor het 'grijpgetal' $\binom{n}{k}$ kan de volgende formule worden gegeven:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k \cdot (k-1) \cdots 2 \cdot 1} \ .$$

(Onder en boven staan k factoren.)
We geven de binomiale verdeling met parameters p en n aan met $\mathrm{Bin}(n, p)$.

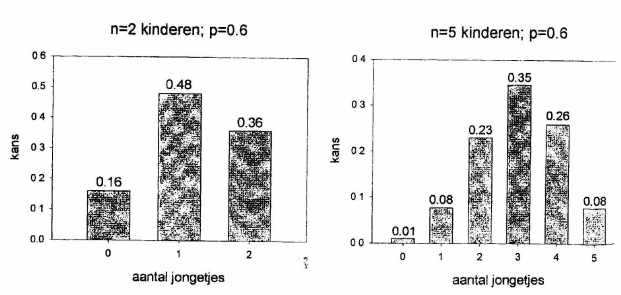




5.2.1 Eigenschappen van de binomiale verdeling:

verwachtingswaarde : $\mu = \sum_{k=1}^{n} \pi_k \cdot k = np$;

variantie : $\sigma^2 = \sum_{k=0}^{n} \pi_k \cdot (k-\mu)^2 = np(1-p)$ (Zie § 4.1.2.)



5.3 De Poisson-verdeling

De Poissonverdeling is als een binomiale verdeling met heel grote n en heel kleine p. Het product $n \times p$, dat we met λ aangeven, hoeft niet speciaal groot of klein te zijn. Een voorbeeld:

Verdeling van typfouten. je typt een aantal pagina's tekst, zeg 2000 letters per pagina. Elke letter kun je zien als een poging om de juiste letter uit een bepaalde bedoelde tekst weer te geven, maar nu en dan gaat het fout: je maakt een typfout. Dit gebeurt misschien bij één op de 1000 letters. Het aantal typfouten per pagina is dan binomiaal verdeeld met $n = 2000$ en $p = \frac{1}{1000}$. Gemiddeld leidt dat tot $\lambda = np = \frac{2000}{1000} = 2$ typfouten per pagina. De kans op k typfouten op pagina 1 is dan:

$$\begin{aligned}
\pi_k &= \binom{2000}{k}\left(\frac{1}{1000}\right)^k\left(1-\frac{1}{1000}\right)^{2000-k} \\
&= \frac{2000\times 1999\times\cdots\times(2000-k+1)}{k!}\cdot\left(\frac{1}{1000}\right)^k\left(1-\frac{1}{1000}\right)^{2000-k} \\
&= \frac{2000}{1000}\cdot\frac{1999}{1000}\cdots\frac{2000-k+1}{1000}\times\frac{1}{k!}\cdot\left(1-\frac{1}{1000}\right)^{2000-k}
\end{aligned}$$

Deze formule begint met k factoren die ieder ongeveer 2 zijn, en eindigt met

$$\left(1-\frac{1}{1000}\right)^{2000} \approx e^{-2}\ .$$

Al met al krijgen we dus

$$\pi_k \approx \frac{2^k}{k!}e^{-2}\ .$$

We noemen dit *de Poissonverdeling* met parameter 2.

Definitie van de Poissonverdeling. We zeggen dat een geheeltallige stochastische variabele Y *Poisson-verdeeld* is met parameter λ als

$$\pi_k := \mathbb{P}[Y = k] = \frac{\lambda^k}{k!} e^{-\lambda} \ . \tag{5.1}$$

Deze verdeling heeft een heel groot geldigheidsgebied. In plaats van typfouten op een pagina kan het gaan om willekeurige gebeurtenissen, geteld per eenheid van tijd of ruimte.
Voorbeelden:

- Als agarplaten gemiddeld drie bacterie-kolonies bevatten, wat is de kans dat een gegeven plaat twee kolonies bevat?

- Als bij een plant per uur gemiddeld drie hommels langskomen, wat is dan de kans dat op een gegeven dag tussen 12.30 en 13.30 uur precies twee hommels langskomen?

- Als een uil per nacht gemiddeld drie muizen vangt, wat is dan de kans dat dat ze in een bepaalde nacht twee muizen vangt?

Het gaat hierbij dus om:

- 'gebeurtenissen' die totaal random verdeeld zijn over tijd of ruimte (het vangen van een muis, het langskomen van een hommel, het ontstaan van een bacteriekolonie);

- terwijl het gemiddelde aantal gebeurtenissen per tijds- of ruimte-eenheid gelijk is aan λ.

Dan wordt de kans dat in een bepaalde tijds- of ruimte-eenheid precies k gebeurtenissen optreden gegeven door de *Poissonverdeling* (5.1) met parameter λ.

Voorbeeld: $\lambda = 3,\ k = 2$. Dan geldt:

$$\mathbb{P}[Y = 0] = e^{-3} \approx 0{,}049787 \ .$$

$$\mathbb{P}(Y = 2) = \frac{3^2}{2!} e^{-3} = \frac{9}{2} e^{-3} \approx 0{,}224$$

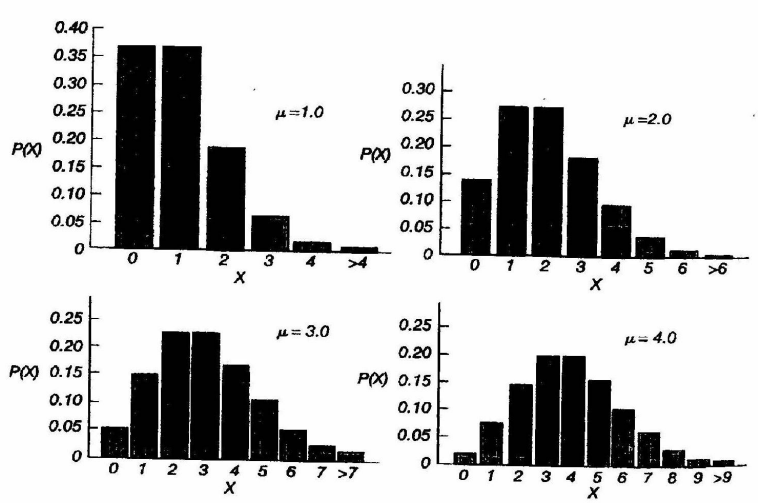


Figure 24.1 The Poisson distribution for various values of μ. These graphs were prepared by using Equation 24.1.

5.3.1 Eigenschappen van de Poissonverdeling.

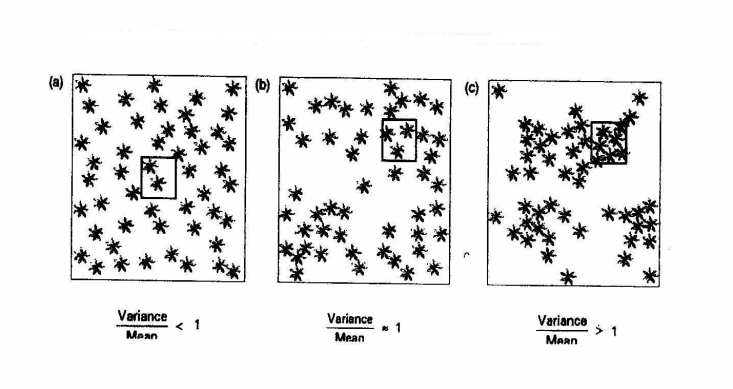
$$\text{Verwachte waarde} \quad : \quad \mu = \sum_k \pi_k \cdot k = \lambda \; ;$$

$$\text{Variantie} \quad : \quad \sigma^2 = \sum_k \pi_k (k-\mu)^2 = \lambda \; .$$

De variantie is dus gelijk aan het gemiddelde! Dit is de aanleiding voor de volgende definitie:

5.3.2 De dispersie-index.

Bij gebeurtenissen die in de tijd of in de ruimte verdeeld zijn wordt de *dispersie-index* $I = \sigma^2/\mu$ gebruikt als maat voor de wijze van verdeling. Als $I < 1$ duidt dit op een regelmatige (min of meer uniforme) verdeling; als $I > 1$ op een geclusterde verdeling. Bij een volkomen random verdeling is $I \approx 1$.



Voorbeelden:

- Verdeling nesten in broedkolonie vogels: regelmatig doordat de vogels afstand houden tot elkaar. Dus $I < 1$.
- Verdeling van eieren van insekten die in pakketjes leggen: deze is geclusterd, dus $I > 1$.
- Verdeling paardebloemen in een weiland: de bloemen trekken zich niets van elkaar aan, zodat ze een Poissonverdeling volgen: $I = 1$.

5.4 'Goodness-of-fit': het toetsen van discrete verdelingen

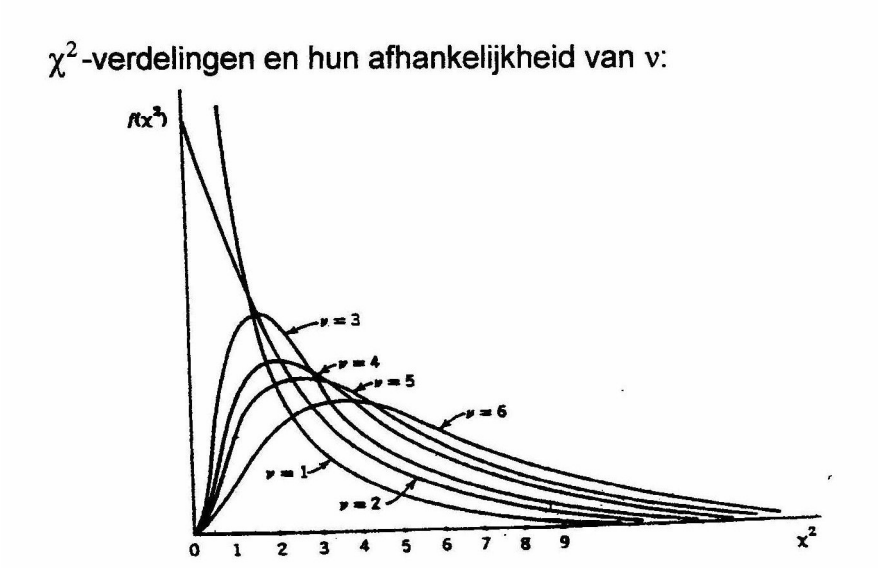
In Hoofsdtuk 2 hebben we gekeken hoe je de hypothese kunt toetsen, dat pissebedden geen voorkeur hebben tussen *twee* mogelijkheden. Nu stellen we ons dezelfde vraag opnieuw, maar nu met *een willekeurig aantal* mogelijkheden. We toetsen dan of de uniforme verdeling over een aantal alternatieven als nulhypothese kan worden gehandhaafd in het licht van experimentele resultaten. Dit is het onderwerp van Paragraaf 5.4.1. Hetzelfde kunnen we ook doen voor de binomiale verdeling: Paragraaf 5.4.3, of de Poissonverdeling: Paragraaf 5.4.5.
Maar eerst hebben we een nieuw theoretisch instrument nodig, een soort meerdimensionale normele verdeling. Dit is het onderwerp van de volgende paragraaf:

5.4.1 De χ^2-verdeling

Definitie: De kansverdeling van de som van de kwadraten van ν onafhankelijke standaard-normaal verdeelde stochasten

$$Z_1^2 + Z_2^2 + \ldots + Z_\nu^2$$

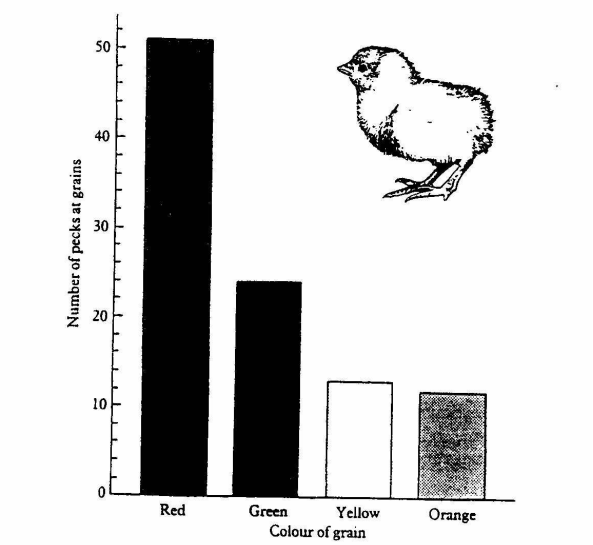
wordt de χ^2-verdeling met ν vrijheidsgraden genoemd.



Berekening van χ^2-kansen met de TI-84 Plus: Zoek de cumulatieve verdelingsfunctie met `2nd DISTR 8`. De kans dat $a \leq X^2 \leq b$ wordt gegeven door als argumenten in te voeren: a, b en d, waarbij d het aantal vrijheidsgraden is. Als je de kans wilt weten dat $X^2 \geq a$ dan vul je voor b in: 10^{99}, via `2nd EE`. De tabel in paragraaf 5.6 geeft de kritische waarden aan van deze verdeling bij diverse vrijheidsgraden. Dat wil zeggen: er wordt aangegeven waar de grenzen van 95% respectievelijk 99% worden overschreden.

5.4.2 Het toetsen van de uniforme verdeling: voorkeur van kuikens.

Kuikens kunnen kiezen tussen graankorrels van vier kleuren: rood, groen, geel en oranje. Bij een experiment kiest een kuiken: 51 rood, 24 groen, 13 geel en 12 oranje.



Als het kuiken geen enkele voorkeur heeft voor een van de kleuren, hoe onwaarschijnlijk is dan de waargenomen verdeling?

Stelling: Als de verdeling van waarnemingen over m kategorieën door een kansproces tot stand komt (bijvoorbeeld de uniforme verdeling), dan is de toetsgrootheid

$$X^2 := \sum_{i=1}^{m} \frac{(o_i - e_i)^2}{e_i} \tag{5.2}$$

bij benadering verdeeld volgens de "χ^2-verdeling met $\nu = m - 1$ vrijheidsgraden".

Hier is o_i het geobserveerde aantal, en e_i het verwachte aantal volgens de veronderstelde kansverdeling (in dit geval volgens de uniforme verdeling, dus $e_i = 25$ voor $i = 1, 2, 3, 4$).
We maken de berekening in een tabel in het geval van het voorbeeld van de kuikens:

kleur graankorrels	o_i	e_i	$\frac{(o_i - e_i)^2}{e_i}$
rood	51	25	27.04
groen	24	25	0.04
geel	13	25	5.76
oranje	12	25	6.76

Dus $X^2 = 27.04 + 0.04 + 5.76 + 6.76 = 39.60$.

Als kuikens geen voorkeur voor een bepaalde kleur hebben dan is X^2 verdeeld volgens de χ^2-verdelingsfunctie met $\nu = 4 - 1 = 3$ vrijheidsgraden. Met behulp van die verdeling kun je bepalen hoe groot de kans P is dat de gevonden X^2-waarde of een grotere puur op grond van het toeval zou optreden. In ons geval: $P = 1.3 \cdot 10^{-8}$. De gevonden uitkomst is dus uiterst onwaarschijnlijk als het pikgedrag kleur-onafhankelijk zou zijn.

5.4.3 Het toetsen van de binomiale verdeling: Geslacht van kinderen.

Hoe toevallig is de geslachtbepaling bij de mens?

- Biologische hypothese: in elk gezin heeft elk kind een kans 0.5 om een jongetje dan wel een meisje te worden, onafhankelijk van het geslacht van zijn broertjes en zusjes.
- Statistische hypothese: de kans dat in een gezin met n nakomelingen k jongetjes geboren worden, wordt gegeven door de binomiale verdeling met $p = 0.5$
- Gegevens die Geissler in 1889 verzamelde: aantallen jongetjes in 6115 gezinnen met twaalf kinderen.

Op grond van de nulhypothese zou het aantal gezinnen met k jongetjes hieruit gelijk zijn aan $6115\pi_k$ met $\pi_k = \binom{12}{k}/2^{12}$.

aantal	kans	verwacht (e)	gevonden (o)	$\frac{(o-e)^2}{e}$
0	0.0002	1.5	3	1.5
1	0.0029	17.9	24	2.1
2	0.0161	98.5	104	0.3
3	0.0537	328.4	286	5.5
4	0.1209	739.0	670	6.4
5	0.1934	1182.4	1033	18.9
6	0.2256	1379.5	1343	1.0
7	0.1934	1182.4	1112	4.2
8	0.1209	739.0	829	11.0
9	0.0537	328.4	478	68.1
10	0.0161	98.5	181	69.1
11	0.0029	17.9	45	41.0
12	0.0002	1.5	7	20.2
totaal	1.0000	6114.9	6115	$X^2 = 249.2$

Bij een χ^2-verdeling met $13 - 1 = 12$ vrijheidsgraden is deze uitkomst uiterst zeldzaam: De kans op $X^2 > 249.2$ is ongeveer $1.4 \cdot 10^{-48}$. We verwerpen de nulhypothese! Maar wat is er dan aan de hand? Een mogelijke verklaring zou zijn: p is niet 0.5 maar een beetje meer, zeg 0.52 (38100 van de $12 \times 6115 = 73380$ kinderen waren immers jongetjes). X^2 berekend op basis van de binomiale verdeling met $p = 0.52$ geeft 116.6. Deze uitkomst is nog steeds uiterst zeldzaam: de p-waarde is ongeveer $8.7 \cdot 10^{-20}$. Deze kans wordt berekend met de χ^2-verdeling met 11 vrijheidsgraden, want de waarde $p = 0.52$ is ontleend aan de steekproef, hetgeen een verlies van één vrijheidsgraad tot gevolg heeft (zie verderop).

aantal	kans	verwacht (e)	gevonden (o)	$\frac{(o-e)^2}{e}$
0	0.0001	0.6	3	9.6
1	0.0019	11.6	24	13.3
2	0.0116	70.9	104	15.5
3	0.0418	255.6	286	3.6
4	0.1020	623.7	670	3.4
5	0.1768	1081.1	1033	2.1
6	0.2234	1366.1	1343	0.4
7	0.2075	1268.9	1112	19.4
8	0.1405	859.2	829	1.1
9	0.0676	413.4	478	10.1
10	0.0220	134.5	181	16.1
11	0.0043	26.3	45	13.3
12	0.0004	2.4	7	8.8
totaal	0.9998	6113.7	6115	$X^2 = 116.6$

De verklaring van Geissler was een andere: in sommige gezinnen is de kans op een jongetje duidelijk groter dan 0.5, maar in andere gezinnen juist kleiner.

5.4.4 Bepaling van het aantal vrijheidsgraden

De volgende regel moet worden gehanteerd: Stel je bepaalt een toetsingsgrootheid Y (bijvoorbeeld $X^2 = \sum \frac{(o-e)^2}{e}$) met behulp van

- m klassen van waarnemingen,
- j statistische grootheden die je uit de waarnemingen al eerder hebt berekend,

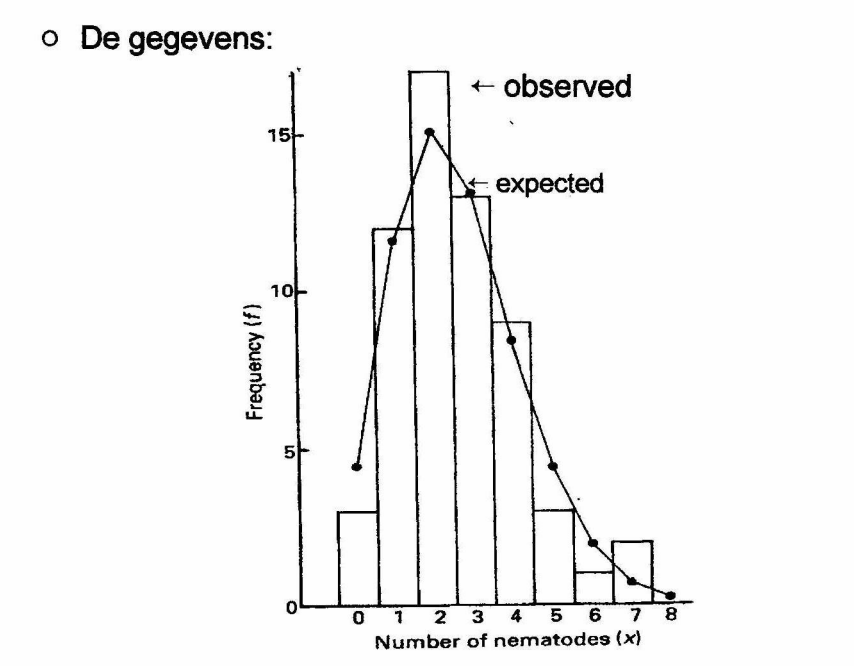
dan is Y een toetsingsgrootheid met $\nu = m - j - 1$ vrijheidsgraden.

