

# Syllabus Statistiek voor Biologen

Hans Maassen

januari 2014



# Contents

<b>1</b>	<b>Inleiding</b>	<b>5</b>
1.1	Meetonnauwkeurigheid . . . . .	5
1.2	Variabiliteit in fenotypen en genotypen . . . . .	6
1.3	Variatie in ruimte en tijd . . . . .	6
1.4	Waarom statistiek voor beschrijving? . . . . .	6
1.5	Het aantonen van causaal verband . . . . .	6
<b>2</b>	<b>Theorievorming en toetsing</b>	<b>9</b>
2.1	Vraagstelling . . . . .	9
2.2	Experiment . . . . .	10
2.2.1	Mogelijkheden . . . . .	10
2.2.2	Opzet van de proef . . . . .	11
2.3	Statistische analyse . . . . .	11
2.3.1	De significantie-drempel . . . . .	12
2.3.2	De nulhypothese . . . . .	12
2.3.3	De kansberekening . . . . .	13
2.3.4	De $P$ -waarde . . . . .	15
2.3.5	Eenzijdige toetsing . . . . .	16
2.3.6	Tweezijdige toetsing . . . . .	16
2.4	De binomiale verdeling . . . . .	17
2.4.1	Ongelijke kansen . . . . .	17
2.5	De binomiale toets: algemeen . . . . .	19
2.5.1	Eenzijdige binomiale toets . . . . .	19
2.5.2	Tweezijdige binomiale toets . . . . .	20
2.6	Berekening met de grafische rekenmachine . . . . .	21
2.6.1	<code>binompdf</code> . . . . .	21
2.6.2	<code>binomcdf</code> . . . . .	21
<b>3</b>	<b>Beschrijvende Statistiek</b>	<b>23</b>
3.1	Typen waarnemingen: . . . . .	23
3.2	Het weergeven van de gegevens . . . . .	24
3.2.1	De afrondingsregel van Ehrenberg. . . . .	24

3.3	Continue verdelingen . . . . .	28
3.4	Steekproeven uit continue verdelingen . . . . .	28
3.4.1	Notatie en lettergebruik . . . . .	29
3.4.2	Gemiddelde en spreiding met de GR . . . . .	29
3.4.3	Toepassing op de vliegende herten . . . . .	29
3.4.4	Schatting: . . . . .	30
3.4.5	Afrondingsregels voor afgeleide variabelen . . . . .	30
<b>4</b>	<b>Schattingen en de centrale limietstelling</b>	<b>31</b>
4.1	Schattingen . . . . .	31
4.1.1	Het schatten van een kans . . . . .	31
4.1.2	Bewijs van de $\sqrt{n}$ -wet* . . . . .	33
4.1.3	Het schatten van een gemiddelde . . . . .	34
4.2	De centrale limietstelling . . . . .	34
4.2.1	De normale verdeling . . . . .	35
4.2.2	Demonstratie van de centrale limietstelling . . . . .	36
4.2.3	Hoe groot moet $n$ zijn? . . . . .	38
4.3	De $z$ -toets . . . . .	39
4.3.1	De $z$ -toets voor een kans . . . . .	39
4.3.2	Betrouwbaarheidsinterval voor een kans . . . . .	40
4.3.3	Theoretisch betrouwbaarheidsinterval voor een gemiddelde . . . . .	41
4.3.4	Definitie van $z_\alpha$ en $z_\gamma$ : . . . . .	42
4.3.5	Een voorbeeld: Isomerase-activiteit . . . . .	43
4.4	De $t$ -verdeling . . . . .	43
4.4.1	De $t$ -verdeling . . . . .	44
4.4.2	De $t$ -toets voor een gemiddelde . . . . .	45
4.4.3	Praktisch betrouwbaarheidsinterval voor een gemiddelde . . . . .	45
4.4.4	Eigenschappen van de $t$ -verdeling . . . . .	46
4.4.5	Voorbeelden . . . . .	46
4.5	Tabel van de standaard-normale verdeling . . . . .	47
4.5.1	Handleiding . . . . .	47
4.5.2	Normale verdeling met de GR . . . . .	48
4.6	Kritieke waarden voor Student $t$ -verdelingen . . . . .	49
4.6.1	Handleiding . . . . .	49
4.6.2	Berekening met de GR . . . . .	50
<b>5</b>	<b>Discrete kansverdelingen</b>	<b>51</b>
5.1	De discrete uniforme verdeling. . . . .	51
5.2	De binomiale verdeling . . . . .	52
5.2.1	Eigenschappen van de binomiale verdeling: . . . . .	53
5.3	De Poisson-verdeling . . . . .	54
5.3.1	Eigenschappen van de Poissonverdeling. . . . .	56

5.3.2	De dispersie-index. . . . .	56
5.4	‘Goodness-of-fit’: het toetsen van discrete verdelingen . . . . .	57
5.4.1	De $\chi^2$ -verdeling . . . . .	57
5.4.2	De uniforme verdeling: voorkeur van kuikens. . . . .	58
5.4.3	De $\chi^2$ -toets van een discrete verdeling . . . . .	60
5.4.4	Bepaling van het aantal vrijheidsgraden . . . . .	60
5.4.5	Het toetsen van de binomiale verdeling: Geslacht van kinderen. . . . .	61
5.4.6	Het toetsen van de Poissonverdeling: nematoden . . . . .	62
5.5	De $\chi^2$ -toets voor onafhankelijkheid . . . . .	64
5.5.1	De $\chi^2$ -toets voor onafhankelijkheid: . . . . .	65
5.5.2	Nog enkele voorbeelden . . . . .	66
5.5.3	Samenvatting . . . . .	68
5.6	Kritieke waarden voor de $\chi^2$ -verdelingen . . . . .	69
<b>6</b>	<b>Twee steekproeven</b> . . . . .	<b>71</b>
6.1	Overlap van betrouwbaarheidsintervallen . . . . .	71
6.2	Ge-poolde varianties . . . . .	72
6.2.1	Grote steekproefomvang: $z$ -toets met gepoolde varianties: . . . . .	73
6.2.2	Kleine steekproefomvang: $t$ -toets met gepoolde varianties . . . . .	73
6.2.3	Voorbeelden . . . . .	74
6.3	Eenzijdig toetsen . . . . .	74
6.4	$t$ -toets voor gepaarde waarnemingen . . . . .	76
6.5	De F-toets voor gelijkheid van varianties . . . . .	77
6.6	Meer dan twee steekproeven: ANOVA! . . . . .	78
6.7	De F-verdeling . . . . .	79
6.8	Kritische waarden: notatie. . . . .	80
6.9	Kritieke waarden van F-verdelingen . . . . .	82
<b>7</b>	<b>Regressie en correlatie</b> . . . . .	<b>83</b>
7.1	Correlatie tussen twee variabelen . . . . .	83
7.2	Regressie-Analyse . . . . .	85
7.2.1	Toetsen of $\beta$ significant afwijkt van nul . . . . .	89
7.2.2	Voorbeeld: Kleur en broedsucces bij stekelbaarsjes . . . . .	89
7.3	Regressie en Correlatie . . . . .	90
7.3.1	Afhankelijke versus onafhankelijke variabele: . . . . .	90
7.3.2	P-waarde en biologische relevantie . . . . .	92
7.4	Kritieke waarden voor $r$ . . . . .	94



# Chapter 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 grootte 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 overall 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 maken: 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.



# Chapter 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 zijn 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 pissebedden: 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.

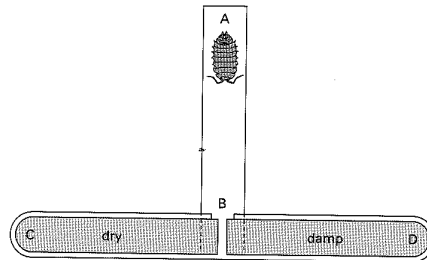


Figure 2.1 A simple choice chamber.

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 zijn 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 toch 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 onwaarschijnlijk moet de uitslag worden voordat onze tegenstander zich gewonnen geeft? Preciezer gezegd: We bepalen de kans  $P$  op *datgene wat we zien gebeuren*, (of iets nog vreemders) onder de aanname dat *de beestjes geen enkele voorkeur hebben*. Met de kritische tegenstander sluiten we het volgende akkoord: hij zal ons gelijk geven als die kans  $P$  onder een bepaalde van tevoren afgesproken drempel  $\alpha$  ligt. Heel gebruikelijk is de keuze:  $\alpha = 5\%$ .

De precieze keuze van  $\alpha$  wordt gemaakt na praktische overwegingen. Als de tegenstander bijvoorbeeld een kans van 1 op miljoen zou blijven eisen, krijgen we weliswaar heel betrouwbare conclusies, maar dan moeten we maanden, zo niet jaren proeven doen. Omgekeerd, als de tegenstander te gauw tevreden is, dan zijn de proeven snel klaar, maar het resultaat is niet te vertrouwen.

Deze zogenaamde *significantiedrempel*  $\alpha$  varieert van het ene wetenschapsgebied tot het andere, en hangt ook af van het belang van het onderzoek en de beschikbare middelen. Misschien ligt  $\alpha = 5\%$  wat aan de hoge kant: je accepteert dan als het ware dat één op de twintig wetenschappelijke resultaten niet deugt. Dat lijkt nogal veel. Toch zullen we deze waarde van  $\alpha$  meestal hanteren, omdat ze in de biologie praktisch 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* of *overschrijdingskans* van de waarneming. Hij wordt vaak aangeduid 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 misschien 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  $\alpha$  gaan vergelijken. Valt hij kleiner uit dan  $\alpha$ , 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  $\alpha$ , dan ... weten we nog niks! In ons experiment hebben de pissebedden niet genoeg voorkeur laten zien om onze strenge criticus te overtuigen. Maar het kan natuurlijk best zo zijn, dat ze er een (misschien lichte) voorkeur voor vochtige plekken op na houden. Nader onderzoek zou kunnen helpen.

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: pissebedden 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 zo uitzien:

$$VVDVDDDDVDVVDVDDVDDV$$

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  $VVVVVVVVVVVVVVVVVVVVV$ , gelijk aan  $1/524288$ . (Laten we deze kans  $\pi_0$  noemen.)

Goed, dus als de pissebedden alle negentien de vochtige gang in waren gegaan, was onze proef zeker geslaagd geweest, want  $\pi_0 = 1/524288 \approx 0,000001907\dots$  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:

$$\begin{aligned} & DVVVVVVVVVVVVVVVVVVV, VDVVVVVVVVVVVVVVVVVVV, \\ & \dots\dots\dots, \\ & VVVVVVVVVVVVVVVVVVDV, VVVVVVVVVVVVVVVVVVVVD. \end{aligned}$$

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:  $\binom{19}{2}$  (spreek uit: “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  $\pi_3$  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

$$VVVD_3VVVD_1VVVVVD_2VVVV .$$

Maar staan deze rijtjes voor allemaal verschillende proefuitslagen? Nee, natuurlijk! De nummertjes die onder de  $D$ 's 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  $\pi_4$  en  $\pi_5$  op 4 respectievelijk 5  $D$ -afslagen:

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

### 2.3.4 De $P$ -waarde

De  $P$ -waarde of *overschrijdingskans* van onze meetuitslag 14, is nu gedefinieerd als de kans onder de nulhypothese op hoogstens 5  $D$ -afslagen:

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

### 2.3.5 Eenzijdige toetsing

Deze kans is nog steeds klein: hij ligt (nog net) beneden onze overeengekomen 5%-grens. Met de kritische opponent hebben we afgesproken dat hij zich in dit geval gewonnen zou geven. Hij geeft toe: pissebedden hebben een voorkeur voor de vochtige gang. We mogen nu stellen dat er een oorzakelijk verband is tussen de vochtigheid van een plek en het aldaar vóórkomen van pissebedden.

De toets is klaar. Dit was een zogenaamde **eenzijdige binomiale toets** van de **nulhypothese** van de opponent dat pissebedden geen voorkeur zouden hebben voor vocht versus onze **alternatieve hypothese** dat zij zo'n voorkeur wél hebben.

Bovenstaande toets heet **eenzijdig**, omdat onze alternatieve hypothese luidde dat de voorkeur van de pissebedden een bepaalde kant opgaat. Bij de toets zijn we er immers van uitgegaan dat pissebedden, als ze al een voorkeur hebben, er een hebben voor de *vochtige* gang. We zouden zeer verbaasd hebben gestaan als de pissebedden in het experiment plotseling in meerderheid de *droge* gang hadden gekozen! Zo'n uitslag zou voor ons een negatief resultaat zijn geweest, een afgang en koren op de molen van onze kritische opponent; zeker geen reden voor hem om zich gewonnen te geven.

### 2.3.6 Tweezijdige toetsing

Maar er zijn ook onderzoekssituaties waarbij we vooraf geen idee hebben 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*. Deze term staat voor een andere afspraak met de opponent.

De voorwaarde voor verwerping van zijn nulhypothese zou dan luiden:  $P \leq \alpha$ , waarbij  $P$  de kans is op een overschrijding in een van beide richtingen:

$$P = 2^{-19} \left( \binom{19}{0} + \binom{19}{1} + \binom{19}{2} + \binom{19}{3} + \binom{19}{4} + \binom{19}{5} \right) \\ + 2^{-19} \left( \binom{19}{14} + \binom{19}{15} + \binom{19}{16} + \binom{19}{17} + \binom{19}{18} + \binom{19}{19} \right)$$

De twee termen waar deze som uit bestaat zijn gelijk omdat

$$\binom{19}{17} = \binom{19}{2},$$

etcetera (zie (2.2) in §2.4 voor de definitie van  $\binom{n}{k}$ ). Daarom vinden we nu de dubbele  $P$ -waarde:

$$P \approx 2 \times 0,03178 = 0,06356 > \alpha .$$

We merken op dat bij dezelfde meetuitslag van 14 ‘vochtige’ keuzes, de nulhypothese *niet* zou zijn verworpen bij  $\alpha = 0,05$  ! Dus bij gelijke  $\alpha$  stelt tweezijdig toetsing een hogere significantie-eis dan éézijdig toetsing.

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  $\pi_k$  op  $k$  enen en  $n - k$  nullen gelijk is aan:

$$\pi_k = 2^{-n} \binom{n}{k}. \quad (2.1)$$

We noemen dit de *kansverdeling* van het aantal enen.

Hierbij is  $\binom{n}{k}$  als volgt gedefinieerd:

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

Uit de tweede uitdrukking lezen we af dat

$$\binom{n}{n-k} = \binom{n}{k}. \quad (2.2)$$

(Natuurlijk, want er zijn evenveel rijtjes met  $k$  enen en  $n - k$  nullen als rijtjes met  $n - k$  enen en  $k$  nullen!)

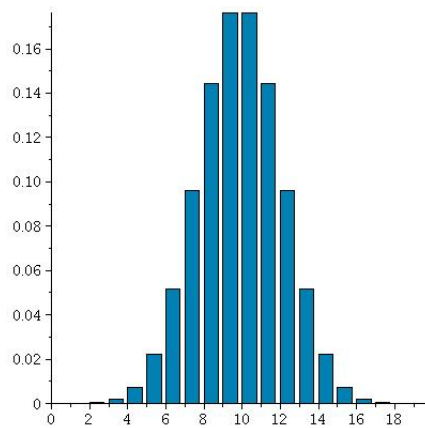
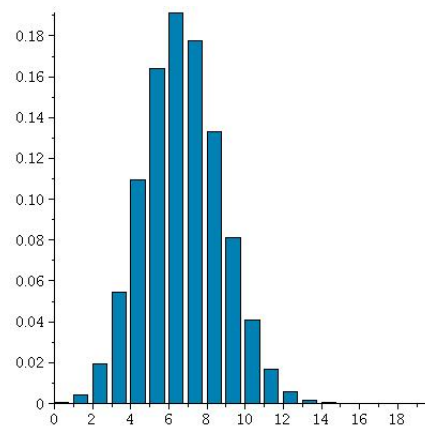
Voor het geval  $n = 19$  vind je de verdeling (2.1) in Tabel 2.1.

### 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  $\pi_0, \pi_1, \dots, \pi_n$  als de ‘succeskans’  $p$  is, en de kans op een mislukking  $q := 1 - p$ ?

Welnu, de kans  $\pi_0$  op 0 successen is dan  $q^n$ . De kans  $\pi_1$  op 1 succes is gelijk aan het aantal mogelijke rijtjes bestaande uit één 1 en  $n - 1$  0-en, vermenigvuldigd met

$k$	$\binom{19}{k}$	$\pi_k = 2^{-19} \binom{19}{k}$
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

Table 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  $p = \frac{1}{3}$

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

$$\pi_k = \binom{n}{k} p^k q^{n-k}. \quad (2.3)$$

De kansen  $\pi_0, \pi_1, \dots, \pi_n$  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. De verdeling (2.3) wordt de *binomiale verdeling* genoemd met parameters  $n$  en  $p$ . Zie Figuur 2.1. Meer hierover in §5.2.

## 2.5 De binomiale toets: algemeen

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

**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 dertig bomen: uit elk daarvan snijden we een stukje bast weg, en analyseren het. De uitslag:  $w$  van deze dertig bomen blijken ziek. (Hierbij is bijvoorbeeld  $w = 16$ ,  $w = 7$ , of  $w = 4$ .) Voor welke waarden van  $w$  stellen wij de boswachter in het ongelijk? En wanneer geven wij hem 'the benefit of the doubt'? (We hanteren een significantieniveau van 5%.)

### 2.5.1 Eenzijdige binomiale toets

Het probleem van de zieke bomen lijkt op dat van de pissebedden. Als *nulhypothese* gebruiken we nu de bewering van de boswachter dat een derde deel van de bomen is aangetast:  $p = \frac{1}{3}$ .

Stel dat we van tevoren een vermoeden hebben: we denken bijvoorbeeld dat het zo erg niet is, en dat een veel kleiner deel van de bomen is aangetast. In dat geval toetsen we de nulhypothese  $p = \frac{1}{3}$  tegen onze *alternatieve hypothese*  $p < \frac{1}{3}$ . (De boswachter krijgt dan gelijk als *nog meer* dan een derde deel van de steekproef aangetast blijkt.)

Dit heet *eenzijdig toetsen*, en het gaat zó in zijn werk: Als de nulhypothese juist is, dan wordt de kans op  $k$  zieke bomen in de steekproef gegeven door (2.3) met  $n = 30$  en  $p = \frac{1}{3}$ . Daarom berekenen we bij de uitslag  $w$  de volgende  $P$ -waarde:

$$P = \sum_{k=0}^w \pi_k . \quad (2.4)$$

We verwerpen de nulhypothese als  $P \leq 5\%$ .

**Dus:** Als  $w = 4$ , dan kunnen we met (2.3) uitrekenen dat

$$P = \pi_0 + \pi_1 + \dots + \pi_4 = 0,012 ,$$

en we verwerpen de nulhypothese: er zijn minder zieke bomen dan de boswachter beweerde.

Als  $w = 7$ , dan is  $P = \pi_0 + \pi_1 + \dots + \pi_7 = 0,167$ , en we geven de boswachter *the benefit of the doubt*.

Als  $w = 16$ , dan is  $P = \pi_0 + \pi_1 + \dots + \pi_{16} = 0,992$ , en we geven de boswachter gelijk: het is behoorlijk mis in het bos!

In het omgekeerde geval, als we denken dat er juist méér bomen ziek zijn dan een derde deel, toetsen we de nulhypothese  $p = \frac{1}{3}$  tegen de alternatieve hypothese dat  $p > \frac{1}{3}$ . In dat geval berekenen we de  $P$ -waarde

$$P = \sum_{k=w}^{30} \pi_k . \quad (2.5)$$

## 2.5.2 Tweezijdige binomiale toets

Als we zelf bij voorbaat geen idee hebben of de boswachter te optimistisch of te pessimistisch is, toetsen we *tweezijdig*. We toetsen de nulhypothese  $p = \frac{1}{3}$  tegen de *alternatieve hypothese* dat  $p \neq \frac{1}{3}$ .

De boswachter verwacht  $30 \times \frac{1}{3} = 10$  zieke bomen in de steekproef. Wijkt het feitelijk gevonden aantal  $w$  hiervan af (wat meestal het geval is!), dan beoordelen we of deze afwijking *significant* is door de volgende  $P$ -waarde uit te rekenen:

$$P := \begin{cases} 2 \sum_{k=0}^w \pi_k & \text{als } w < 10; \\ 2 \sum_{k=w}^{30} \pi_k & \text{als } w > 10. \end{cases} \quad (2.6)$$

Dus als  $w = 10$ , is er geen sprake van een  $P$ -waarde, omdat er geen afwijking van het gemiddelde is. Als  $w < 10$ , dan hebben we een afwijking *naar beneden*, en moeten we voor  $P$  de kansen bij elkaar tellen voor minstens zo grote afwijkingen naar beneden, en daarbij de kans optellen op een minstens zo zeldzame afwijking *naar boven*. Dat laatste is de reden voor de factor 2. Is  $w > 10$ , dan doen we het omgekeerde.

We verwerpen de nulhypothese als  $P \leq 0,05$ .

**Dus:** Als  $w = 4$ , dan is  $P = 2(\pi_0 + \pi_1 + \dots + \pi_4) = 0,024$ , en we verwerpen de nulhypothese: er zijn minder zieke bomen dan de boswachter dacht.

Als  $w = 7$ , dan is  $P = 2(\pi_0 + \pi_1 + \dots + \pi_7) = 0,334$ , en we geven de boswachter *the benefit of the doubt*.

Als  $w = 16$ , dan is  $P = \pi_{16} + \pi_{17} + \dots + \pi_{30} = 0,037$ , en we verwerpen de nulhypothese: er zijn meer zieke bomen dan de boswachter dacht.

## 2.6 Berekening met de grafische rekenmachine

Het met de hand uitrekenen van  $P$ -waarden is een heel karwei. We kunnen gebruik 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  $\pi_k$  uit formule (2.3), 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(`.

Voer nu je argumenten in: eerst de waarde van  $n$ , dan die van  $p$ , en tenslotte een waarde voor  $k$ , zeg  $k = 10$ . Sluit 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  $\pi_k$  moeten kleiner zijn, want de som is 1. Dat  $\pi_{10}$  de grootste is, is wel te begrijpen: 10 is immers  $30 \times \frac{1}{3}$ , de verwachting van de binomiale verdeling met  $n = 30$  en  $p = \frac{1}{3}$ .

### 2.6.2 binomcdf

Dit betekent: *binomial cumulative distribution function*, dus *binomiale cumulatieve verdelingsfunctie*. Dat cumulatieve is precies wat wij willen! In formule (2.4) staat immers de som van alle binomiale kansen van  $\mathbb{P}[K = 0]$  tot en met  $\mathbb{P}[K = w]$ .

(Onder  $\mathbb{P}[\dots]$  verstaan wij: ‘de kans dat  $\dots$ ’.) 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 gemeten waarde  $w$ ; zeg  $w = 4$ .

Er komt:

.0122297238

De kans inderdaad kleiner dan 5%. De nulhypothese wordt verworpen (zie boven).

Bij eenzijdig toetsen naar boven, en ook bij tweezijdig toetsen als  $w > 10$ , willen we uitrekenen:

$$\sum_{k=w}^{30} \pi_k . \quad (2.7)$$

Probleem hierbij is dat onze GR alleen cumulatieve kansen *van onder af* wil geven. We kunnen dit probleem omzeilen door gebruik te maken van het feit dat de kans dat  $K \geq w$  gelijk is aan  $1 -$  (de kans dat  $K < w$ ), en dit weer aan  $1 -$  (de kans dat  $K \leq w - 1$ ). Dus we berekenen `binomcdf` van  $w - 1$ , en trekken het resultaat van 1 af.

$$\mathbb{P}[K \geq 16] = 1 - \mathbb{P}[K \leq 15] = .0187950012 .$$

# Chapter 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 ook een absoluut 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.

### 3.2.1 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

Table 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

Table 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 Fig. 3.4) of een box-whiskerplot, Fig. 3.3. 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 rechts liggen, en de rest links. Het interval tussen het eerste en het derde kwartiel heet het *interkwartielbereik*.

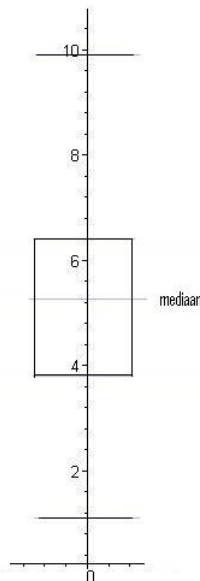


Fig. 3.3: Box-whiskerplot. Mediaan: 5,25, interkwartielbereik: [3,8; 6,4]

**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.

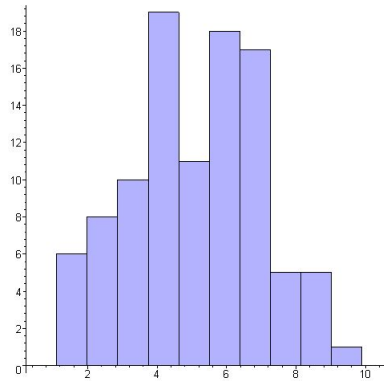


Fig. 3.4: Histogram lengtes vliegende herten

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 het totale aantal gegevens, krijgen we het *relatieve frequentie histogram* met totale oppervlakte 1.

**De relatieve frequentiehoek** 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.

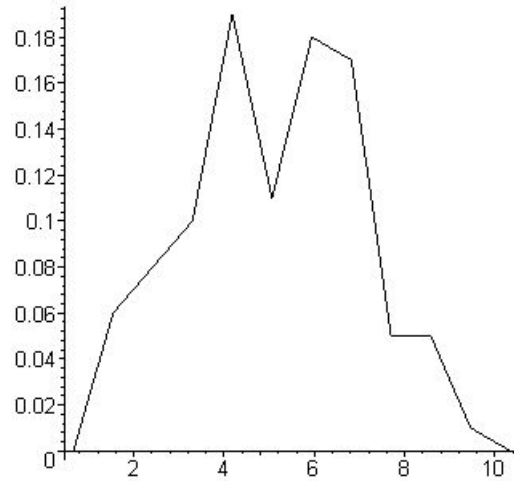


Fig. 3.4: Relatieve frequentievelhoek voor lengtes vliegende herten

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 frequentievelhoek 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 3.4.)

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

**Mediaan:** Bij een oneven aantal gegevens  $x_1 \leq x_2 \leq \dots \leq x_{2m-1}$ : neem de middelste waarde  $x_m$ . Bij een even aantal gegevens  $x_1 \leq x_2 \leq \dots \leq x_{2m}$ : 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:

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

De populatie-variantie is:

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

En de populatie-standaarddeviatie:

$$\sigma = \sqrt{\sigma^2} . \quad (3.3)$$

De variantie is een *kwadratische grootheid*, in tegenstelling tot de standaard-deviatie.

### 3.4 Steekproeven uit continue verdelingen

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

Het steekproef-gemiddelde

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k ; \quad (3.4)$$

de steekproef-variantie

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

en de steekproef-standaarddeviatie

$$s = \sqrt{s^2} . \quad (3.6)$$

### 3.4.1 Notatie en lettergebruik

1. Populatieparameters worden aangegeven met Griekse letters, steekproefvariabelen ('statistieken') met Romeinse letters.
2. Vóórdat we de steekproef nemen, is het steekproefgemiddelde natuurlijk nog onbekend. We geven zo'n *onbekende door het toeval bepaalde variabele* of *stochastische variabele* of kortweg *stochast* aan met Romeinse hoofdletters. Zo is bijvoorbeeld het steekproefgemiddelde  $\bar{x}$  meestal verschillend van  $\mu$ . Maar *gemiddeld over alle steekproeven* is  $\bar{X}$  wèl gelijk aan  $\mu$ . Zie hierover Hoofdstuk 4.
3. 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  $\sum_{i=1}^n (X_i - \bar{X})^2$  niet  $n\sigma^2$  oplevert, maar  $(n - 1)\sigma^2$ ! Er gaat één vrijheidsgraad verloren doordat het gemiddelde  $\bar{X}$  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  $\sigma^2$ . 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 ← 3. Er verschijnt nu:

$$\text{mean}( \quad ,$$

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

$$\text{StDev}( \quad ,$$

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

$$\bar{x} = \frac{1}{100} \cdot (4,816 + \dots + 1,423) = 5,100 \quad (3.7)$$

- variantie

$$s^2 = \frac{1}{99} \cdot [(4,816 - 5,100)^2 + \dots + (1,432 - 5,100)^2] = 3,690 ;$$

- standaarddeviatie

$$s = \sqrt{3,690} = 1,921 . \quad (3.8)$$

### 3.4.4 Schatting:

- $\mu$ : dit is de ‘typische lengte’ van een vliegend hert.  
 $\bar{x}$ : schatter van  $\mu$  op basis van de steekproefgegevens.
- $\sigma$ : de spreiding in de populatie rond het gemiddelde  $\mu$ ;  
 $s$ : schatter van  $\sigma$  op basis van steekproefgegevens.