χ^2-analyse: Je verliest altijd één vrijheidsgraad omdat je de *totale* omvang van de steekproef nodig hebt om de verwachte frequenties te bepalen.
Als de kansverdeling waartegen je wilt toetsen *van tevoren* bekend is ("extrinsieke" hypothese), dan hoef je verder geen parameters meer te schatten en kom je uit op een χ^2-verdeling met $\nu = m - 1$ vrijheidsgraden. Dit is het geval bij toetsing van de uniforme verdeling, of de binomiale verdeling met gegeven p.
Als je daarentegen voor deze kansverdeling nog j parameters aan de hand van de gegevens moet schatten ("intrinsieke hypothese"), dan verlies je nog eens j vrijheidsgraden, en moet je een χ^2-verdeling hanteren met $\nu = m - j - 1$ vrijheidsgraden. Dit is het geval bij toetsing van de binomiale verdeling met onbekende p, en meestal bij toetsing van de Poissonverdeling: je haalt de te toetsen p respectievelijk λ *uit de gegevens*.

5.4.5 De verdeling van nematoden over de ruimte

- Biologische nulhypothese: nematoden verdelen zich random over de ruimte.

- Statistische nulhypothese: de kans dat per oppervlakte-eenheid k nematoden worden aangetroffen wordt gegeven door een Poissonverdeling waarvan de parameter λ overeenkomt met het *gemiddelde* aantal nematoden per oppervlakte-eenheid.



Het plaatje suggereert: er is maar een geringe afwijking van een Poisson-verdeling, die waarschijnlijk wel aan het toeval is toe te schrijven. Dit gaan we nu precies bekijken.

- 156 nematoden waargenomen $(= 3 \times 0 + 12 \times 1 + 17 \times 2 + \ldots)$

- gemiddelde aantal per cel: $156/60 = 2.6$, dus we gaan een Poissonverdeling met $\lambda = 2.6$ gebruiken.

- Na poolen van de kategorieën 5,6,7: $X^2 = 1.04$

- $\nu = 6 - 1 - 1 = 4$ vrijheidsgraden

aantal	gevonden	Poissonkans	verwacht	$\frac{(o-e)^2}{e}$
0	3	0.0743	4.5	0.50
1	12	0.1931	11.6	0.01
2	17	0.2510	15.1	0.24
3	13	0.2176	13.1	0.00
4	9	0.1414	8.5	0.03
5/6/7	6	0.1226	7.4	0.26
totaal:	60	1.000	60.2	$X^2 = 1.04$

Conclusie: Dit is niet uitzonderlijk. De kans op deze waarde of hoger is 0.9.

Opmerkingen

- Bij de χ^2-toets van 'goodness-of-fit' vul je alleen aantallen in, nooit percentages, proporties of meetuitslagen.
- Een χ^2-verdeling is niet van toepassing als de verwachte frequenties van bepaalde kategorieën te klein zijn. Vuistregel: geen verwachte frequentie mag kleiner zijn dan 1 en niet meer dan 20% van de verwachte frequenties mogen kleiner zijn dan 5. Zonodig dien je kategorieën te poolen.

5.5 De χ^2-toets op onafhankelijkheid

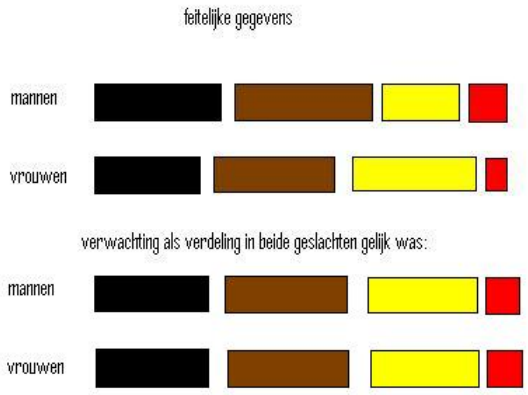
We kunnen de χ^2-verdeling ook gebruiken wanneer we *twee* geheeltallige toevalsgrootheden hebben gemeten, en we willen weten of er een verband, een statistische afhankelijkheid tussen deze grootheden bestaat.
De nulhypothese zal steeds zijn: er is geen verband. Consequentie hiervan zou zijn, dat als grootheid A bijvoorbeeld de waarde 1 heeft, de kansverdeling van de grootheid B dezelfde is, als wanneer A de waarde 2 heeft.
Het toetsen van onafhankelijkheid betekent dus eigenlijk: nagaan of *twee of meer* gevonden frequentieverdelingen overeenkomen *met dezelfde kansverdeling.*

5.5.1 Hebben mannen en vrouwen dezelfde verdeling van haarkleur?

	zwart:	bruin:	blond:	rood:	totaal:
mannen:	32	43	16	9	100
vrouwen:	55	65	64	16	200
totaal:	87	108	80	25	300

Percentages:

	zwart:	bruin:	blond:	rood:	totaal:
mannen:	32.0%	43.0%	16.0%	9.0%	33.3%
vrouwen:	27.5%	32.5%	32.0%	8.0%	66.7%



Nulhypothese: de verdeling van haarkleuren is bij mannen en vrouwen hetzelfde; de gevonden verschillen zijn door toeval ontstaan.

Verwachte frequenties op basis van nulhypothese:

- aantal mannen: 100
  fractie individuen met blond haar: $0.267 = 80/300$

- Als geslacht en haarkleur *onafhankelijk* van elkaar zijn, zou je $0.267 \cdot 100 = 26.7$ mannen met blond haar verwachten.

- Merk op dat
$$26.7 = \frac{100}{300} \cdot \frac{80}{300} \cdot 300 = \frac{r_1 \cdot k_3}{N}$$

Algemene rekenformule: Als de rij- en kolomeigenschap *onafhankelijk* van elkaar zijn (zoals de nulhypothese veronderstelt), dan wordt de verwachte frequentie in cel c_{ij} gegeven door $\frac{r_i \cdot k_j}{N}$. Hierbij is r_i het totaal van de i-de rij, en k_j het totaal van de j-de kolom, en N is de steekproefgrootte.

We gaan als volgt te werk:

De χ^2-statistiek X^2 voor onafhankelijkheid:

- maak een "kruistabel" (contingency table) die naast de gegevens ook de rijtotalen, de kolomtotalen en de steekproefgrootte bevat
- bereken voor elke cel de verwachte frequentie op basis van de nulhypothese (onafhankelijkheid): verwacht=rijtotaal x kolomtotaal /steekproefgrootte
- vorm de X^2-waarde

Als rij- en kolomeigenschap onafhankelijk zijn, is X^2 verdeeld volgens de χ^2-verdeling met $\nu = (r-1)(k-1)$ vrijheidsgraden.

Het voorbeeld: sekse en haarkleur

	zwart:	bruin:	blond:	rood:	totaal:
mannen:	32/29.0	43/36.0	16/26.7	9/8.3	100
vrouwen:	55/58.0	65/72.0	64/53.3	16/16.7	200
totaal:	87	108	80	25	300

$$X^2 = \frac{(32-29.0)^2}{29.0} + \frac{(43-36.0)^2}{36.0} + \cdots + \frac{(16-16.7)^2}{16.7} = 8.987$$

Aantal vrijheidsgraden $\nu = (2-1)\cdot(4-1) = 3$, dus kans op $X^2 \geq 8.987$ is 0.029.

Conclusie: sekse en haarkleur zijn bij de mens waarschijnlijk *niet* onafhankelijk van elkaar; ten opzichte van de toevalsverwachting zijn vrouwen met blond haar en mannen met zwart of bruin haar oververtegenwoordigd.

5.5.2 Meniscusmuggen (*Dixa*) als indicatoren voor waterkwaliteit

	nebulosa	submaculata	dilatata	nubilipennis	total
oligotrphic	12/15.93	7/6.53	5/8.88	17/9.66	41
mesotrophic	14/19.82	6/8.12	22/11.04	9/12.02	51
eutrophic	35/25.25	12/10.35	7/14.08	11/15.32	65
total:	61	25	34	37	157

$X^2 = 31.0$, 6 vrijheidsgraden, kans op $X^2 \geq 31.0$ is $2.5{\cdot}10^{-5}$. Conclusie: de gevonden uitkomst is uiterst onwaarschijnlijk indien de *Dixa*-soorten niet verschillen in hun verdeling over de waterkwaliteiten.

Inspectie van de kruistabel toont aan dat er een positieve associatie bestaat tussen:

- *Dixa nubilipennis* en oligotroof water
- *Dixa dilatata* en mesotrrof water
- *Dixa nebulosa* en eutroof water

5.5.3 Zaadkieming afhankelijk van temperatuur en bodemtype

	clay soil	sandy soil	total:
5^oC	40/57.97	100/82.03	140
25^0C	131/113.03	142/159.97	273
total:	171	242	413

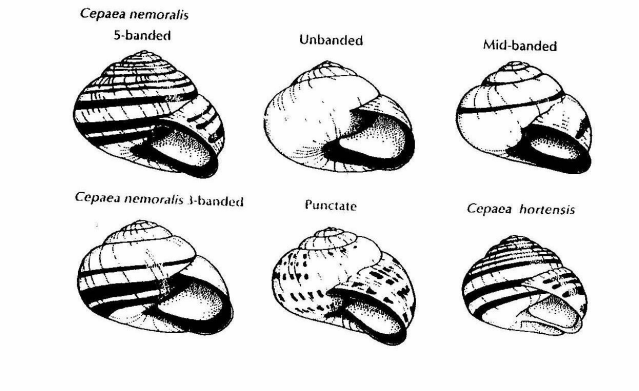
1 vrijheidsgraad, dus

$$X^2 = \frac{(|40 - 57.97| - 0.5)^2}{57.97} + \cdots = 13.6$$

en de kans op $X^2 \geq 13.6$ is $2.3 \cdot 10^{-4}$.

Conclusie: de effecten van temperatuur en bodemtype zijn *niet* onafhankelijk van elkaar ($P < 10^{-3}$ op een uitkomst van X^2 minstens 13.6). Er is een positieve associatie tussen lage temperatuur en zandgrond en tussen hoge temperatuur en kleigrond.

5.5.4 Schelpkleur en bandering bij *Cepaea nemoralis*



De gegevens:

	geel	rose	bruin	totaal
gebandeerd	60	30	20	110
ongebandeerd	30	40	70	140
totaal	90	70	90	250

Vraagstelling: komen alle kleurvarianten even vaak voor?
Zou dit het geval zijn, dan moeten de gevonden frequenties overeenkomen met een uniforme verdeling:

	geel	rose	bruin	totaal
gevonden:	90	70	90	250
verwacht:	83.33	83.33	83.33	250

$$X^2 = 0.53 + 2.13 + 0.53 = 3.19$$

Bij twee vrijheidsgraden is de kans op $X^2 > 3.19$ gelijk aan 0.20. Conclusie: de verdeling van de kleurvarianten wijkt niet significant af van een uniforme verdeling.

Vraagstelling: zijn schelpkleur en bandering onafhankelijk van elkaar? Vergelijk hiertoe de beide gevonden frequenties met elkaar.

	geel	rose	bruin	totaal
gebandeerd	60/39.6	30/30.8	20/39.6	110
ongebandeerd	30/50.4	40/39.2	70/50.4	140
totaal	90	70	90	250

$X^2 = \frac{(60-39.6)^2}{39.6} + \cdots = 36.13$ en twee vrijheidsgraden, dus een zeer zeldzame uitkomst (kans $1.4 \cdot 19^{-8}$ indien er tussen kleur en bandering geen verband bestaat). Conclusie: kleur en bandering zijn met elkaar geassocieerd ($P < 0.05$)

Samenvatting

1. $X^2 = (\text{gevonden} - \text{verwacht})^2/\text{verwacht}$ geeft *afwijking* weer tussen de gevonden frequenties en de te verwachten frequenties.

2. Alleen een *grote* waarde van X^2 kan tot verwerping van de nulhypothese leiden.

3. De χ^2-verdelingen hebben alleen betrekking op aantallen (absolute frequenties) en *niet* op percentages, proporties, fracties of metingen.

4. De verwachte frequenties mogen niet te klein zijn: geen verwachte frequentie kleiner dan 1 en niet meer dan 20% kleiner dan 5. Zo niet, dan kategorieën poolen (bij kruistabellen hele rijen of kolommen).

5.6 Kritieke waarden voor de Chi-kwadraat verdelingen

Voor testen van onafhankelijkheid en goodness-of-fit wordt bij $\alpha = 0.05$ en 0.01 de waarde χ^2_α gegeven met de eigenschap dat $P(X^2 > \alpha) = \alpha$.

df	$\chi^2_{0.05}$	$\chi^2_{0.01}$	df	$\chi^2_{0.05}$	$\chi^2_{0.01}$	df	$\chi^2_{0.05}$	$\chi^2_{0.01}$
1	3.841	6.635	11	19.675	24.725	25	37.652	44.314
2	5.991	9.210	12	21.026	26.217	30	43.773	50.892
3	7.815	11.345	13	22.362	27.688	35	49.802	57.342
4	9.488	13.277	14	23.685	29.141	40	55.758	63.691
5	11.070	15.086	15	24.996	30.578	50	67.505	76.154
6	12.592	16.812	16	26.296	32.000	60	79.082	88.379
7	14.067	18.457	17	27.587	33.409	70	90.531	100.425
8	15.507	20.090	18	28.869	34.805	80	101.879	112.329
9	16.919	21.666	19	30.144	36.191	90	113.145	124.116
10	18.307	23.209	20	31.410	37.566	100	124.324	135.807

Hoofdstuk 6

Twee steekproeven

We hebben twee steekproeven met steekproefgrootten n_1 en n_2 uit twee populaties. De steekproefgemiddelden $\overline{x}_1$ en $\overline{x}_2$ verschillen van elkaar. Is dit verschil significant? Dat wil zeggen: liggen $\overline{x}_1$ en $\overline{x}_2$ voldoende uiteen om te concluderen dat $\mu_1 \neq \mu_2$? Dit is het onderwerp van de paragrafen 6.1 tot en met 6.4 Een soortgelijke vraag is te stellen over de varianties: paragraaf 6.6.

6.1 Een eerste criterium

als de 95% betrouwbaarheidsintervallen voor μ_1 en μ_2 *niet* overlappen, lijkt het redelijk, aan te nemen dat $\mu_1 \neq \mu_2$. Hoewel dit criterium veel gebruikt wordt, is het eigenlijk niet zo efficiënt. We zullen in de volgende paragraaf een beter criterium geven.

Voorbeeld: Effect van eiwit op de groei van ratten. We hebben twee groepen vrouwelijke ratten, één op proteïnerijk dieet en één op proteïnearm dieet. Gemeten: gewichtstoename in grammen van dag 28 tot dag 84:

proteïnerijk	116.1, 101.1, 99.3, 128.7, 102.4, 94.6, 131.6
	130.7, 114.6, 115.5, 105.2, 116, 129.3, 112
proteïnearm	125.8, 85.8, 108.5, 99.8, 113.9, 102.8, 85.5
	93.4, 106.5, 105.6

	proteïnerijk	proteïnearm
steekproefgrootte	14	10
gemiddelde	114.08	102.76
standaarddeviatie	12.48	12.44
s.e. $(= s/\sqrt{n})$	3.34	3.93
t-waarde $(t_{0.95})$	2.16	2.262
$t_{0.95} \cdot$ s.e.	7.2	8.9
betrouwbaarheidsinterval	$[106.9, 121.3]$	$[93.9, 111.7]$

De betrouwbaarheidsintervallen overlappen. Toch is er een reden om aan te nemen dat $\mu_1 \neq \mu_2$! We komen op dit voorbeeld terug in paragraaf 6.3

6.2 Een beter criterium

Een scherper criterium wordt geleverd door de volgende stelling.

Stelling 3. *Stel dat twee* onafhankelijke *steekproeven ter grootte n_1 en n_2 worden getrokken uit twee populaties met gemiddelden μ_1 en μ_2 en varianties σ_1^2 en σ_2^2. Als de steekproeven groot genoeg zijn, dan geldt:*

Het verschil tussen de steekproefgemiddelden, $\Delta := \overline{X}_1 - \overline{X}_2$, is bij benadering normaal verdeeld, met gemiddelde $\mu_\Delta = \mu_1 - \mu_2$ en variantie $\sigma_\Delta^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

Gevolg: de stochast

$$\frac{\Delta - \mu_\Delta}{\sigma_\Delta} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

is bij benadering standaard normaal verdeeld.

Een toepassing. Lengte (in mm) van pissebedden (*Sphaeroma rugicauda*) afkomstig van twee locaties: De rivieren Deben en Colne:

	Deben	Colne
gemiddelde (mm)	$\overline{Y}_D = 4.37$	$\overline{Y}_C = 4.12$
stddeviatie (mm)	$s_D = 0.837$	$s_C = 0.739$
sample size	$n_D = 100$	$n_C = 100$

Nulhypothese: De populatiegemiddelden verschillen niet van elkaar ($\mu_D = \mu_C$) de gevonden verschillen ($\overline{Y}_D \neq \overline{Y}_C$) zijn het gevolg van toeval.
Volgens Stelling 3 is, onder de nulhypothese, de stochast

$$Z = \frac{\overline{Y}_D - \overline{Y}_C}{\sqrt{(\sigma_D^2/n_D) + (\sigma_C^2/n_C)}}$$

standaard normaal verdeeld. We zouden daarom Z voor ons voorbeeld willen uitrekenen.

Probleem: de populatievarianties σ_D^2 en σ_C^2 zijn niet bekend!

Oplossing (Grote steekproefomvang): omdat onze steekproeven groot zijn (vuistregel: $n_i \geq 100$), worden de populatievarianties heel goed benaderd door de steekproefvarianties, dus in ons geval is

$$z \sim \frac{\overline{y}_D - \overline{y}_C}{\sqrt{(s_D^2/n_D) + (s_C^2/n_C)}} = 2.24$$

terwijl de kritieke z-waarde $z_{\alpha/2}$ bij significantieniveau $\alpha = 0.05$ gelijk is aan 1.96. (Zie paragraaf 6.8 voor notatie.)

Conclusie: op basis van de nulhypothese $\mu_D = \mu_C$ is het gevonden resultaat erg onwaarschijnlijk. We verwerpen dus de nulhypothese: de populaties verschillen van elkaar.

Probleem: Wat te doen bij kleine steekproeven? Bij kleine steekproeven ($n_i < 100$) zijn de steekproefvarianties geen betrouwbare schatters voor de populatievarianties. We komen dan uit op een t-verdeling in plaats van een z-verdeling (als $n_1 \sim n_2$ of $\sigma_1 \sim \sigma_2$).

Stelling 4. (Verschil van gemiddelden bij onbekende σ) *Laten twee steekproeven van grootten n_1 en n_2 met gemiddelden $\overline{x}_1$ en $\overline{x}_2$ en steekproefvarianties s_1^2 en s_2^2 uit twee normaal verdeelde populaties met gemiddelden μ_1 en μ_2 en varianties σ_1^2 en σ_2^2 gegeven zijn. Dan is de statistiek*

$$t_\Delta = \frac{\overline{x}_1 - \overline{x}_2}{s_\Delta}$$

een trekking uit de t-verdeling met $n_1 + n_2 - 2$ vrijheidsgraden. Hierbij is, afhankelijk van de situatie, s_Δ als volgt te berekenen: $s_\Delta \geq 0$ en

- *Als $n_1 = n_2 = n$:* $\quad s_\Delta^2 = (s_1^2 + s_2^2)/n$;
- *Als $n_1 \simeq n_2$:* $\quad s_\Delta^2 = s_1^2/n_1 + s_2^2/n_2$;
- *Als $\sigma_1 = \sigma_2$:* $\quad s_\Delta^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \cdot \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$.

6.3 Voorbeelden

Effect van medicijn op de bloedstolling.

	Bloedstoller	Geen medicijn
steekproefomvang	15	15
gemiddelde	8.75	9.63
standaarddeviatie	1.02	1.38

(Bloedstollingstijd in minuten.)
Nulhypothese: het medicijn heeft geen effect: $\mu_B = \mu_G$.
We vinden de volgende t-waarde:

$$t_\Delta = \frac{\overline{x}_B - \overline{x}_G}{\sqrt{(s_B^2 + s_G^2)/n}} = \frac{8.75 - 9.63}{\sqrt{(1.02^2 + 1.38^2)/15}} = -1.986$$

Kritieke t-waarde ($\alpha = 0.05;\ \nu = 2(n-1) = 28$):

$$t_{\alpha/2} = 2.048\ .$$

Omdat $t_\Delta \in [-t_{\alpha/2}, t_{\alpha/2}]$, hebben we geen reden om de nulhypothese te verwerpen. Er is geen effect van het medicijn aangetoond.