Dit is het onderwerp van Hoofdstuk 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

# Chapter 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 levende 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  $\bar{x}$  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 van de lengtes, dus het *populatiegemiddelde*  $\mu$  uit (3.1) en de populatie-standaarddeviatie  $\sigma$  uit (3.2). Deze willen we uit onze steekproefgrootheden  $\bar{x}$  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 opinie, 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 bomen erbij zijn. Als dit aantal  $k$  is, dan *schatten* we onze fractie  $p$  met

$$\widehat{p} := \frac{k}{n}. \quad (4.1)$$

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 (met teruglegging) van  $n$  bomen*?

We zouden dan heel veel schattingen van de kans  $p$  krijgen:

$$\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \dots, \widehat{p}_N. \quad (4.2)$$

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  $\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \dots, \widehat{p}_N$  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

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

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

$$\sigma^2 = np(1-p). \quad (4.3)$$

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

$$\sigma_{\hat{p}}^2 = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Dus

$$\sigma_{\hat{p}} = \frac{c}{\sqrt{n}}. \quad (4.4)$$

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:** De spreiding in het gemiddelde van steekproeven ter grootte  $n$  is evenredig met  $1/\sqrt{n}$ .

### 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.

Eerst 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  $\sigma_k^2$  van het aantal zieke bomen in de steekproef is

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

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

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

Dus is

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

Daarmee hebben we (4.3) bewezen.

### 4.1.3 Het schatten van een gemiddelde

Als we een trekking  $x_1, x_2, \dots, x_n$  uit een populatie hebben gedaan, dan ligt het voor de hand als schatter van het populatiegemiddelde  $\mu$  het steekproefgemiddelde  $\bar{x}$  te nemen. Maar ook hier is het duidelijk dat  $\mu$  best ook iets groter of kleiner kan zijn dan  $\bar{x}$ . Hoe betrouwbaar is onze schatting van  $\mu$ ?

Ditmaal zonder bewijs herhalen we de  $\sqrt{n}$ -wet:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} . \quad (4.5)$$

Deze fout in het steekproefgemiddelde  $\sigma_{\bar{X}}$  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  $\bar{X}$ . Dit doen we bij voorkeur als de steekproef nog niet is genomen, en het steekproefgemiddelde dus nog onbekend of **stochastisch** is. Alleen dan kun je immers zinvol van een *spreiding* spreken.

## 4.2 De centrale limietstelling

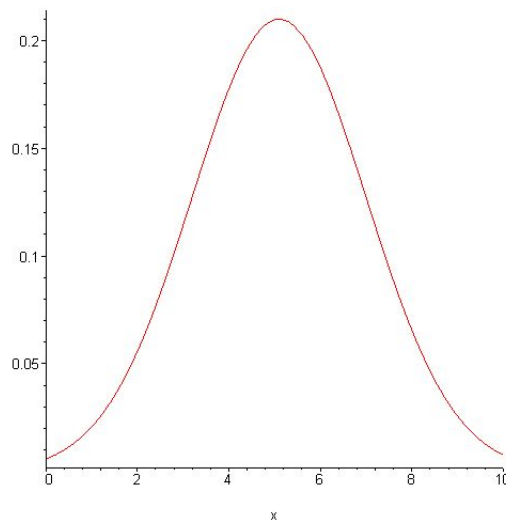
Behalve op de  $\sqrt{n}$ -wet is de statistiek gebaseerd op nog een tweede wonder: Terwijl bij toenemende steekproefgrootte de kansverdeling van het gemiddelde  $\bar{X}$  steeds smaller wordt, neemt zij bovendien steeds meer een vaste vorm aan, een vorm die altijd dezelfde is, en die niet afhangt van de vorm van de kansverdeling die in de populatie heerst: de klokvorm of **normale verdeling**. Dit wonder wordt uitgedrukt in 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  $f$  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 1 bedraagt. Bij zo'n kansdichtheidsfunctie kun je voor elk interval een kans definiëren: de kans dat een grootte 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 als het interval groot wordt (naar beide kanten), nadert die kans tot 1.

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 grootte  $X$  die deze functie als kansdichtheid heeft, noemen we *normaal verdeeld*. Zo'n stochastische variabele heeft gemiddelde  $\mu$  en variantie  $\sigma^2$ . 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  $\mu$  en variantie  $\sigma$ , dan is de grootte

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

standaard-normaal verdeeld. Er is dus voor de oppervlaktes onder de grafiek van *alle* normale verdelingsfuncties maar één tabel nodig. Dit is Tabel 4.5 aan het einde van dit hoofdstuk. In deze tabel staan de waarden van de functie  $\Phi$  getabelleerd, die wordt gegeven door

$$\Phi(z) := \mathbb{P}[Z \leq z] = \int_{-\infty}^z \phi_{0,1}(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{1}{2}x^2} dx .$$

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 Demonstratie van de centrale limietstelling

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

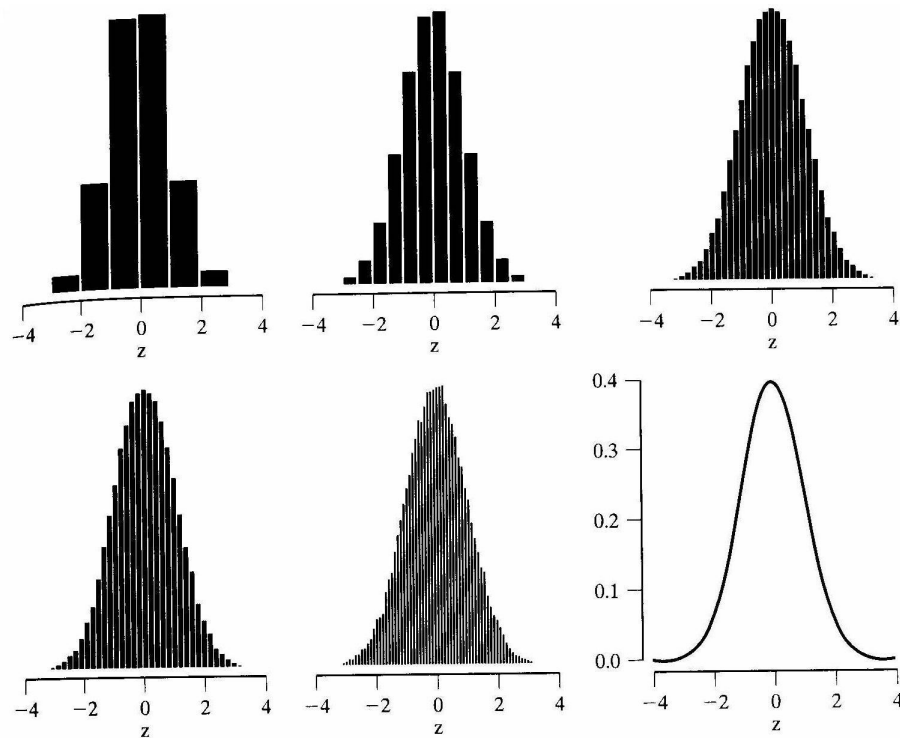
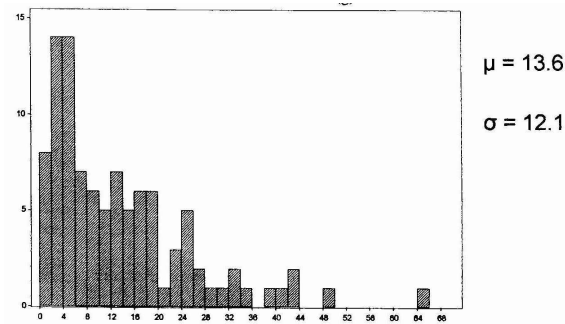


FIGURE 4.2: The progression of a histogram to a continuous distribution as the sample size increases.

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

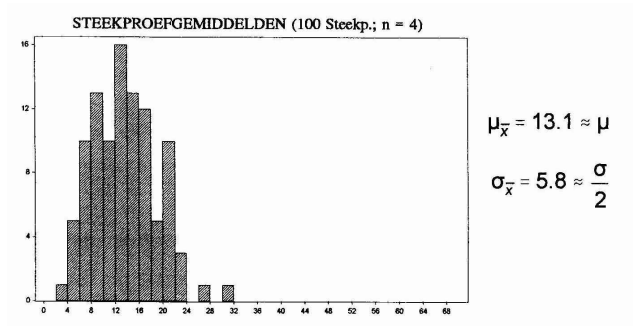
**Een computereperiment** We nemen een willekeurige dataset, en tekenen het histogram:



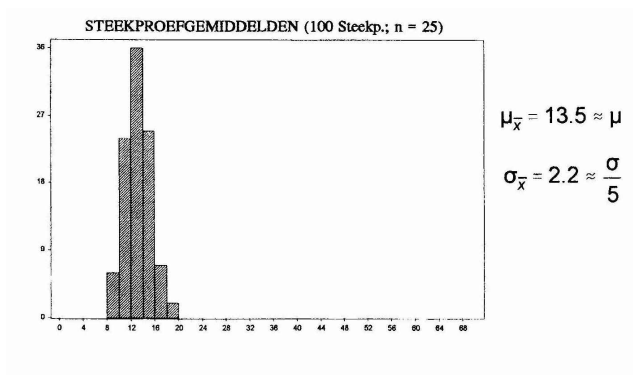
Aan het histogram zien we dat de verdeling sterk asymmetrisch is. We nu doen even net of dit een *gehele populatie* is. Dan zijn de populatieparameters:

$$\mu = 13,58, \quad \sigma^2 = 146,66, \quad \text{dus} \quad \sigma = 12,11 .$$

We nemen een aantal steekproeven van verschillende grootten uit deze populatie: 100 steekproeven met  $n = 4$  en 100 steekproeven met  $n = 25$ . Hier is de uitslag van deze series steekproeven:



voor  $n = 4$  en



voor  $n = 25$ . We merken op:

- De steekproefgemiddelden  $\bar{X}$  zijn bij benadering normaal verdeeld rond het populatiegemiddelde  $\mu$ . Dit is het geval ondanks de asymmetrische populatieverdeling.

- De standaarddeviatie van deze verdeling, die de **standaardfout** wordt genoemd, is een maat voor hoever de steekproefgemiddelden van het populatiegemiddelde verwijderd zijn. Hij is evenredig met de spreiding  $\sigma$  van de onderliggende populatie en omgekeerd evenredig met de wortel uit de steekproefgrootte. Dit is de genoemde  $\sqrt{n}$ -wet:

$$\text{Standaardfout} = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

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

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

- *bij benadering normaal verdeeld*
- *met gemiddelde  $\mu_{\bar{X}} = \mu$  en variantie  $\sigma_{\bar{X}}^2 = \sigma^2/n$*

### 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 hier in de praktijk mee om te gaan.

1. Als de onderliggende populatie *normaal* verdeeld is, dan is  $\bar{X}$  altijd — onafhankelijk van de steekproefgrootte — normaal verdeeld met gemiddelde  $\mu$  en standaarddeviatie  $\sigma/\sqrt{n}$ .
2. Als de verdeling van de onderliggende populatie niet normaal is, maar wel *symmetrisch*, dan is  $\bar{X}$  al bij relatief kleine steekproeven ( $n \geq 10$ ) ‘in goede benadering’ normaal verdeeld met gemiddelde  $\mu$  en standaarddeviatie  $\sigma/\sqrt{n}$ .

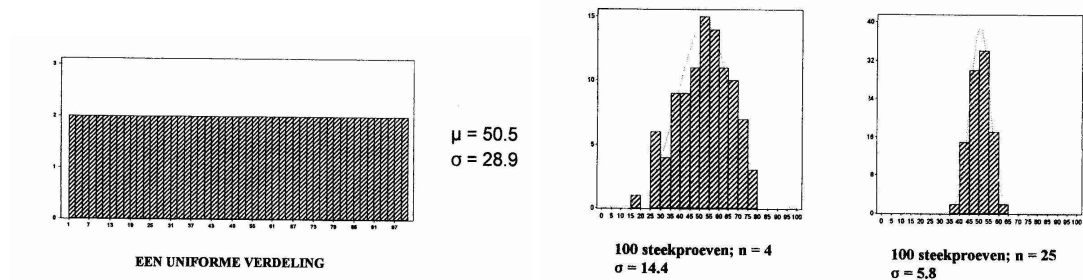
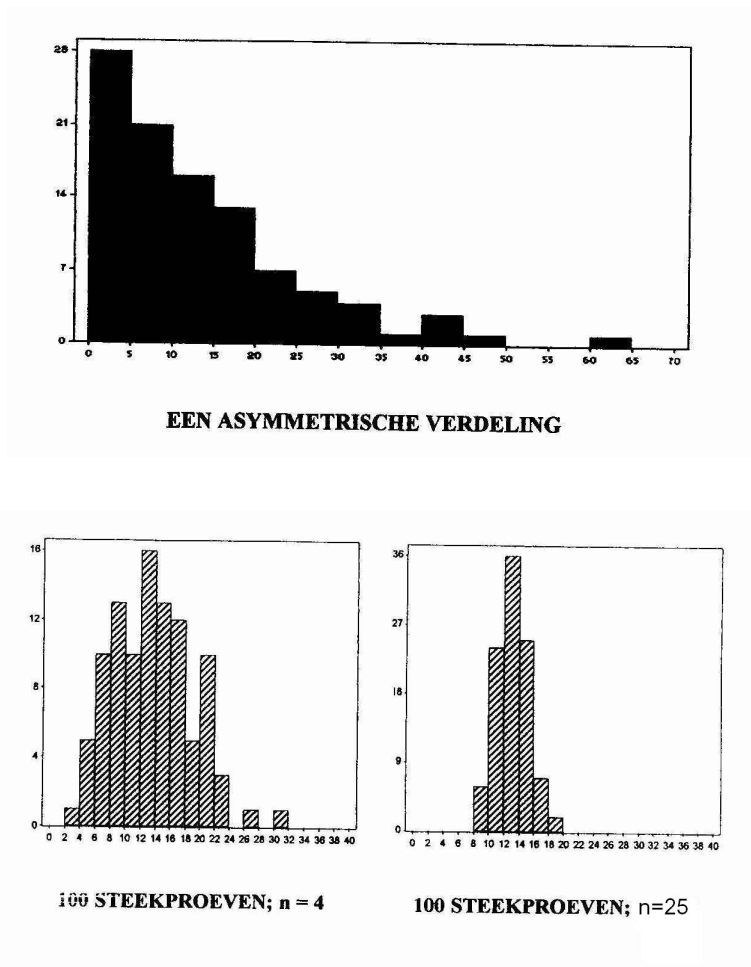


Fig. 4.2.3: Steekproeven uit de uniforme verdeling

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



## 4.3 De z-toets

### 4.3.1 De z-toets voor een kans

In paragraaf 2.5 zagen we hoe je een bewering over een kans kunt toetsen met een steekproef. Het ging toen over een steekproef van 30 bomen, en we gebruikten de binomiale toets om te kijken of een derde deel was aangetast door een ziekte. Maar soms zijn de steekproeven veel groter: we houden bijvoorbeeld een opiniepeiling onder 1000 kiesgerechtigden om de fractie van de bevolking te achterhalen die op een bepaalde politieke partij zou stemmen. De `binomcdf` van je rekenmachine is dan niet zo handig meer. We gebruiken daarom de centrale limietstelling.