Voorbeeld: (Rattengroei opnieuw bekeken) Nulhypothese: geen verschil ($\mu_1 = \mu_2$).
De gevonden waarde voor t_Δ is:

$$\frac{114.08 - 102.76}{\sqrt{\left(\frac{(14-1)\cdot 12.48^2 + (10-1)\cdot 12.44^2}{14+10-2}\right) \cdot \left(\frac{1}{14} + \frac{1}{10}\right)}} = 2.194$$

Kritieke t-waarde bij significantieniveau $\alpha = 0.05$, 14+10-2=22 vrijheidsgraden:

$$t_{\alpha/2} = 2.074\ .$$

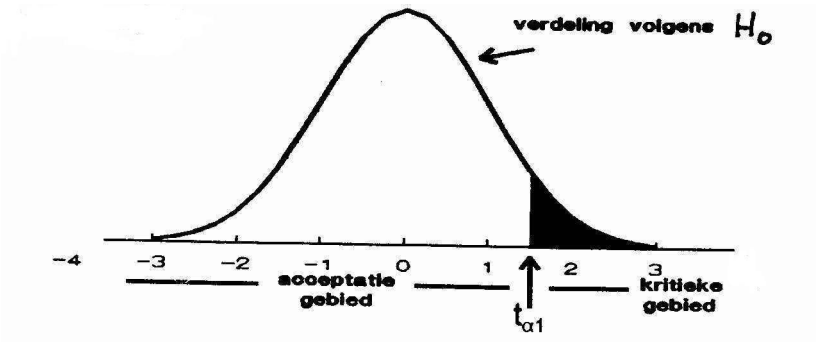
Conclusie: nulhypothese verwerpen: blijkbaar heeft toevoeging van eiwit aan het voer wèl een significant effect ($\alpha = 0.05$) op de groei van ratten (ondanks het feit dat de 95% betrouwbaarheidsintervallen voor de populatiegemiddelden overlappen, zie paragraaf 6.1).

6.4 Eenzijdig toetsen

In het voorbeeld over het effect van een medicijn op de bloedstolling hebben we de nulhypothese niet verworpen.
De situatie verandert als alle partijen het er bij voorbaat over eens zijn, dat een *verhogende* werking van het medicijn op de bloedstollingstijd onmogelijk is. Het verwerpingsgebied van de nulhypothese wordt dan verkleind: we verwerpen alleen als $\overline{x}_B$ voldoende kleiner is dan $\overline{x}_G$ (dus als het medicijn lijkt te werken als bedoeld). We gaan *eenzijdig* toetsen.

Eenzijdig toetsen: Een éénzijdige toets houdt alleen rekening met éénzijdige afwijkingen van de toetsgrootheid (bijvoorveeld de t-waarde). Het "kritieke gebied" ligt daarom maar aan één kant en wordt begrensd door een kritieke waarde; in het geval van een t-toets is dit t_α.
Er moet voor eenzijdig toetsen aan twee voorwaarden zijn voldaan:

1. De keuze moet worden gemaakt voordat de toetsresultaten bekend zijn.
2. De keuze moet goed kunnen worden gemotiveerd, bijvoorbeeld op grond van een biologische theorie of kennis van de literatuur over soortgelijke situaties.



De nulhypothese wordt bij een eenzijdige toets alleen verworpen als de toetsgrootheid in de van tevoren afgesproken richting van 0 afwijkt.
Als de toetsgrootheid in de "juiste" richting van 0 afwijkt, dan wordt H_0 bij een éénzijdige toets sneller verworpen dan bij de analoge tweezijdige toets.

Opnieuw het voorbeeld: Effect van medicijn op bloedstolling. o nulhypothese: er is geen effect ($\mu_B = \mu_G$);
o gevonden t-waarde: -1.986;
o kritieke t-waarde ($\alpha = 0.05,\ \nu = 28$): $t_\alpha = -1.701$.

Conclusie: Nulhypothese wèl verwerpen.

Gevaar: Het is verleidelijk om stelselmatig éénzijdige toetsen te gebruiken, want bij deze wordt de nulhypothese sneller verworpen.
Maar een "significante" conclusie die getrokken wordt op basis van een éénzijdige toets is volstrekt *onbetrouwbaar* als er niet bij voorbaat (*a priori*) duidelijke redenen zijn waarom afwijking maar in één richting zou kunnen optreden.

6.5 Gepaarde Steekproeven

In dit hoofdstuk behandelen we toetsen ter vergelijking van twee steekproeven die aan elkaar gekoppeld zijn. Er is dan sprake van *paren van waarnemingen*.

De t-toets voor gepaarde waarnemingen

- n paren van waarnemingen $\Rightarrow$ n verschillen d_i
- gemiddeld verschil: $\overline{d}$
- standaarddeviatie van de verschillen: s_d
- Nulhypothese (H_0):
  Het gevonden verschil $\overline{d}$ berust alleen op toeval: in werkelijkheid is er geen systematisch verschil ($\mu_d = 0$)

Stelling 5. (t-toets voor gepaarde waarnemingen) *Als de nulhypothese $\mu_d = 0$ klopt, en de waarnemeingen komen uit een normale verdeling, dan is de steekproefgrootheid*

$$t = \frac{\overline{d}}{s_d/\sqrt{n}}$$

een trekking uit een t-verdeling met $\nu = n - 1$ vrijheidsgraden.

Conclusie: We verwerpen de nulhypothese als de berekende t-waarde groter is dan de kritieke t-waarde $t_{\alpha/2}$.

Voorbeeld:

persoon	controle	medicijn	verschil (d_i)
1	112	125	+13
2	115	118	+3
3	115	109	-6
4	120	114	-6
5	121	135	+14
6	125	155	+30
7	129	143	+14
8	130	139	+9
9	131	145	+14
10	148	152	+4
gemiddelde:	124.6	133.5	8.9
stddeviatie:	10.6	16.2	10.8

- berekende t-waarde:
$$t = \frac{8.9}{10.8/\sqrt{10}} = 2.61$$
- kritieke t-waarde bij $\alpha = 0.05$ en 9 vrijheidsgraden:
$$t_{\alpha/2} = 2.262$$

Dus het geneesmiddel leidt wel tot een significante ($P < 0.05$) verhoging van de bloeddruk!

6.6 De F-toets voor gelijkheid van varianties

Hier is de vraagstelling: wordt het verschil tussen twee steekproefvarianties veroorzaakt door toevalseffecten of door het feit dat de varianties van de onderliggende populaties verschillen?
Nulhypothese (H_0): het gaat alleen om toevalseffecten; de varianties van de onderliggende populaties verschillen niet van elkaar: $\sigma_1^2 = \sigma_2^2$.

Toetsingsprocedure:

- Kies een significantieniveau α. Voor het gebruik van Tabel 6.9 is het niveau $\alpha = 0.1$ vereist.

- Bereken het quotiënt van de twee steekproefvarianties (F-waarde)

$$F = s_1^2/s_2^2, \text{ waarbij } s_1^2 > s_2^2 \, .$$

  Als de nulhypothese klopt zal dit quotiënt een trekking zijn uit een F-verdeling met $\nu_1 = n_1 - 1$ en $\nu_2 = n_2 - 1$ vrijheidsgraden.

- Haal uit Tabel 6.9 de "kritieke waarde" $f_{\alpha/2}$. Omdat we kunstmatig de grootste steekproefvariantie in de teller hebben gezet, heeft de berekende toetsgrootheid onder H_0 een kans α om groter te zijn dan $f_{\alpha/2}$.

- Verwerp H_0 als de gevonden F-waarde groter is dan de kritieke waarde $f_{\alpha/2}$, want zo'n grote F-waarde is op basis van de nulhypothese alleen met kleine kans ($P < \alpha$) te verwachten.

Merk op dat we met bovenstaande procedure *tweezijdig* toetsen!

Voorbeeld: Eiproductie in twee stekelbaarspopulaties

	brak water	zoet water
sample size	12	17
gem. eiproductie	127.2	103.4
standaarddeviatie	29.1	20.2

Nulhypothese: varianties zijn gelijk: $\sigma_b^2 = \sigma_z^2$.
Alternatieve hypothese: varianties zijn ongelijk: $\sigma_b^2 \neq \sigma_z^2$.
Gevonden F-waarde: $F = s_b^2/s_z^2 = (29.1)^2/(20.2)^2 = 2.075$
kritieke F-waarde ($\alpha = 0.05,\ \nu_b = 11,\ \nu_z = 16$): $F_\alpha = 2.93$

Conclusie: de gevonden F-waarde is op basis van de nulhypothese niet onwaarschijnlijk genoeg om de nulhypothese te verwerpen. We gaan er dus voorlopig van uit dat de varianties van beide populaties gelijk zijn.

- De t-toets ('Is $\mu_1 = \mu_2$?') en de F-toets ('Is $\sigma_1^2 = \sigma_2^2$?') toetsen of de *parameters* van twee populaties aan elkaar gelijk zijn.
- Beide toetsen zijn in eerste instantie alleen van toepassing als de onderliggende populaties normaal verdeeld zijn.
- De t-toets is echter robuust en mag ook worden toegepast op
  - *symmetrische* verdelingen, indien $n_1, n_2 > 10$
  - *asymmetrische* verdelingen, indien $n_1, n_2 > 30$

6.7 De F-verdeling

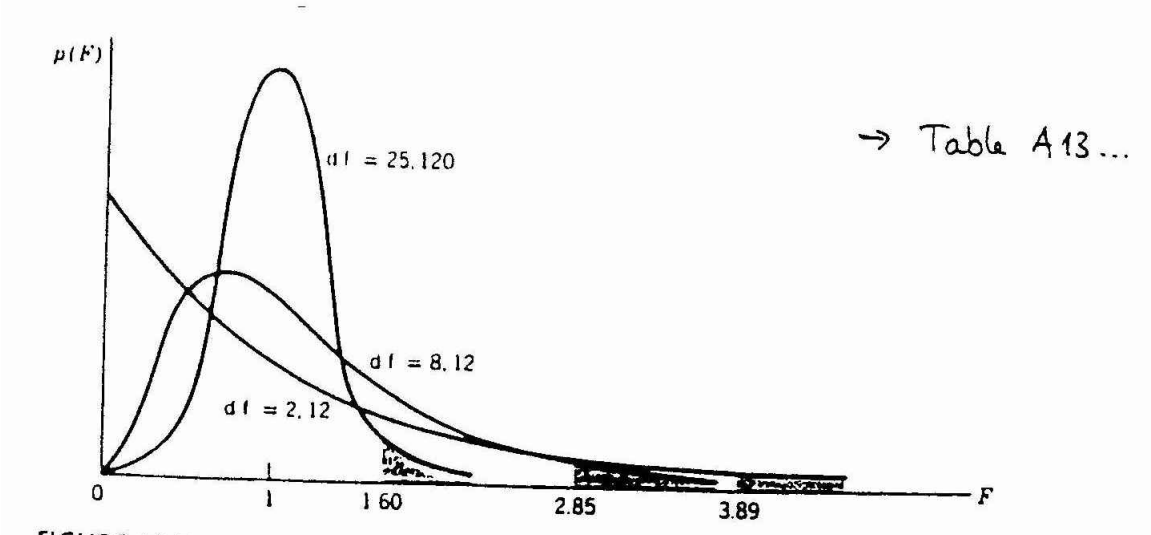


FIGURE 10-3
Some typical F distributions, with various d.f. in the numerator and denominator. Note how the 5% critical point (beyond which H_0 is customarily rejected) moves left toward 1 as d.f. increase.