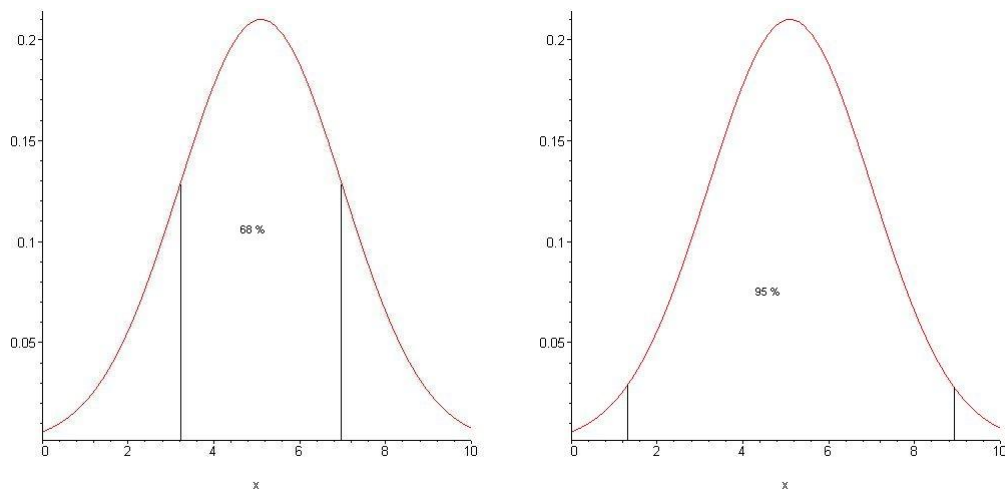


Fig. 4.3: 68% van de gegevens bij een normale verdeling liggen in het interval  $[\mu - \sigma, \mu + \sigma]$  en 95% in het interval  $[\mu - 2\sigma, \mu + 2\sigma]$

**De procedure.** De willen de (nul)hypothese toetsen dat een fractie  $p_0$  van de populatie eigenschap  $A$  heeft. We nemen een steekproef van  $n$  uit de populatie. (Hierbij moet  $n$  zeker boven de 10 liggen; als  $p_0$  ver van  $\frac{1}{2}$  af ligt, dan boven de 30.) In de steekproef hebben  $k$  individuen de eigenschap  $A$ . We bepalen de *schatting*  $\hat{p} := \frac{k}{n}$  van  $p_0$ . Volgens de centrale limietstelling is deze (onder de nulhypothese) een trekking uit een normale verdeling met gemiddelde  $p_0$  en variantie  $\frac{\sigma^2}{n} = \frac{p_0(1-p_0)}{n}$ . Omdat we  $p_0$  niet kennen, maar  $\hat{p}$  daar behoorlijk dicht bij ligt, nemen we voor de *standaardfout* s.e. niet  $\frac{p_0(1-p_0)}{\sqrt{n}}$  maar  $\frac{\hat{p}(1-\hat{p})}{\sqrt{n}}$ .

Onze  $P$ -waarde wordt, bij tweezijdige toetsing,

$$P = \mathbb{P}\left[|Z| > \frac{|\hat{p} - p_0|}{\text{s.e.}}\right] = 2 \left(1 - \Phi\left(\frac{|\hat{p} - p_0|}{\text{s.e.}}\right)\right).$$

Bij éézijdige toetsing tegen de alternatieve hypothese  $p > p_0$  wordt de  $P$ -waarde:

$$P = \mathbb{P}[Z > \hat{p} - p_0] = 1 - \Phi\left(\frac{\hat{p} - p_0}{\text{s.e.}}\right).$$

### 4.3.2 Betrouwbaarheidsinterval voor een kans

We zullen de hypothese dat  $p = p_0$  accepteren (d.w.z. niet verwerpen) als de tweezijdige  $P$ -waarde groter is dan  $1 - \gamma$ . Uit het bovenstaande volgt dat dit het geval is wanneer

$$\mathbb{P}\left[|Z| < \frac{|\hat{p} - p_0|}{\text{s.e.}}\right] = 1 - P < \gamma.$$

Omdat  $\mathbb{P}[|Z| < z_\gamma] = \gamma$ , is dit equivalent met

$$|\hat{p} - p_0| < z_\gamma \cdot \text{s.e.},$$

oftewel  $p_0$  ligt in het interval

$$(\hat{p} - z_\gamma \cdot \text{s.e.}, \hat{p} + z_\gamma \cdot \text{s.e.})$$

waarbij  $z_\gamma$  zó gekozen is dat  $1 - \Phi(z_\gamma) = \frac{1}{2}(1 - \gamma)$ .

Bovenstaand interval is het *betrouwbaarheidsinterval* voor de kans  $p_0$  bij een *betrouwbaarheid*  $\gamma$  dat door onze uitslag wordt bepaald.

### 4.3.3 Theoretisch betrouwbaarheidsinterval voor een gemiddelde

Veronderstel dat we een populatie hebben waarin van ieder individu de grootte  $X$  kan worden gemeten. We nemen een steekproef; de gemiddelde waarde van  $X$  in de steekproef is  $\bar{X}$ .

Stel verder dat je van de populatie **wel** de variantie  $\sigma^2$  kent, maar **niet** het gemiddelde  $\mu$ . (Dit is tamelijk uitzonderlijk! Zie paragraaf 4.4 voor als hier *niet* aan is voldaan.)

Met behulp van de centrale limietstelling kunnen we dan quantificeren hoe groot de verwachte afwijking is tussen de schatter  $\bar{X}$  en de te schatten waarde  $\mu$ :

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

Dus:  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  is standaardnormaal verdeeld. Kies eens even  $\gamma = 95\%$ , dan is  $z_\gamma = 1,96$ .

$$\mathbb{P} \left[ -1,96 < Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1,96 \right] = 0,95$$

$$\mathbb{P} \left[ -1,96\sigma/\sqrt{n} < \bar{X} - \mu < 1,96\sigma/\sqrt{n} \right] = 0,95$$

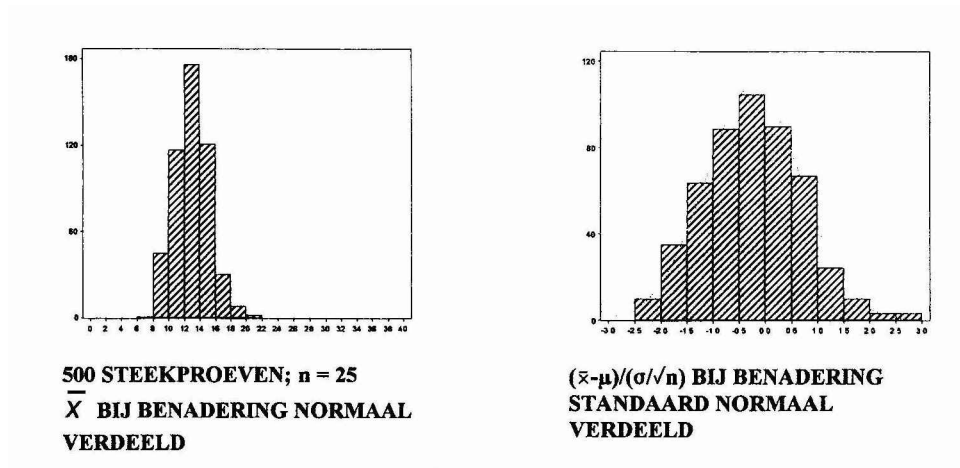
$$\mathbb{P} \left[ \bar{X} - 1,96\sigma/\sqrt{n} < \mu < \bar{X} + 1,96\sigma/\sqrt{n} \right] = 0,95$$

Dus met waarschijnlijkheid 95% ligt de echte waarde  $\mu$  in het interval

$$[\bar{X} - 1,96\sigma/\sqrt{n}, \bar{X} + 1,96\sigma/\sqrt{n}]$$

Dus een 95% betrouwbaarheidsinterval wordt gegeven door:

$$\mu = \bar{X} \pm 1,96 \cdot \sigma/\sqrt{n}$$



Gewichten vliegende herten:  $\mu = 13,58, \sigma/\sqrt{n} = 2,42$

#### 4.3.4 Definitie van $z_\alpha$ en $z_\gamma$ :

Onder  $z_\alpha$  verstaan we het positieve getal waarvoor geldt:

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

Dit is de grenswaarde waar volgens de nulhypothese de waarneming met kans  $\alpha$  boven ligt. Dus ook de waarde **waarboven** we bij een éézijdige toets met significantieniveau  $\alpha$  de nulhypothese verwerpen. We merken nog op dat, wegens de symmetrie van de normale verdeling:

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

Onder  $z_\gamma$  verstaan we het positieve getal waarvoor geldt:

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

Wij zullen ook van deze conventie gebruik maken, meestal in de context van betrouwbaarheidsintervallen. De reden is dat in de literatuur bijvoorbeeld vaak wordt gesproken van  $z_{95\%}$  in plaats van  $z_{2,5\%}$ .

**Vuistregel:**  $\gamma$  ligt dicht bij 1, en  $\alpha$  dicht bij 0. De waarde van  $z_\gamma$  is *dezelfde als* die van  $z_\alpha$  mits

$$\gamma + 2\alpha = 1; .$$

**Berekening van  $z_\alpha$  en  $z_\gamma$  met de GR:** Roep middels 2nd DISTR 3 de functie `invNorm` op. Typ de waarde van  $1 - \alpha$  in. Uitkomst:  $z_\alpha$ . ( $z_\gamma$  vind je door eerst  $\alpha = \frac{1}{2}(1 - \gamma)$  uit te rekenen, en dan daarbij  $z_\alpha$ .)

### 4.3.5 Een 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  $\mu$  bepaald worden; alleen moet dan de waarde  $z_{0,95} = 1,96$  vervangen worden:

- 95%-betrouwbaarheidsinterval:  $\mu = \bar{X} \pm 1,96\sigma/\sqrt{n}$
- 99%-betrouwbaarheidsinterval:  $\mu = \bar{X} \pm 2,58\sigma/\sqrt{n}$
- 99,9%-betrouwbaarheidsinterval:  $\mu = \bar{X} \pm 3,29\sigma/\sqrt{n}$

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

- Stel:  $\sigma = 2$ ,  $n = 25$  en  $\bar{X} = 12,1$
- 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: In de praktijk is de populatie-standaarddeviatie  $\sigma$  niet bekend!

**Oplossing voor grote  $n$ :**

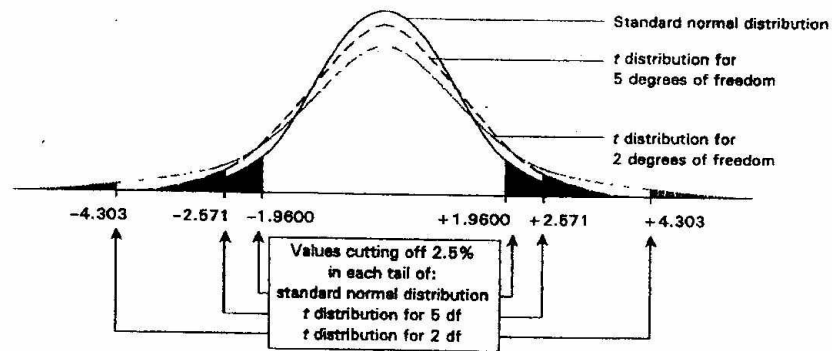
- Voor grote steekproeven ( $n \geq 100$ ) geldt met zeer grote kans:  $S \approx \sigma$ . Dan is  $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$  is bij benadering standaard normaal verdeeld, en dus wordt een  $\gamma$ -betrouwbaarheidsinterval gegeven door:  $\mu = \bar{x} \pm z_\gamma \cdot s/\sqrt{n}$
- Voor kleine steekproeven hebben we een geheel nieuwe stelling nodig!

**4.4.1 De  $t$ -verdeling**

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

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

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



Grafieken van  $t$ -verdelingsfunctie bij  $\nu$  vrijheidsgraden

**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 ( $n \geq 30$ ).

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

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

#### 4.4.2 De $t$ -toets voor een gemiddelde

Veronderstel dat we een populatie hebben waarin van elk individu een grootte  $X$  kan worden gemeten. We nemen een steekproef ter grootte  $n$ . Stel verder dat  $X$  in deze populatie normaal verdeeld is met gemiddelde  $\mu$  en variantie  $\sigma^2$ , òfwel  $n$  is tamelijk groot (zie voor de meer precieze voorwaarden paragraaf 4.4.1). We willen de hypothese toetsen dat het gemiddelde een bepaalde waarde  $\mu_0$  heeft. (Een veel voorkomend voorbeeld is  $\mu_0 = 0$ .) De toets werkt als  $\mu$  en  $\sigma$  beide onbekend zijn.