- Stel dat twee populaties normaal verdeeld zijn met dezelfde variantie: $\sigma_1^2 = \sigma_2^2$.
- Als we van elke populatie een steekproef nemen met steekproefomvang n_1 resp. n_2, dan zal het *quotiënt* s_1^2/s_2^2 van de steekproefvarianties ongeveer gelijk zijn aan 1.
- Op grond van toevalseffecten zal het quotiënt echter nooit *precies* gelijk zijn aan 1, maar rond 1 fluctueren.

- Deze fluctuatie wordt beschreven door de *F-verdeling* met $\nu_1 = n_1 - 1$ en $\nu_2 = n_2 - 1$ vrijheidsgraden. Hierbij is n_1 de omvang van de steekproef met de grootste variantie.

Berekening van de kans op een bepaalde of extremere uitkomst met de `TI-84 Plus`: De toetsen `2nd DISTR 0` geeft `Fcdf(`, de cumulatieve verdelingsfunctie voor de F-verdeling. Intypen van a, b, ν_1 en ν_2 geeft de kans dat $a \leq F \leq b$ voor een F-verdeelde stochast met ν_1 en ν_2 vrijheidsgraden.

Voor ons eiproductievoorbeeld: bij 11 en 16 vrijheidsgraden is de kans op $F \geq 2.075$ gelijk aan Fcdf(2.075,E99,11,16)=0.09. Dit is dus niet erg onwaarschijnlijk.

6.8 Kritische waarden: notatie.

We geven het *significantieniveau* van een toets meestal met α aan, en het *betrouwbaarheidsniveau* met γ. Hierbij is α een klein getal, en γ ligt juist dicht bij 1=100%.

Voeren we een toets uit met behulp van een standaard-normale statistiek Z, dan hebben we de keus tussen

- éénzijdig toetsen: verwerp H_0 als $Z \geq z_\alpha$, waarbij z_α zó gekozen is, dat $\mathbb{P}[Z \geq z_\alpha] = \alpha$.
- tweezijdig toetsen. Dit is het meest gebruikelijk, en ook het meest aan te raden. In dit geval verwerpen we H_0 als $|Z| \geq z_{\alpha/2}$, waarbij $z_{\alpha/2}$ natuurlijk zó gedefineerd is dat $\mathbb{P}[Z \geq \alpha/2] = \alpha/2$ (en dus $\mathbb{P}\big[|Z| \geq \alpha/2\big] = \alpha$!).

Bijvoorbeeld:

- $\mathbb{P}[Z \leq 1.96] = \Phi(1.96) = 0.975$; (Zie Tabel 4.5)
- dus $\mathbb{P}[Z \geq 1.96] = 1 - 0.975 = 0.025$, en dus $z_{0.025} = 1.96$.
- Symmetrie: $\mathbb{P}[Z \leq -1.96] = \mathbb{P}[Z \geq 1.96] = 0.025$;
- dus $\mathbb{P}\big[|Z| \geq 1.96\big] = \mathbb{P}[Z \geq 1.96] + \mathbb{P}[Z \leq -1.96] = 2 \times 0.025 = 0.05$;
- en $\mathbb{P}[|Z| \leq 1.96] = 1 - 0.05 = 0.95$.
- Dat wil zeggen: $z_{95\%} = 1.96$ $(= z_{0.025}!)$.

Bij de t-verdeling gaat het net zo: de waarde t_α is zó gekozen dat voor een t-verdeelde stochast T geldt:

$$\mathbb{P}[T \geq t_\alpha] = \alpha \ .$$

Dus, als T verdeeld is volgend de t-verdeling met bijvoorbeeld $\nu = 7$ vrijheidsgraden, dan is steeds $\mathbb{P}[T \geq t_\alpha] = \alpha$. Bijvoorbeeld

$$t_{0.025} = 2.365 \ , \quad \text{want} \quad \mathbb{P}[T \geq 2.365] = 0.025 \ .$$

En ook hier schrijven we soms (nog steeds voor $\nu = 7$),

$$t_{95\%} = 2.365 \; .$$

Dus bij een tweezijdige t-toets met 7 vrijheidsgraden en significantieniveau 5% verwerpen we de nulhypothese als de statistiek $|T|$ de waarde 2.365 overschrijdt.
Eenzelfde notatie wordt gebruikt bij de F-verdeling: als de stochast F Fisher-verdeeld is met $\nu_1 = 5$ vrijheidsgraden in de noemer en $\nu_2 = 10$ vrijheidsgraden in de teller, dan is

$$f_{5\%} = 3.33, \quad \text{want} \quad \mathbb{P}[F \geq 3.33] = 5\% \; .$$

Maar omdat in de stochast S_1^2/S_2^2 steeds per definitie $S_1 \geq S_2$ is gekozen, had met dezelfde kans een omgekeerde overschrijding op kunnen treden. Onze significantie bij deze tweezijdige toets is daarom slechts 10%.

6.9 Kritieke waarden van F-verdelingen

In de eerste kolom staat het aantal vrijheidsgraden van de noemer, in de eerste rij het aantal vrijheidsgraden van de teller. In de tabel staat de waarde f zó, dat $\mathbb{P}[F \geq f] = 0.05$.

	1	2	3	4	5	6	7	8	9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	9.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.12	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21

Hoofdstuk 7

Regressie en correlatie

In dit hoofdstuk behandelen we twee verschillende situaties, die veel met elkaar gemeen hebben.

- We hebben een steekproef genomen uit een populatie, en aan ieder individu zijn *twee* grootheden x_i en y_i gemeten.
- We hebben bij verschillende instellingen, plaatsen of tijden x_i één grootheid y_i gemeten.

In elk van deze situaties hebben we een rijtje getallenparen

$$(x_1, y_2), (x_2, y_2), \ldots, (x_n, y_n) \ .$$

En in elk ervan interesseert ons de vraag of er tussen x_i en y_i een verband bestaat. Maar in het eerste geval interesseert ons de *correlatie* of samenhang, die niet per se causaal hoeft te zijn. In het tweede geval interesseert ons meer de vraag *of* y_i *van* x_i *afhangt*, bijvoorbeeld of y stijgt als functie van x. Dit is een causaal verband. Natuurlijk hebben deze twee vragen veel met elkaar te maken. Maar omdat de voorstelling erbij anders is, behandelen we ze in twee verschillende paragrafen: 7.1 en 7.2.

7.1 Correlatie tussen twee metrische variabelen

We berekenen de variantie in x-waarden en die in de y-waarden:

$$s_x^2 = \frac{1}{n-1} \sum_i (x_i - \overline{x})^2,$$

$$s_y^2 = \frac{1}{n-1} \sum_i (y_i - \overline{y})^2.$$

De *covariantie* tussen x en y wordt nu gegeven door:

$$s_{xy} = \frac{1}{n-1}\sum_i (x_i - \overline{x})(y_i - \overline{y}) \ .$$

Je kunt deze overigens gemakkelijker berekenen met de formule:

$$s_{xy} = \frac{1}{n-1}\left(\sum_i x_i y_i - n\overline{x}\cdot\overline{y}\right) .$$

Op basis van deze grootheden wordt de **correlatiecoëfficiënt** gedefinieerd:

$$r := \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y} \ .$$

Voorbeeld: Kleur en reproductief succes van stekelbaarsjes. (x_i: kleurindex; y_i: legselgewicht)

visnr	x_i	y_i	$x_i y_i$
1	0.15	5.0	0.75
2	0.23	12.3	2.83
3	0.28	11.9	3.33
4	0.45	10.1	4.55
5	0.47	17.2	8.08
6	0.52	25.1	13.05
7	0.59	19.5	11.51
8	0.62	20.1	12.46
9	0.81	29.0	23.49
10	0.90	17.2	15.48
som	5.02	167.4	95.53
gem	0.502	16.74	
S^2	0.058	51.62	

$$s_x^2 = 0.058,\ s_y^2 = 51.62,\ s_{xy} = \frac{1}{9}(95.53 - 10\cdot 0.502\cdot 16.74) = 1.277$$

$$r = \frac{1.277}{\sqrt{0.058\cdot 51.62}} = 0.735$$

Correlatietoets: De nulhypothese dat er in de populatie geen verband bestaat tussen de beide grootheden X en Y wordt getoetst door na te gaan of de gevonden waarde van r al dan niet groter is dan de kritieke waarde r_α in Tabel 7.4.
Voor een tweezijdige toets met $\nu = 10-2 = 8$ vrijheidsgraden lezen we in deze tabel af:

$$\alpha = 0.05 \Rightarrow r_{\alpha/2} = 0.632;$$

$$\alpha = 0.01 \Rightarrow r_{\alpha/2} = 0.765$$

Dus: er is een significante correlatie geconstateerd tussen kleur en broedsucces. De p-waarde ligt tussen 0.01 en 0.05.

Eigenschappen van de correlatiecoëfficiënt r

- Het teken $(-, +)$ geeft de richting van het verband: dalend of stijgend.
- De grootte staat voor de sterkte van het verband.
- r neemt waarden aan tussen -1 en $+1$.