We kunnen nu stelling 2 toepassen. Dit komt erop neer dat we  $\sigma$  schatten uit de steekproef, en deze schatting gebruiken om de nauwkeurigheid van onze schatting van  $\mu$  mee te bepalen.

**Toetsingsprocedure.** We bepalen het steekproefgemiddelde  $\bar{x}$  en de standaardfout s.e. gegeven door

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

We verwerpen de hypothese dat  $\mu = \mu_0$  als

$$|\bar{x} - \mu_0| > t_{\alpha/2} \cdot \text{s.e.}$$

Dit is een tweezijdige toets met significantie  $\alpha$ , waarbij de nulhypothese  $\mu = \mu_0$  wordt getoetst tegen de alternatieve hypothese dat  $\mu \neq \mu_0$ . De eenzijdige toets, waarbij  $\mu = \mu_0$  wordt getoetst tegen  $\mu > \mu_0$ , gaat zó: we verwerpen de nulhypothese als

$$\bar{x} > \mu_0 + t_\alpha \cdot \text{s.e.}$$

#### 4.4.3 Praktisch betrouwbaarheidsinterval voor een gemiddelde

Direct hiermee verbonden is het betrouwbaarheidsinterval voor het gemiddelde  $\mu$ : met kans  $\gamma$  ligt  $\mu$  in het interval

$$(\bar{x} - t_\gamma \cdot \text{s.e.}, \bar{x} + t_\gamma \cdot \text{s.e.}).$$

#### 4.4.4 Eigenschappen van de t-verdeling

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

#### 4.4.5 Voorbeelden

$\gamma$ -betrouwbaarheidsinterval voor de activiteit van het enzym isomerase in het spierweefsel van verschillende diersoorten

$\gamma = 0,95$	$n$	$s$	s.e.	$t_\gamma$	$\bar{X} \pm t_\gamma \cdot \text{s.e.}$
muis	4	0,18	0,09	3,182	$0,76 \pm 0,29$
kikker	6	0,16	0,06	2,571	$1,53 \pm 0,17$
kreeft	6	0,24	0,10	2,571	$1,06 \pm 0,25$
forel	10	0,60	0,19	2,262	$4,22 \pm 0,43$

**Voorbeeld:** Een etholoog 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:  $\text{s.e.} = \frac{40}{\sqrt{25}} = 8$
- 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

$z$	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

### 4.5.1 Handleiding

Deze tabel geeft de kans  $\Phi(z)$  dat  $Z \leq z$  voor een standaard-normaal verdeelde stochast  $Z$ . In de eerste kolom staan de cijfers van  $z$  aan weerszijden van de komma. Het tweede cijfer van  $z$  achter de komma staat boven de kolom met de uitkomst. Voor de cijfers hiervan moet '0,' worden gezet. 893 betekent dus: 0,893. De tabel kan omgekeerd ook gebruikt worden om de  $z$ -waarden te vinden bij een kans  $\alpha$ : dit is het getal  $z_\alpha$  waarvoor  $\Phi(z_\alpha) = \alpha$ . Zoek  $\alpha$  op in de tabel; dan staat (een benadering

van)  $z_\alpha$  in de marge.

Voorbeeld:  $\mathbb{P}[Z \leq 1,24] = 0,893$  en  $z_{0,05} = 1,96$ .

### 4.5.2 Normale verdeling 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.

## 4.6 Kritieke waarden voor Student t-verdelingen

$\nu$	$t_{0,05}$	$t_{0,025}$	$t_{0,01}$	$t_{0,005}$
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	2,528	2,845
21	1,721	2,080	2,516	2,831
22	1,717	2,074	2,507	2,819
23	1,714	2,069	2,499	2,807
24	1,711	2,064	2,492	2,797
25	1,708	2,060	2,485	2,787
26	1,706	2,056	2,479	2,779
27	1,703	2,052	2,473	2,771
28	1,701	2,048	2,467	2,763
29	1,699	2,045	2,462	2,756
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
$\infty$	1,645	1,960	2,326	2,576

### 4.6.1 Handleiding

Voor de gevallen  $\alpha = 0,05, 0,025, 0,01, 0,005$  wordt de grenswaarde  $t_\alpha$  gegeven met de eigenschap dat  $\mathbb{P}[T \geq t_\alpha] = \alpha$  voor een Student- $t$ -verdeelde stochast  $T$  met  $\nu$  vrijheidsgraden. De standaard normale verdeling is de  $t$ -verdeling met oneindig veel

vrijheidsgraden:  $\nu = \infty$ .

### **4.6.2 Berekening met de GR**

Typ 2nd DISTR 4 en geef de parameterwaarden  $1 - \alpha$  en  $\nu$  in, gescheiden door een komma en gevolgd door ) ENTER. De grenswaarde  $t_\alpha$  verschijnt.

# Chapter 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.

Vlakbij een aquarium met stekelbaarsjes staat een tank waarin een exemplaar van de roofvis *Cichlide* rondzwemt. In het aquarium zelf zijn voederbakjes met *Tubifex*-wormpjes aangebracht op verschillende afstanden van de tank met de roofvis. De stekelbaarsjes lijken minder te eten uit de bakjes dicht bij de *Cichlide*. Mijden zij deze roofvis? Om deze vraag te beantwoorden, moeten we de nulhypothese toetsen dat hun fouragegedrag niet door predatoren wordt beïnvloed, zodat het feitelijk *uniform* over de bakjes verdeeld is.

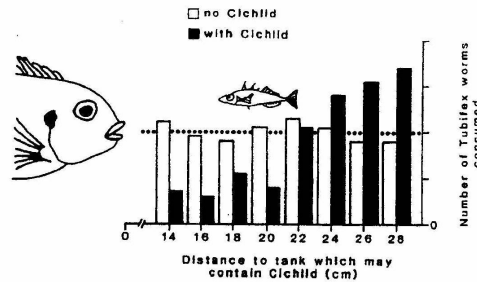


Fig. 12.1 The relationship between the number of *Tubifex* worms eaten by sticklebacks and the distance to a tank which sometimes contains a cichlid predator. (After Milinski 1985.)

Een ander voorbeeld van de uniforme verdeling is die van een zuivere dobbelsteen:

- Mogelijke uitkomsten  $k = 1, \dots, 6$  uniform verdeeld:  $\pi_k = \frac{1}{6}$ .

De algemene discreet-uniforme verdeling ziet er zó uit:

- Mogelijke uitkomsten:  $k = 1, \dots, n$ .
- Kans op uitkomst  $k$ :  $\pi_k = \frac{1}{n}$ .

## 5.2 De binomiale verdeling

Deze hebben 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:

- We hebben een binaire klassificatie: De populatie is verdeeld in twee categorieën A en B. (bijv. man/vrouw, blauw/bruin, geïnfecteerd/niet geïnfecteerd).
- We doen  $n$  onafhankelijke trekkingen uit de populatie. Deze zijn elk òf A òf B zijn (in het vlindervoorbeeld:  $n = 5$ ).
- Elke getrokken individu heeft dezelfde kans  $p$  om A te zijn. (In het vlindervoorbeeld staat A voor ‘blauw’ en  $p = 0,4$ .)

De kans  $\pi_k$  dat precies  $k$  van de  $n$  individuen van type A zijn, wordt gegeven door de *binomiale verdeling*:

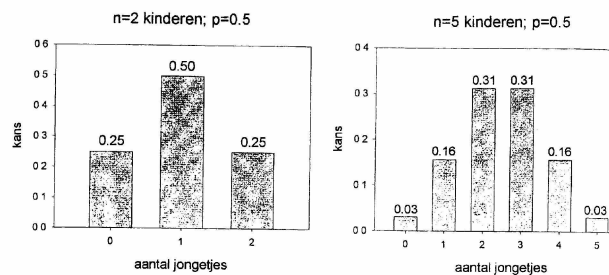
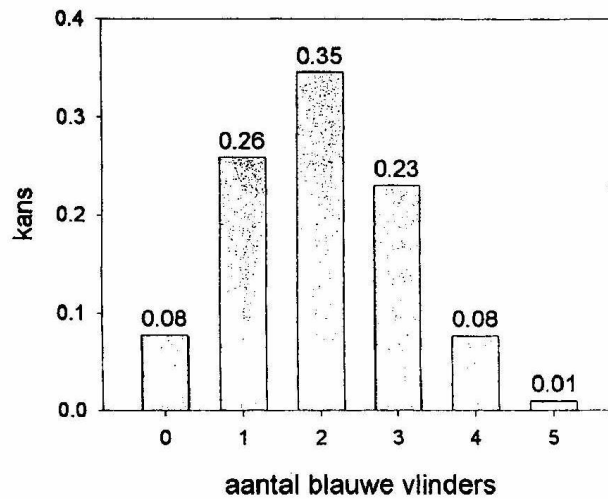
$$\pi_k = \binom{n}{k} p^k (1-p)^{n-k},$$

waarbij het ‘grijpgetal’  $\binom{n}{k}$  wordt gevonden met de volgende formule:

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

(Onder en boven staan  $k$  factoren.)

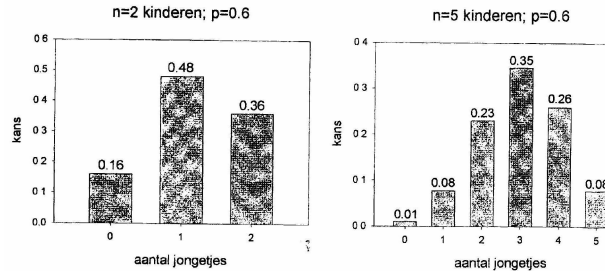
We geven de binomiale verdeling met parameters  $p$  en  $n$  aan met  $\text{Bin}(n, p)$ .



### 5.2.1 Eigenschappen van de binomiale verdeling:

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

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



### 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  $\lambda$  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:

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

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  $\lambda$  als

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

(Hierbij is  $0!$  gedefinieerd als 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 van vragen waar de Poissonverdeling het antwoord op geeft:

- Als agarplaten gemiddeld tien bacterie-kolonies bevatten, wat is de kans dat een gegeven plaat twaalf kolonies bevat?
- Als bij een plant per uur gemiddeld 3,5 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 vijf muizen vangt, wat is dan de kans dat dat ze in een bepaalde nacht zeven 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-, lengte-, oppervlakte- of volume-eenheid gelijk is aan  $\lambda$ .

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

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

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

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

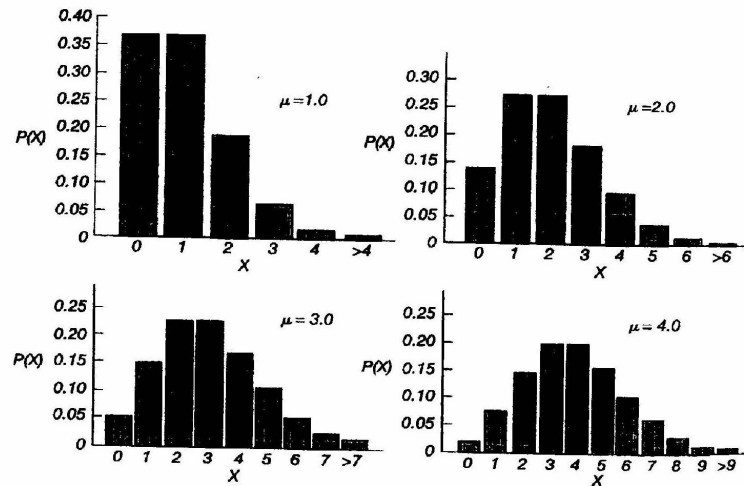


Figure 24.1 The Poisson distribution for various values of  $\mu$ . These graphs were prepared by using Equation 24.1.

### 5.3.1 Eigenschappen van de Poissonverdeling.

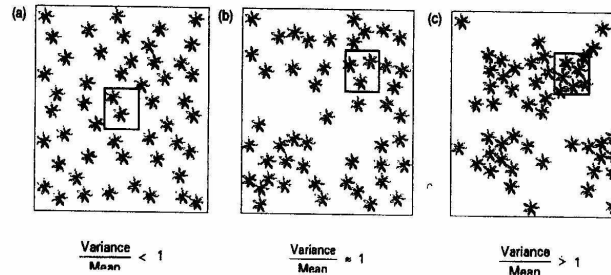
$$\text{Gemiddelde} : \quad \mu = \sum_k \pi_k \cdot k = \lambda ;$$

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

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 insecten 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 Hoofdstuk 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 5.4.5, of de Poissonverdeling: Paragraaf 5.4.6.

Maar eerst hebben we een nieuw theoretisch instrument nodig, een soort meerdimensionale normale verdeling. Dit is het onderwerp van de volgende paragraaf:

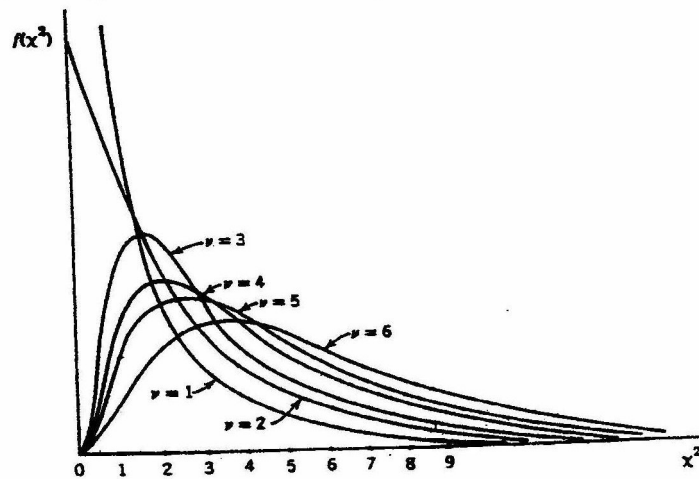
### 5.4.1 De $\chi^2$ -verdeling

**Definitie:** De kansverdeling van de som van de kwadraten van  $\nu$  onafhankelijke standaard-normaal verdeelde stochasten

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

wordt de  $\chi^2$ -verdeling met  $\nu$  vrijheidsgraden genoemd.

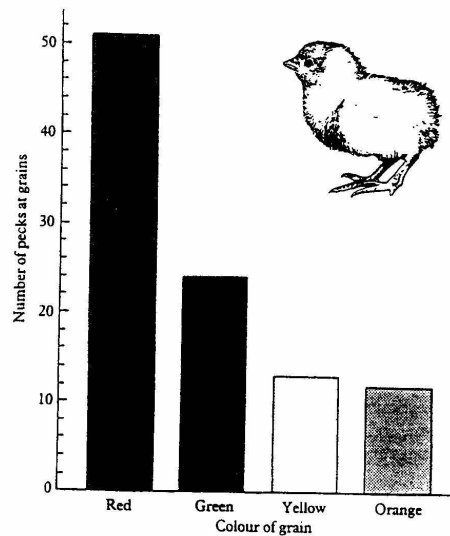
$\chi^2$ -verdelingen en hun afhankelijkheid van  $\nu$ :



**Berekening van  $\chi^2$ -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 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  $m$  categorieën door een kansproces tot stand komt (bijvoorbeeld de uniforme verdeling), dan is de toetsgrootheid

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

bij benadering verdeeld volgens de “ $\chi^2$ -verdeling met  $\nu = m - 1$  vrijheidsgraden”.

De toetsingsgrootheid  $X^2$  noemen we de  $\chi^2$ -waarde (‘*chi-kwadraat-waarde*’);  $X$  is bedoeld als een *hoofdletter chi*!

Hier is  $o_i$  het geobserveerde aantal, en  $e_i$  het verwachte aantal volgens de veronderstelde kansverdeling. In ons voorbeeld is dit de uniforme verdeling: we veronderstellen de kansen op alle kleuren gelijk, zodat  $e_i = 25$  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  $X^2 = 27,04 + 0,04 + 5,76 + 6,76 = 39,60$ .

### 5.4.3 De $\chi^2$ -toets van een discrete verdeling

We toetsen de *nulhypothese* dat de kuikens geen voorkeur hebben voor een bepaalde kleur. Onder deze hypothese is  $X^2$  verdeeld volgens de  $\chi^2$ -verdelingsfunctie met  $\nu = 4 - 1 = 3$  vrijheidsgraden.

Eerst berekenen we met formule (5.2) de  $\chi^2$ -waarde van onze waarneming. (In het kuikenvoorbeeld:  $X^2 = 39,6$ .) Dan bepalen we met de  $\chi^2$ -verdeling onze  $P$ -waarde: dit is de kans  $P$  dat, terwijl de nulhypothese waar is, onze  $\chi^2$ -waarde of een grotere zou optreden. In het voorbeeld vertelt de rekenmachine dat  $P = 1,3 \cdot 10^{-8}$ . Omdat dit zo'n klein getal is, verwerpen we de nulhypothese. In het algemeen doen we dit als  $P < \alpha$ , waar  $\alpha$  het significantieniveau aanduidt waarmee we deze  $\chi^2$ -toets willen uitvoeren. (Vaak is dit 5%.) De kuikens hebben dus wel degelijk een voorkeur.

In plaats van de  $P$ -waarde, die we met computer of rekenmachine kunnen vinden, kunnen we ook gebruik maken van de *kritieke waarde* voor de toetsgrootheid  $X^2$ , die in de tabel in paragraaf 5.6 kan worden opgezocht. Dit is de waarde van  $X^2$  waarbij de  $P$ -waarde onder ons gekozen significantieniveau  $\alpha$  duikt. Deze waarde hangt af van  $\alpha$  en van het aantal vrijheidsgraden  $\nu$  in onze proef. In de tabel staan kritieke waarden voor twee significantieniveaux:  $\alpha = 0,05$  en  $\alpha = 0,01$ . In ons kuikenvoorbeeld is  $\nu = 4 - 1 = 3$ , zodat de kritieke  $\chi^2$ -waarde 7,815 is als we  $\alpha = 5\%$  kiezen, en 11,345 als  $\alpha = 1\%$ . Allebei liggen ruim onder 39,6.

### 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.

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  $\chi^2$ -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  $\chi^2$ -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  $\lambda$  *uit de gegevens*.

### 5.4.5 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  $\chi^2$ -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 met kracht!

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). Met dit gegeven kunnen we opnieuw een  $\chi^2$ -goodness-of-fit-toets doen, nu met de binomiale verdeling met parameterwaarden  $n = 6115$  en  $p = 0,52$ .  $X^2$  berekend op basis van deze verdeling geeft 116,6. (Zie tweede tabel.) Deze  $P$ -waarde moet nu worden berekend met de  $\chi^2$ -verdeling met slechts 11 vrijheidsgraden. Eén vrijheidsgraad gaat verloren omdat de waarde  $p = 0,52$  is ontleend aan de steekproef,

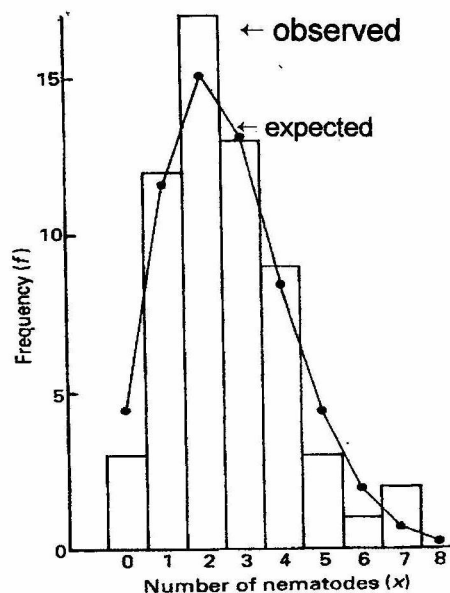
niet aan ons theoretisch model, zoals in paragraaf 5.4.4 is uitgelegd. Het blijkt dat de  $\chi^2$ -waarde nog steeds uiterst zeldzaam is: de  $P$ -waarde is ongeveer  $8,7 \cdot 10^{-20}$ .

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	$\chi^2 = 116,6$

De verklaring van Geissler was een andere: in sommige gezinnen is de kans op een jongetje groter dan 0,5, in andere gezinnen juist kleiner.

### 5.4.6 Het toetsen van de Poissonverdeling: nematoden

o De gegevens:



#### 5.4. 'GOODNESS-OF-FIT': HET TOETSEN VAN DISCRETE VERDELINGEN 65

We hebben een preparaat van nematoden (rondwormen) onder de microscoop liggen, en vragen ons af of deze diertjes zich gelijkmatig over het oppervlak hebben verspreid. Dit zou aanleiding geven tot een Poisson-verdeling (paragraaf 5.3.2). We gaan deze verdeling toetsen met een  $\chi^2$ -goodness-of-fit-toets. En wel als volgt:

- Biologische nulhypothese: nematoden verdelen zich random over het oppervlak.
- Statistische nulhypothese: de kans dat in een bepaalde eenheid van oppervlak  $k$  nematoden worden aangetroffen, wordt gegeven door een Poissonverdeling waarvan de parameter  $\lambda$  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 ( $= 3 \times 0 + 12 \times 1 + 17 \times 2 + \dots$ )
- gemiddelde aantal per cel:  $156/60 = 2,6$ . Het lijkt daarom een goed idee om een Poissonverdeling met  $\lambda = 2,6$  te gebruiken.
- Na 'poolen' van de categorieën 5, 6, en 7 vinden we  $X^2 = 1,04$ .
- Het aantal vrijheidsgraden is  $\nu = 6 - 1 - 1 = 4$ .

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  $\chi^2$ -toets van 'goodness-of-fit' vul je alleen aantallen in, nooit percentages, proporties of meetuitslagen.
- Een  $\chi^2$ -verdeling is niet van toepassing als de verwachte frequenties van bepaalde categorieën 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 categorieën te 'poolen'.

## 5.5 De $\chi^2$ -toets voor onafhankelijkheid

We kunnen de  $\chi^2$ -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.

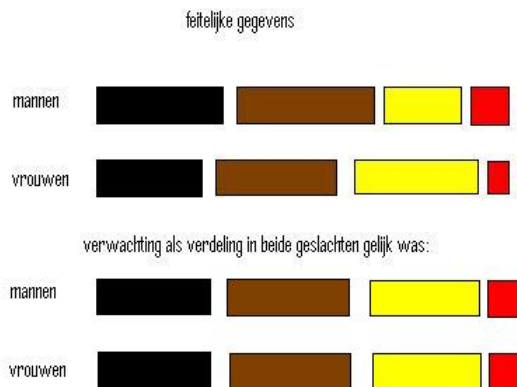
Het toetsen van onafhankelijkheid betekent dus eigenlijk: nagaan of *twee of meer* gevonden frequentieverdelingen overeenkomen *met dezelfde kansverdeling*.

**Voorbeeld: 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

$$e_{ij} = \frac{r_i \cdot k_j}{N} \quad (5.3)$$

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.

Hier is een tabel van de verwachte frequenties:

	zwart:	bruin:	blond:	rood:	totaal:
mannen:	$e_{11} = 29,0$	$e_{12} = 36,0$	$e_{13} = 26,7$	$e_{14} = 8,3$	$r_1 = 100$
vrouwen:	$e_{21} = 58,0$	$e_{22} = 72,0$	$e_{23} = 53,3$	$e_{24} = 16,7$	$r_2 = 200$
totaal:	$k_1 = 87$	$k_2 = 108$	$k_3 = 80$	$k_4 = 25$	$N = 300$

**5.5.1 De  $\chi^2$ -toets voor onafhankelijkheid:**

Ga als volgt te werk:

- Maak een *kruistabel* ('contingency table'), die naast de gegevens ook de rijtotaal, de kolomtotalen en de steekproefgrootte bevat.
- Bereken voor elke cel de verwachte frequentie op basis van de nulhypothese (5.3) van onafhankelijkheid.
- Bereken de  $\chi^2$ -waarde volgens (5.2)

$$X^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(o_{ij} - e_{ij})^2}{e_{ij}} .$$

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

$$P = \mathbb{P}_\nu[X^2 > \chi^2] .$$

Deze kunnen we berekenen we met de GR. Andere mogelijkheid: vergelijk de  $\chi^2$ -waarde met de kritieke waarde in Tabel 5.6, bepaald door  $\nu$  en de gekozen  $\alpha$ .

**Het voorbeeld:** Sekse en haarkleur: We zetten de gemeten frequenties en de verwachte frequenties bij elkaar in een tabel:

	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

Gevonden  $\chi^2$ -waarde:

$$\frac{(32 - 29,0)^2}{29,0} + \frac{(43 - 36,0)^2}{36,0} + \dots + \frac{(16 - 16,7)^2}{16,7} = 8,987 .$$

Aantal vrijheidsgraden  $\nu = (2 - 1) \cdot (4 - 1) = 3$ . De  $P$ -waarde is

$$P = \mathbb{P}[X^2 \geq 8,987] = 0,029 .$$

Dit is meer dan 5%. 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 Nog enkele voorbeelden

**Meniscusmuggen (*Dixa*) als indicatoren voor waterkwaliteit.**

	nebulosa	submaculata	dilatata	nubilipennis	total
oligotrophic	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

$\chi^2$ -waarde: 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 correlatie bestaat tussen:

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

**Zaadkieming afhankelijk van temperatuur en bodemtype.**

	clay soil	sandy soil	total:
5°C	40/57,97	100/82,03	140
25°C	131/113,03	142/159,97	273
total:	171	242	413

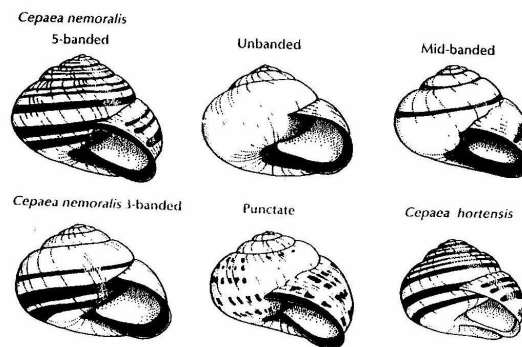
1 vrijheidsgraad, dus

$$X^2 = \frac{(40 - 57,97)^2}{57,97} + \dots = 14,4$$

en de kans op  $X^2 \geq 14,4$  is  $1,5 \cdot 10^{-4}$ .

Conclusie: de effecten van temperatuur en bodemtype zijn *niet* onafhankelijk van elkaar ( $P < 10^{-3}$  op een uitkomst van  $X^2$  minstens 14,4). Er is een positieve associatie tussen lage temperatuur en zandgrond en tussen hoge temperatuur en kleigrond.

### 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 door hun  $\chi^2$ -waarde te bepalen.

	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

$X^2 = \frac{(60-39,6)^2}{39,6} + \dots = 36,13$  bij twee vrijheidsgraden, dus dit is een zeer zeldzame uitkomst, met een kans  $1,4 \cdot 10^{-8}$  indien er tussen kleur en bandering geen verband bestaat.

Conclusie: kleur en bandering zijn met elkaar gecorreleerd.

### 5.5.3 Samenvatting

1. De  $\chi^2$ -waarde, dat is de som van  $\frac{(\text{gevonden} - \text{verwacht})^2}{\text{verwacht}}$  uit (5.2) geeft de *afwijking* weer tussen de gevonden frequenties en de te verwachten frequenties.
2. Alleen een *grote*  $\chi^2$ -waarde kan tot verwerping van de nulhypothese leiden. Er bestaat dus geen ‘tweezijdige  $\chi^2$ -toets’.
3. De  $\chi^2$ -verdelingen hebben alleen betrekking op aantallen (absolute frequenties) en *niet* op percentages, proporties of fracties.
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 moeten categorieën ‘gepooled’ worden: Hele rijen of kolommen in de kruistabel moeten worden samegevoegd, en hun frequenties opgeteld. Het aantal vrijheidsgraden wordt navenant verminderd.

## 5.6 Kritieke waarden voor de $\chi^2$ -verdelingen

Ten behoeve van toetsen van goodness-of-fit en onafhankelijkheid wordt bij  $\alpha = 0,05$  en  $\alpha = 0,01$  de waarde  $\chi_\alpha^2$  gegeven met de eigenschap dat  $\mathbb{P}[X^2 > \chi_\alpha^2] = \alpha$  voor een  $\chi^2$ -verdeelde stochast  $X^2$  met  $\nu$  vrijheidsgraden.