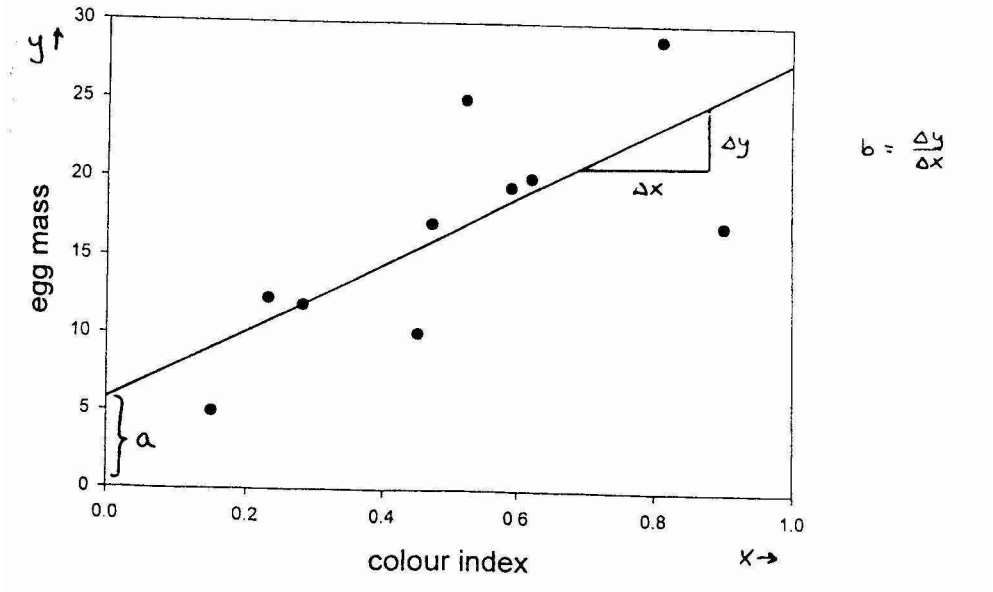
$r = 0$:	er bestaat *geen* verband tussen x en y;
$r > 0$:	er bestaat een *positieve* relatie tussen x en y;
$r < 0$:	er bestaat een *negatieve* relatie tussen x en y;
$r = +1$:	er bestaat een *lineair positief* verband;
$r = -1$:	er bestaat een *lineair negatief* verband.

7.2 Regressie-Analyse

Lineaire regressie probeert het verband tussen twee variabelen zo goed mogelijk door middel van een *rechte lijn* (lineaire functie) weer te geven:

$$\hat{y} = a + bx$$

De "regressiecoëfficiënt" b kenmerkt de helling van de lijn (is de richtingscoëfficiënt). De constante a kenmerkt het snijpunt met de y-as.



Welke rechte lijn past het beste bij de gegevens?

De "kleinste kwadraten methode":

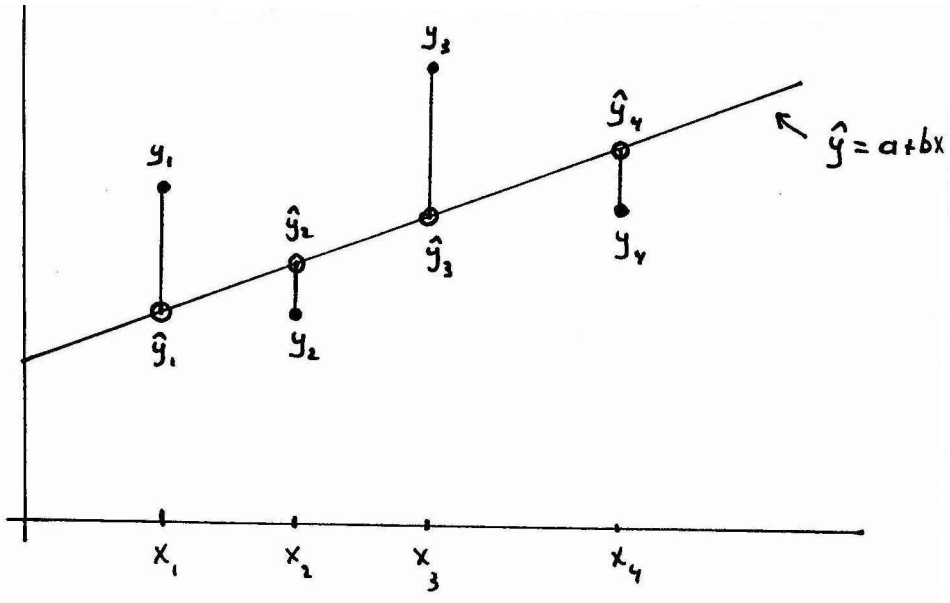
- de rechte $y = a + bx$ wordt bepaald door de twee coëfficiënten a en b
- gegeven a en b, vergelijk de waargenomen punten (x_i, y_i) met de bijbehorende punten op de lijn: $(x_i, \hat{y}_i) = (x_i, a + bx_i)$
- de verschillen in de verticale richting (y) zijn een maat voor de afwijking tussen de rechte lijn en de datapunten

$$y_i - \hat{y}_i = y_i - (a + bx_i)$$

- de 'best passende' lijn krijgen we door a en b zo te kiezen dat de som van de kwadratische afwijkingen geminimaliseerd wordt:

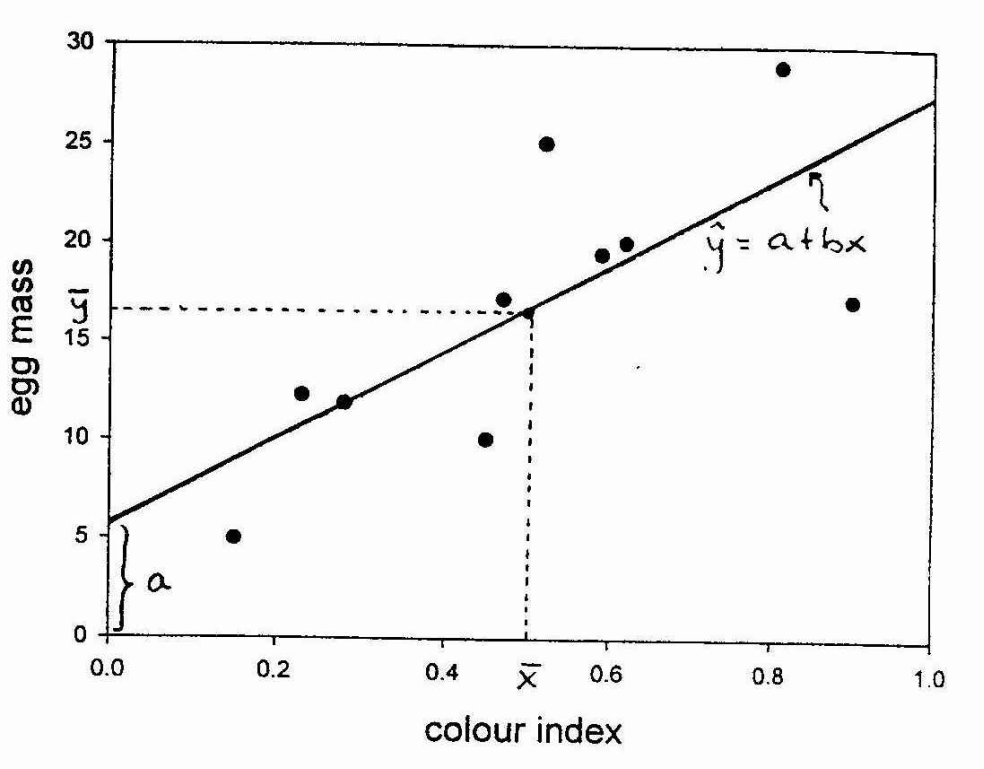
Resultaat:

$$b = \frac{s_{xy}}{s_x^2} \quad \text{en} \quad a = \overline{y} - b\overline{x}$$



Voorbeeld: Kleur en reproductief succes van Stekelbaarsjes (x_i: kleurindex, y_i: legselgewicht)

visnr	x_i	y_i	$x_i y_i$
1	0.15	5.0	0.75
2	0.28	11.9	2.83
3	0.47	17.2	3.33
4	0.59	19.5	4.55
5	0.81	29.0	8.08
6	0.23	12.3	13.05
7	0.45	10.1	11.51
8	0.52	25.1	12.46
9	0.62	20.1	23.49
10	0.90	17.2	15.48
Som	5.02	167.4	95.53
Gem	0.502	16.74	
S^2	0.058	51.62	



$$\overline{x} = 0.502,\ \overline{y} = 16.740,\ s_x^2 = 0.058,\ s_y^2 = 51.620,\ s_{xy} = 1.277$$

Dus

$$b = \frac{s_{xy}}{s_x^2} = 21.845$$

$$a = \overline{y} - b\overline{x} = 5.774$$

LET OP: grote afrondingseffecten bij regressie en correlatie

Vraag: Kunnen we bij een stekelbaarsje met een bepaalde kleurindex voorspellen wat zijn legselgewicht zal zijn? B.v. wat is de beste schatting voor het legselgewicht van een stekelbaarsje met een kleurindex van 0.70?

$$\text{Antwoord: } \hat{y} = a + bx = 5.774 + 21.845 \cdot 0.7 = 21.07$$

Nut van de regressielijn: Op basis van x kun je de bijbehorende y-waarde voorspellen.

Eigenschappen:

- De regressielijn gaat altijd door het punt $(\overline{x}, \overline{y})$

- $b > 0$: er bestaat een positief verband tussen x en y;
- $b < 0$: er bestaat een negatief verband tussen x en y;
  $b = 0$: er bestaat geen verband tussen x en y.

Betrouwbaarheid van een regressie-analyse:

- Stel dat het verband tussen x en y in de *hele* populatie wordt weergegeven door $\hat{y} = \alpha + \beta x$.

- Door middel van een regressie-analyse probeer je dit verband te achterhalen, maar de regressiecoëffeiciënten

  $$b = \frac{s_{xy}}{s_x^2} \text{ en } a = \overline{y} - b\overline{x}$$

  worden bepaald op basis van een steekproef.

- Andere steekproeven uit *dezelfde* populatie zouden tot iets andere resultaten leiden: a en b zijn alleen maar schatters voor de "echte" parameters α en β.

- De betrouwbaarheid van die schattingen kan gekwantificeerd worden: de standaardfout die we bij de schatting van β (d.i. b) maken wordt gegeven door

  $$s.e._\beta = \frac{s_e}{s_x\sqrt{n-1}}$$

  Hierbij is s_e de wortel uit de "residual variance"

  $$s_e^2 = \frac{1}{n-2}\sum(y_i - \widehat{y}_i)^2 = \frac{n-1}{n-2}(s_y^2 - b^2 s_x^2),$$

  Dit is een maat voor de afwijking tussen datapunten en regressielijn

7.2.1 Toetsen of β significant afwijkt van nul

Methode 1: betrouwbaarheidsinterval . De nulhypothese $H_0 : \beta = 0$ kan worden verworpen als 0 niet binnen het $(1 - \alpha)$-betrouwbaarheidsinterval van β valt.

$$\beta = b \pm t_{\alpha;\nu} \cdot s.e._\beta$$

Hier is $t_{\alpha;\nu}$ de kritieke waarde bij α van een t-verdeling met $\nu = n-2$ vrijheidsgraden.