$\nu$	$\chi_{0,05}^2$	$\chi_{0,01}^2$	$\nu$	$\chi_{0,05}^2$	$\chi_{0,01}^2$	$\nu$	$\chi_{0,05}^2$	$\chi_{0,01}^2$
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



# Chapter 6

## Twee steekproeven

We hebben twee steekproeven met steekproefgrootten  $n_1$  en  $n_2$  uit twee populaties. De steekproefgemiddelden  $\bar{x}_1$  en  $\bar{x}_2$  verschillen van elkaar. Is dit verschil significant? Dat wil zeggen: liggen  $\bar{x}_1$  en  $\bar{x}_2$  voldoende uiteen om te concluderen dat  $\mu_1 \neq \mu_2$ ? Dit is het onderwerp van de paragrafen 6.1 tot en met 6.3. Een soortgelijke vraag is te stellen over de varianties: paragraaf 6.5.

### 6.1 Overlap van betrouwbaarheidsintervallen

Als de 95% betrouwbaarheidsintervallen voor  $\mu_1$  en  $\mu_2$  *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, zodat de nulhypothese vaak ten onrechte wordt gehandhaafd. 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
$t_{0,95} \cdot \text{s.e.}$	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 6.2.3

## 6.2 Ge-poolde varianties

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  $\mu_1$  en  $\mu_2$  en varianties  $\sigma_1^2$  en  $\sigma_2^2$ . Als de steekproeven groot genoeg zijn, dan geldt:*

*Het verschil tussen de steekproefgemiddelden,  $\Delta := \bar{X}_1 - \bar{X}_2$ , 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

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

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)	$\bar{Y}_D = 4,37$	$\bar{Y}_C = 4,12$
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 ( $\bar{Y}_D \neq \bar{Y}_C$ ) zijn het gevolg van toeval.

Volgens Stelling 3 is, onder de nulhypothese, de stochast

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

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

**Probleem:** de populatievarianties  $\sigma_D^2$  en  $\sigma_C^2$  zijn niet bekend!

### 6.2.1 Grote steekproefomvang: $z$ -toets met gepoolde varianties:

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

$$z \approx \frac{\bar{y}_D - \bar{y}_C}{\sqrt{(s_D^2/n_D) + (s_C^2/n_C)}} = 2,24 ,$$

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.

### 6.2.2 Kleine steekproefomvang: $t$ -toets met gepoolde varianties

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, **mits**  $n_1 \approx n_2$  **of**  $\sigma_1 \approx \sigma_2$ .

**Stelling 4. (Verskil van gemiddelden bij onbekende  $\sigma$ )** *Laten twee steekproeven van grootten  $n_1$  en  $n_2$  met gemiddelden  $\bar{x}_1$  en  $\bar{x}_2$  en steekproefvarianties  $s_1^2$  en  $s_2^2$  uit twee normaal verdeelde populaties met gemiddelden  $\mu_1$  en  $\mu_2$  en varianties  $\sigma_1^2$  en  $\sigma_2^2$  gegeven zijn. Dan is de  $t$ -waarde met gepoolde varianties*

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

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

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

### 6.2.3 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.)

Nullhypothese: het medicijn heeft geen effect:  $\mu_B = \mu_G$ .

We vinden de volgende  $t$ -waarde:

$$t_{\Delta} = \frac{\bar{x}_B - \bar{x}_G}{\sqrt{(s_B^2 + s_G^2)/n}} = \frac{8,75 - 9,63}{\sqrt{(1,02^2 + 1,38^2)/15}} = -1,986$$

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.

**Rattengroei opnieuw bekeken.** Nullhypothese:  $\mu_1 = \mu_2$ .

Omdat de steekproefgrootten gelijk zijn, kunnen we de  $t$ -toets met *gepoolede varianties* toepassen. De gevonden waarde voor  $t_{\Delta}$  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).

## 6.3 Eenzijdig toetsen

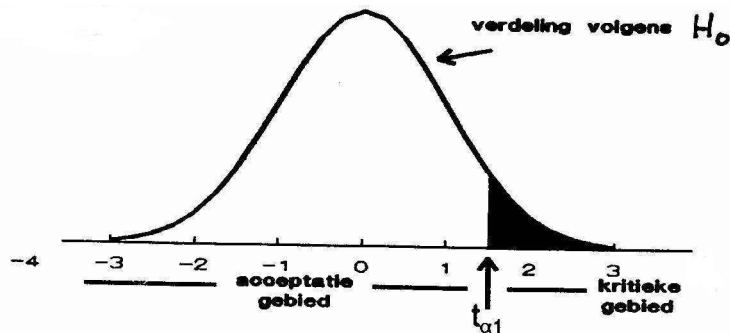
In bovenstaand voorbeeld (paragraaf 6.2.3) over het effect van een medicijn op de bloedstolling hebben we geen effect van het medicijn aangetoond, want de nulhypothese is niet verworpen.

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 gewijzigd: het gaat geheel naar één kant, maar wordt daar groter (zodat de totale verwerpingskans gelijk blijft). We verwerpen dus alleen als  $\bar{x}_B$  voldoende *kleiner* is dan  $\bar{x}_G$ , namelijk als het medicijn lijkt te werken als bedoeld. Dit is *eenzijdig* toetsen:

**Eenzijdig toetsen:** Een éézijdige toets houdt alleen rekening met éézijdige afwijkingen van de toetsgrootte, bijvoorbeeld de  $t$ -waarde. Het “kritieke gebied” ligt maar aan één kant en wordt begrensd door een kritieke waarde; in het geval van een  $t$ -toets is dit  $t_\alpha$ .

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 toetsgrootte in de van tevoren afgesproken richting van 0 afwijkt.

Als de toetsgrootte in de “juiste” richting van 0 afwijkt, dan wordt  $H_0$  bij een éézijdige toets gemakkelijker verworpen dan bij de analoge tweezijdige toets.

**Opnieuw het voorbeeld: Effect van medicijn op bloedstolling.**

- Nulhypothese: er is geen effect ( $\mu_B = \mu_G$ );
- Gevonden  $t$ -waarde:  $-1,986$ ;
- Kritieke  $t$ -waarde ( $\alpha = 0,05$ ,  $\nu = 28$ ):  $t_\alpha = -1,701$ .

**Conclusie:** Nulhypothese wèl verwerpen.

**Gevaar:** Het is verleidelijk om stelselmatig éézijdige toetsen te gebruiken, want bij deze wordt de nulhypothese sneller verworpen.

Maar een “significante” conclusie die getrokken wordt op basis van een éézijdige toets is volstrekt *onbetrouwbaar* als er niet bij voorbaat (*a priori*) duidelijke redenen zijn waarom afwijking maar in één richting zou kunnen optreden.

## 6.4 $t$ -toets voor gepaarde waarnemingen

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

- $n$  paren van waarnemingen leveren  $n$  verschillen  $d_i$ ;
- steekproefgemiddelde van het verschil:  $\bar{d}$ ;
- steekproef-standaarddeviatie van de verschillen:  $s_d$ ;
- Nulhypothese ( $H_0$ ):  
Het gevonden verschil in de steekproef  $\bar{d}$  berust alleen op toeval: het werkelijke gemiddelde verschil is nul ( $\mu_d = 0$ ).

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

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

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

**Conclusie:** We verwerpen de nulhypothese als de berekende  $t$ -waarde in absolute 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!

## 6.5 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 significantiedrempel  $\alpha$ . Gebruik van Tabel 6.9 leidt tot significantiedrempel  $\alpha = 0,1$ .
- 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 toetsgrootte onder  $H_0$  een kans  $\alpha$  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  $n_1, n_2 > 30$

## 6.6 Meer dan twee steekproeven: ANOVA!

Soms doet zich de situatie voor dat er steekproeven zijn genomen uit meerdere populaties met gemiddelden, zeg,  $\mu_1, \mu_2, \dots, \mu_k$  en varianties  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ , en dat we willen toetsen of er verschillen zijn tussen de  $\mu_i$ , of dat ze allemaal gelijk zijn.

Het zou dan misschien voor de hand liggen om de steekproeven paarsgewijs te gaan vergelijken door de methoden te gebruiken uit dit hoofdstuk (§6.2 en §6.3). **Dit is niet aan te raden!** Op de eerste plaats: Omdat er  $\frac{1}{2}n(n+1)$  paren steekproeven moeten worden vergeleken, is de kans dat per toeval een significant resultaat ontstaat niet  $\alpha$ , maar in de buurt van  $\frac{1}{2}n(n+1)\alpha$ , zodat de betrouwbaarheid slecht is. Wanneer je daarvoor zou corrigeren, wat wel kan, dan wordt de werkwijze is zeer inefficiënt: ze vreet meer proefnemingen dan noodzakelijk.

Er bestaat een veel betere methode om meerdere steekproeven te vergelijken, die ANOVA heet, een afkorting van 'ANalysis Of Varince', (Nederlands: variantie-analyse). Bij deze methode wordt gekeken naar de spreiding *in de gemiddelden* van de steekproeven. Als deze spreiding kan worden verklaard uit de spreiding van de individuen, die binnen elke steekproef apart kan worden gemeten, dan kunnen we aannemen dat de  $\mu_i$  allemaal gelijk zijn. Is ze groter, dan zijn de  $\mu_i$  verschillend. (En als ze kleiner is . . . . ., dan heeft iemand met de cijfers zitten knoeien!)

We zullen in dit college ANOVA niet behandelen, omdat de tijd tekort schiet. Bedenk echter dat zo'n methode bestaat, en in alle statistiek-pakketten is opgenomen, op het ogenblik dat je in bovenstaande situatie terechtkomt.

## 6.7 De F-verdeling

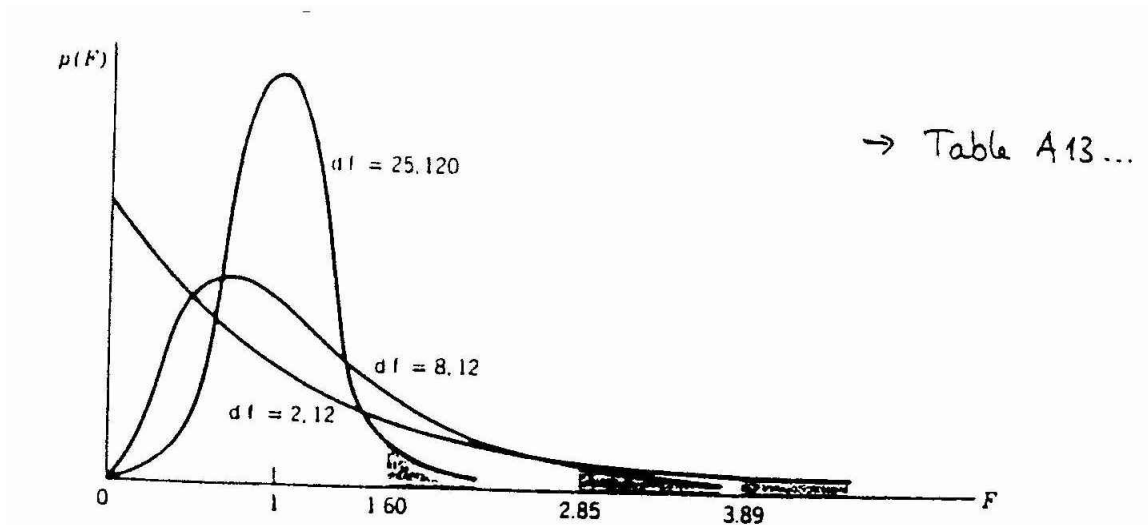


FIGURE 10-3

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  $a$ ,  $b$ ,  $\nu_1$  en  $\nu_2$  geeft de kans dat  $a \leq F \leq b$  voor een F-verdeelde stochast met  $\nu_1$  en  $\nu_2$  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  $\alpha$  aan, en het *betrouwbaarheidsniveau* met  $\gamma$ . Hierbij is  $\alpha$  een klein getal, en  $\gamma$  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

- éézijdig toetsen: verwerp  $H_0$  als  $Z \geq z_\alpha$ , waarbij  $z_\alpha$  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 gedefinieerd is dat  $\mathbb{P}[Z \geq z_{\alpha/2}] = \alpha/2$  (en dus  $\mathbb{P}[|Z| \geq z_{\alpha/2}] = \alpha$ ).

Bijvoorbeeld:

- $\mathbb{P}[Z \leq 1,96] = \Phi(1,96) = 0,975$ ; (Zie Tabel 4.5).
- Dus  $\mathbb{P}[Z \geq 1,96] = 1 - 0,975 = 0,025$ , dat wil zeggen  $z_{0,025} = 1,96$ .
- Symmetrie:  $\mathbb{P}[Z \leq -1,96] = \mathbb{P}[Z \geq 1,96] = 0,025$ ;
- dus  $\mathbb{P}[|Z| \geq 1,96] = \mathbb{P}[Z \geq 1,96] + \mathbb{P}[Z \leq -1,96] = 2 \times 0,025 = 0,05$ ;
- en  $\mathbb{P}[|Z| \leq 1,96] = 1 - 0,05 = 0,95$ .
- Dat wil zeggen:  $z_{95\%} = 1,96$  ( $= z_{0,025}$ !). Zie de vuistregel in paragraaf 4.3.4:  $z_\alpha = z_\gamma$  als  $\alpha$  dicht bij 0,  $\gamma$  dicht bij 1, en  $\gamma + 2\alpha = 1$ .

Bij de  $t$ -verdeling gaat het net zo: de waarde  $t_\alpha$  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 teller en  $\nu_2 = 10$  vrijheidsgraden in de noemer, 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  $\mathbb{P}[F \geq f] = 0,05$ . Het aantal vrijheidsgraden in de teller is  $\nu_1$ , dat in de noemer  $\nu_2$ .

$\downarrow \nu_2 \setminus \nu_1 \rightarrow$	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

# Chapter 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. Dit is de situatie van §6.4: een steekproef van *gepaarde waarnemingen*.
- 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

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) .$$

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 variabelen

We berekenen de steekproefvariantie in  $x$ -waarden en die in de  $y$ -waarden:

$$s_x^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2,$$
$$s_y^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2.$$

De (steekproef-)covariantie tussen  $x$  en  $y$  wordt nu gegeven door:

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

Je kunt deze overigens gemakkelijker berekenen met de formule:

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

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	

$$s_x^2 = 0,058, \quad s_y^2 = 51,62, \quad s_{xy} = \frac{1}{9}(95,53 - 10 \cdot 0,502 \cdot 16,74) = 1,277$$

$$r = \frac{1,277}{\sqrt{0,058 \cdot 51,62}} = 0,735$$

**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  $r_\alpha$  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  $+1$ .

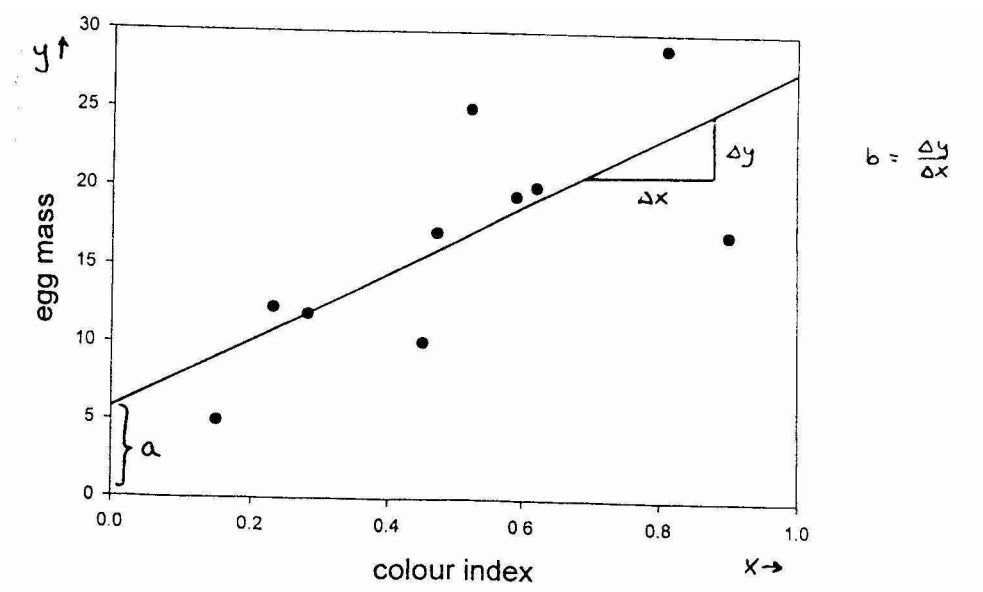
$r = 0$ :	er bestaat <i>geen</i> verband tussen $x$ en $y$ ;
$r > 0$ :	er bestaat een <i>positieve</i> relatie tussen $x$ en $y$ ;
$r < 0$ :	er bestaat een <i>negatieve</i> relatie tussen $x$ en $y$ ;
$r = +1$ :	er bestaat een <i>lineair positief</i> verband;
$r = -1$ :	er bestaat een <i>lineair negatief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

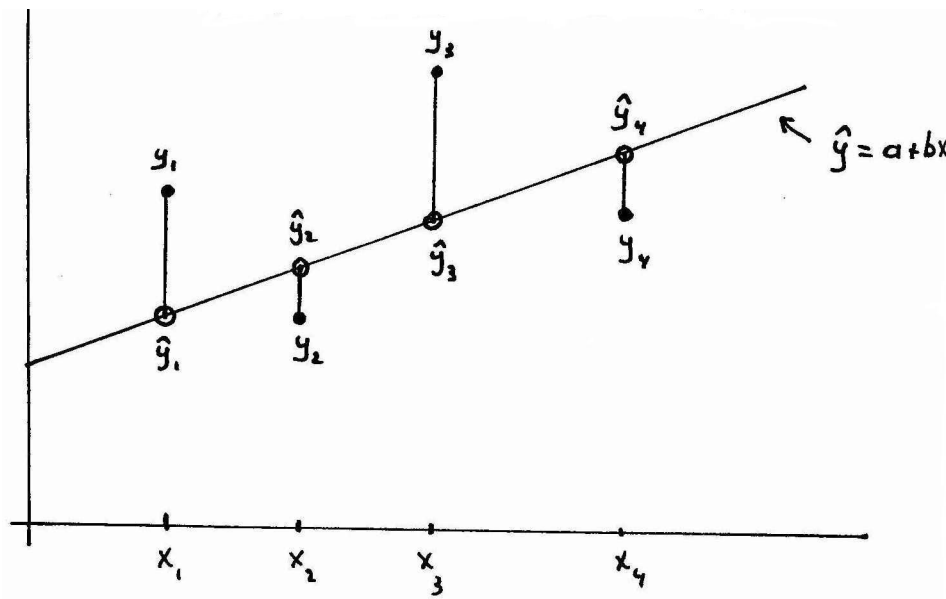
$$y_i - \hat{y}_i = y_i - (a + bx_i)$$

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

Resultaat:

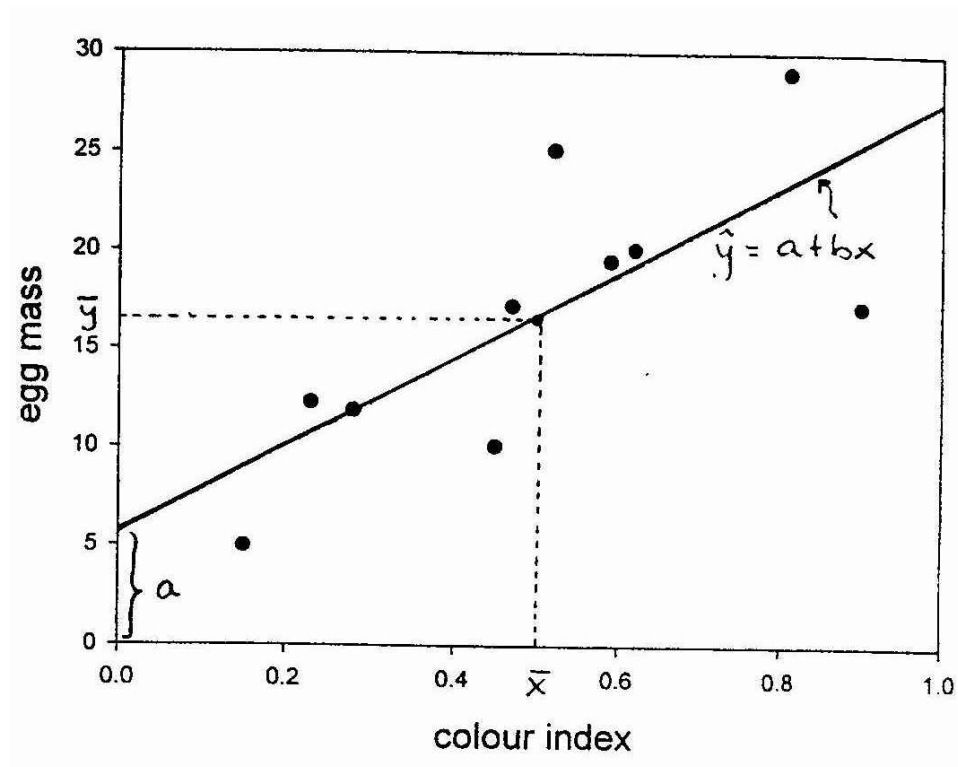
$$b = \frac{s_{xy}}{s_x^2} \quad \text{en} \quad a = \bar{y} - b\bar{x}$$

$b$  is de helling van de lijn, en  $a$  de ‘intercept’, de hoogte van het snijpunt met de  $y$ -as.



**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	



$$\bar{x} = 0,502, \bar{y} = 16,740, s_x^2 = 0,058, s_y^2 = 51,620, s_{xy} = 1,277$$

Dus

$$b = \frac{s_{xy}}{s_x^2} = 21,845$$

$$a = \bar{y} - b\bar{x} = 5,774$$

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 schatten.

**Eigenschappen:**

- De regressielijn gaat altijd door het punt  $(\bar{x}, \bar{y})$
- $b > 0$ : 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 in de *hele* populatie het *gemiddelde*  $\hat{y}$  van alle  $y$ -waarden bij gegeven  $x$  steeds gegeven wordt door  $\hat{y} = \alpha + \beta x$ .
- Door middel van een regressie-analyse probeer je dit verband te achterhalen, maar de regressiecoëfficiënten

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

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  $\alpha$  en  $\beta$ .
- De betrouwbaarheid van die schattingen kan gekwantificeerd worden: de standaardfout die we bij de schatting  $b$  van  $\beta$  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”

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

Dit is een maat voor de afwijking tussen datapunten en regressielijn

### 7.2.1 Toetsen of $\beta$ 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  $\beta$  valt.

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

Hier is  $t_{\alpha;\nu}$  de kritieke waarde bij  $\alpha$  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 de berekende toetsgrootheid groter is dan de kritieke waarde  $t_{\alpha;\nu}$ .

### 7.2.2 Voorbeeld: Kleur en broedsucces bij stekelbaarsjes

$$s_x^2 = 0,058, \quad s_y^2 = 51,620, \quad s_{xy} = 1,277$$

$$\Rightarrow b = 21,845, \quad a = 5,774$$

- residual variance  $s_e^2 = \frac{9}{8}(51,620 - 21,845^2 \cdot 0,058) = 26,686$
- standaardfout van de schatting van  $\beta$ :

$$s.e.\beta = \frac{s_e}{s_x \sqrt{n - 1}} = \frac{5,166}{0,24\sqrt{9}} = 7,122$$

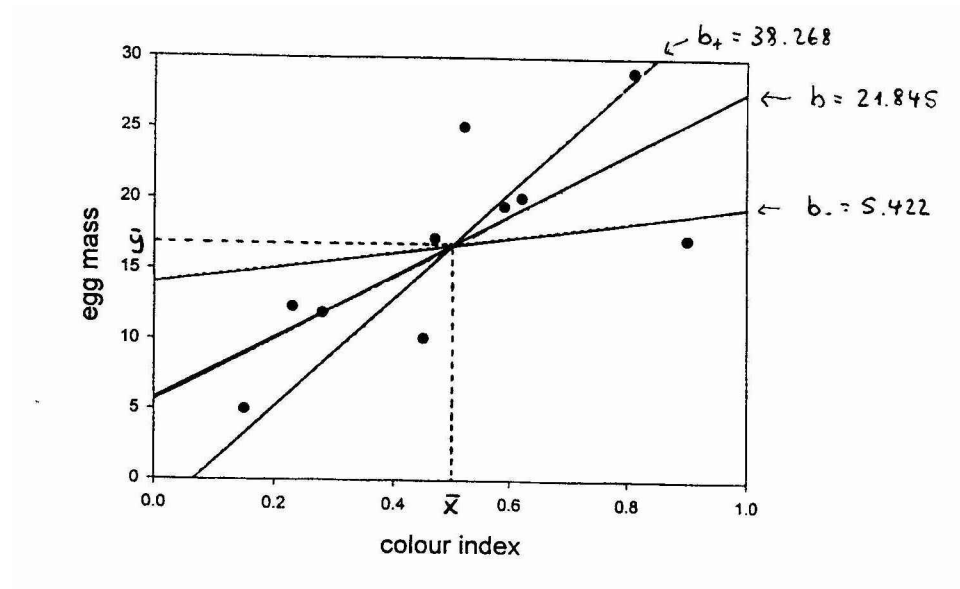
#### Betrouwbaarheidsinterval voor $\beta$

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  $\beta$  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*.

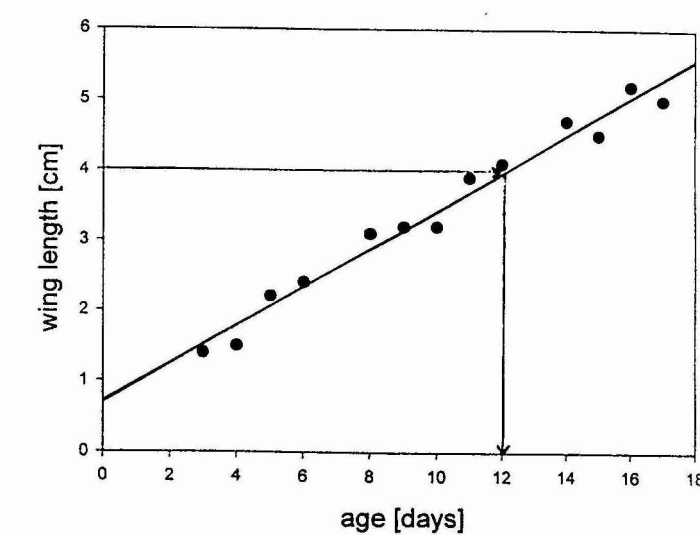
#### 7.3.1 Afhankelijke versus onafhankelijke variabele:

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

- 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

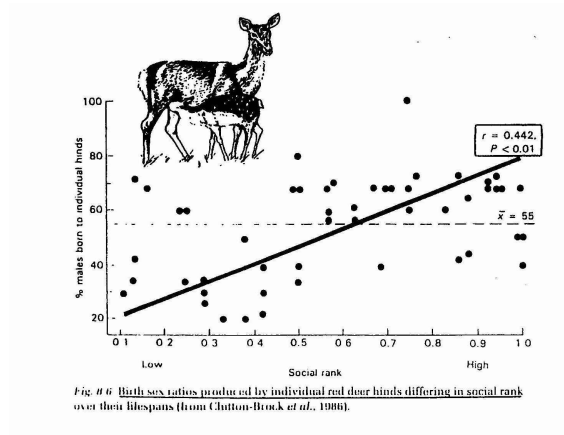


$$b = 0,270, a = 0,713, r = 0,987$$

$r = 0,987$  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).



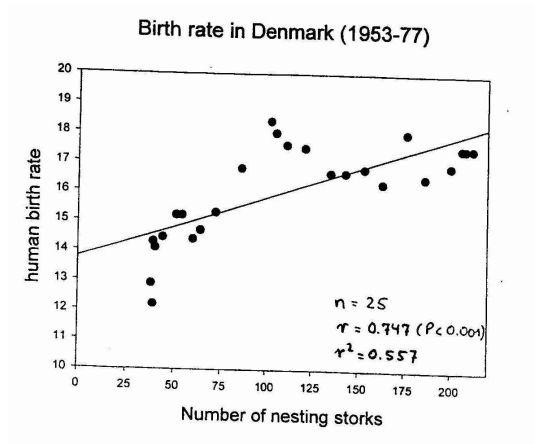
### 7.3.2 P-waarde en biologische relevantie

**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  $\nu = n-2$ .

$\alpha_1$	5%	2,5%	$\alpha_1$	5%	2,5%	$\alpha_1$	5%	2,5%
$\nu=1$	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