Methode 2: toetsgrootheid .
Als de nulhypothese $H_0 : \beta = 0$ waar is, dan volgt

$$t = \frac{b}{s.e._\beta}$$

een t-verdeling met $\nu = n - 2$ vrijheidsgraden. We kunnen de nulhypothese dus verwerpen als e berekende toetsgrootheid groter is dan de kritieke waarde $t_{\alpha;\nu}$.

7.2.2 Voorbeeld: Kleur en broedsucces bij stekelbaarsjes

$$s_x^2 = 0.058,\ s_y^2 = 51.620,\ s_{xy} = 1.277$$

$$\Rightarrow\ b = 21.845,\ a = 5.774$$

- residual variance $s_e^2 = \frac{9}{8}(51.620 - 21.845^2 \cdot 0.058) = 26.686$
- standaardfout van de schatting van β:

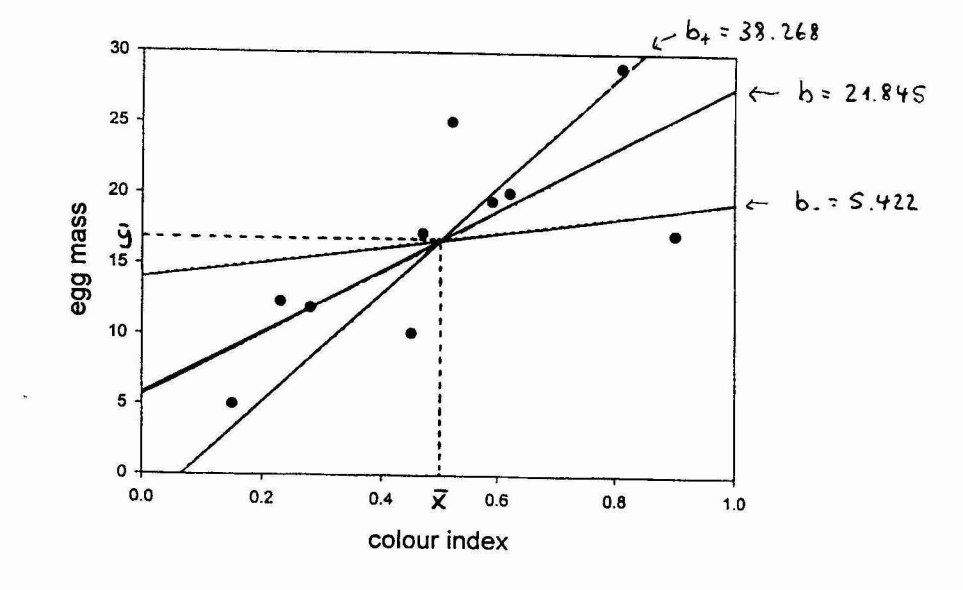
$$s.e._\beta = \frac{s_e}{s_x\sqrt{n-1}} = \frac{5.166}{0.24\sqrt{9}} = 7.122$$

Betrouwbaarheidsinterval voor βToetsen of β

95% betrouwbaarheidsinterval (we gebruiken dat $t_{0.05;8} = 2.306$):

$$\beta = 21.845 \pm 2.306 \cdot 7.122 = 21.845 \pm 16.423$$

95% betrouwbaarheidsinterval = (5.422;38.268).
Conclusie: met 95% zekerheid kunnen we er dus van uitgaan dat de echte regressiecoëfficiënt tussen 5.422 en 38.268 in ligt.



Toetsen of β significant afwijkt van nul.

Hier is $t = \frac{21.845}{7.138} = 3.067$ en $t_{0.05} = 2.306$. Conclusie: H_0 verwerpen.

7.3 Regressie en Correlatie

- Beide geven een dalende / stijgende richting
- Bij de correlatie zijn x en y verwisselbaar ($r = s_{xy}/s_x s_y$)
- Bij regressie zijn x en y *niet* verwisselbaar ($b = s_{xy}/s_x^2$). Nu is x de *onafhankelijke* variabele en y is de *afhankelijke* variabele.
- Vanwege onderstaande relatie zijn beide gerelateerd aan het begrip *verklaarde variantie*:

$$b^2 s_x^2 = r^2 s_y^2$$

$$s_e^2 \sim s_y^2 - b^2 s_x^2 = (1 - r^2) s_y^2$$

Een fractie $1-r^2$ van de totale variantie in de y-waarden bestaat uit fluctuaties rond de regressielijn, de rest (een fractie r^2 dus) wordt verklaard door het lineaire verband met x. Daarom heet r^2 de *coefficient of determination.*

7.3.1 Afhankelijke versus onafhankelijke variabele:

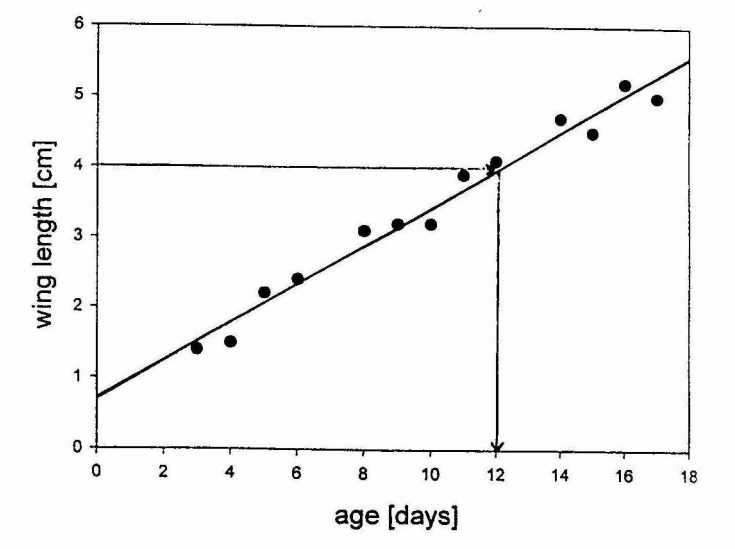
- hoofddoel van een regressie-analyse is het voorspellen van y op basis van x: $\widehat{y} = a + bx$

- regressie veronderstelt daarom impliciet dat y *causaal* afhankelijk is van x (en niet andersom)

- y wordt daarom de "afhankelijke" variabele genoemd en x de "onafhankelijke" variabele

Voorbeeld: Leeftijd en vleugellengte bij mussen. Gegevens:

x_i	3	4	5	6	8	9	10	11	12	14	15	16	17
y_i	1.4	1.5	2.2	2.4	3.1	3.2	3.2	3.9	4.1	4.7	4.5	5.2	5.0



$b = 0.270,\ a = 0.713,\ r = 0.987$

$r = 0.987$ dus $r^2 = 0.973$, en 97.3% van de variatie in de vleugellengte van de mussen wordt verklaard door hun verschillen in leeftijd.

Gevolg

- Dit geeft ons ook een snelle manier om een indruk te krijgen van de grootte van de correlatie: vergelijk de puntenwolk met de bijpassende cirkel.

- Als r^2 groot is liggen de punten zo goed als op één rechte lijn en is dus "inverse prediction" (omgekeerde voorspelling) enigszins verantwoord (zie het vorige voorbeeld: de vleugellengte van de mussen geeft een goed indicatie van hun leeftijd).

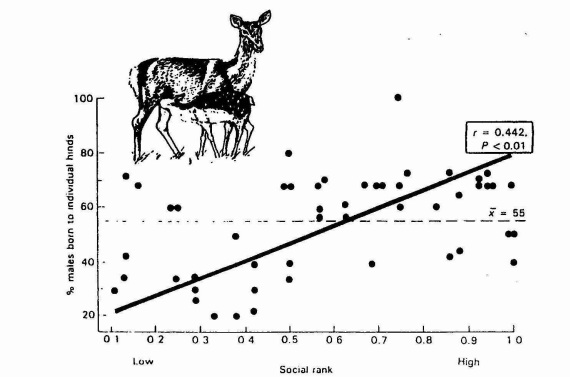


Fig. 8 6 Birth sex ratios produced by individual red deer hinds differing in social rank over their lifespans (from Clutton-Brock *et al.*, 1986).

7.3.2 P-waarde en biologische relevantie

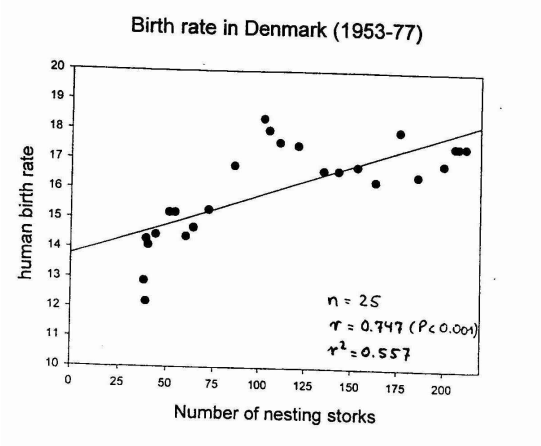
Voorbeeld Hier is een significante correlatie ($P < 0.01$) maar is die ook biologisch relevant ($r^2 < 0.20$)?

Algemeen: uit een kleine r^2-waarde (b.v. $r^2 < 0.5$) kun je concluderen dat in de gegeven situatie *lineaire regressie niet erg zinvol* is:

- òf er wordt maar weinig variatie in y door x verklaard, omdat y nauwelijks van x afhangt,

- òf er is sprake van een *niet-lineair* verband tussen de variabelen

Voorbeeld: Ooievaars en geboortecijfer



Zowel het aantal nesten van ooievaars als het aantal geboortes neemt in de tijd toe. Dus zijn deze grootheden wel gecorreleerd, maar is er toch geen oorzakelijk verband.

7.4 Kritieke waarden voor r

Correlatiecoëfficiënt, eenzijdige toetsing. Aantal vrijheidsgraden ν = n-2.

α_1	5%	2.5%	α_1	5%	2.5%	α_1	5%	2.5%
ν=1	0.988	0.997	11	0.476	0.553	21	0.352	0.413
2	0.900	0.950	12	0.457	0.532	22	0.344	0.404
3	0.805	0.878	13	0.441	0.514	23	0.337	0.396
4	0.729	0.811	14	0.426	0.497	24	0.330	0.388
5	0.669	0.755	15	0.412	0.482	25	0.323	0.381
6	0.621	0.707	16	0.400	0.468	26	0.317	0.374
7	0.582	0.666	17	0.389	0.456	27	0.311	0.367
8	0.549	0.632	18	0.378	0.444	28	0.306	0.361
9	0.521	0.602	19	0.369	0.433	29	0.301	0.355
10	0.497	0.576	20	0.360	0.423	30	0.296	0.349