# 1 Kansen en waarschijnlijkheid

## 1.1 Kansvariabelen

1 Het aantal van 5 klassen is zo klein dat we voor een zo realistisch mogelijke uitwerking ook 'in' de klassen moeten kijken.

a Niet aan de vraag voldoen betekent dat de vraag groter is dan 40. De kans daarop kan worden gesplitst in 41,42 en groter dan 42. Groter dan 42, dus 43, 44, 45 enz valt in de klassen 45 en 50:

$$P(x > 42) = P(k = 45) + P(k = 50) = 0{,}25$$

De vraag 40 en 41 valt in de klasse 40, met vraag: 38, 39, 40, 41 en 42.

$$P(x = 41{,}42) = \frac{2}{5}P(k = 40) = 0{,}4\,0{,}30 = 0{,}12$$

De gevraagde kans op neeverkoop is dus totaal 0,37.

b,c Dit intikken in de rekenmachine. Denk er wel aan, dat er *kansen* staan, en geen aantallen. De machine denkt dat dit aantallen zijn, zodat de er totaal 'n=1' getallen in de tabel staan. Nu moet je $\sigma = \sigma_n$ nemen; bij indrukken van $\sigma_{n-1}$ krijg je een foutmelding of zoiets (delen door $1 - 1 = 0$).
Een ruwe schatting levert: $\mu = 38 + (5/30)5 = 38{,}7$ (mediaan) en $\sigma = (50 - 30)/4 = 5$ (bereik/4).

d Volg de suggesties in de opgave.

e De krantenverkoper wil graag weten hoeveel kranten hij moet inkopen om een maximale winst te halen. Daartoe zouden we voor ieder mogelijk inkoopaantal $x$ de verwachte winst kunnen bepalen, zoals in d is gedaan voor $x = 40$, en dan nagaan waar dat een maximum heeft. Die berekening is echter omslagtig.
Het geeft veel minder werk als we dit probleem anders aanpakken, en wel als volgt. We stellen ons de vraag met hoeveel de winst *toeneemt* als we 1 krant meer inkopen. (In wiskundige taal: wat is het diffentiequotiënt van de winst.) Zolang de winsttoename positief is kopen we een krant meer in, totdat de winsttoename negatief is. Dan is het maximum bereikt voor de winst.
De winsttoename is heel eenvoudig te vinden. Immers die ene extra krant kost 1,40 en levert 2,25 op mits de vraag maar meer dan x is. De *verwachtte winsttoename* is:

$$-1{,}40 + 2{,}25P(v > x)$$

Maximale winst wordt behaald als de kans op neeverkoop $P(v > x)$ juist $1{,}40/2{,}25 = 0{,}622$ is. We zien dat $P(v > 37) = 0{,}55$ net iets te

weinig is, maar $P(v > 32) = 0{,}90$ veel te veel. In eerste benadering kiezen we $x = 37{,}5 - 5(0{,}622 - 0{,}55)/0{,}35 = 37{,}2$ Merk op, dat dit aantal niet veel verschilt met de gemiddelde vraag: 38,75. Een goede strategie is dus: inspelen op de gemiddelde vraag.

# 2 Frequentieverdelingen en hun karakteristieke grootheden

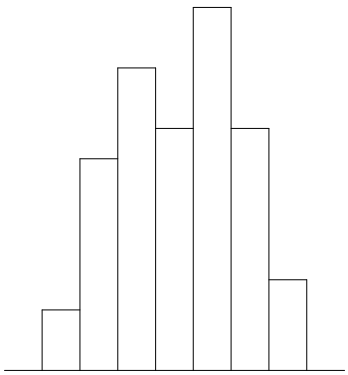
## 2.6 Uitwerkingen

### Frequentieverdeling

1 b De *spreidingsbreedte* $R = 73 - 43 = 30$. We starten met de *klassebreedte* $b = R/\sqrt{n} = 30/\sqrt{50} = 4{,}24$, afgerond tot $b = 5$. Dan wordt het *aantal klassen* in eerste instantie $k = R/b = 30/5 = 6$, waaruit we $k = 7$ halen (minimum aantal).

c Kies als (heeltallige) klassegrenzen 40, 45, ..., 70, 75. Dan is de eerste klasse vanaf 40 totenmet 44, ezovoort. De frequentietabel wordt daarmee:

| klasse | midden | aantal |
|---|---|---|
| 40–44 | 42 | 2 |
| 45–49 | 47 | 7 |
| 50–54 | 52 | 10 |
| 55–59 | 57 | 8 |
| 60–64 | 62 | 12 |
| 65–69 | 67 | 8 |
| 70–74 | 72 | 3 |



d Uit de frequentietabel blijkt dat de modus *mo* bij 62 ligt.

e Het gemiddelde $m$ van de *frequentieverdeling* is:

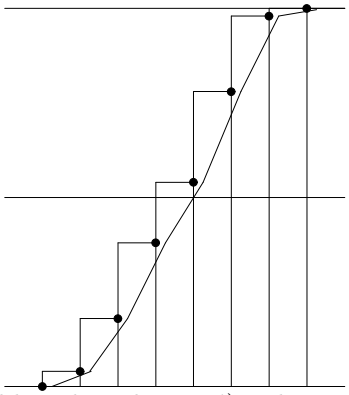
$$m = (2{\cdot}42 + 7{\cdot}47 + \cdots + 8{\cdot}67 + 3{\cdot}72)/50 = 57{,}7$$

De standaardafwijking $\sigma$ van de frequentieverdeling (van de populatie):

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{7} f_i{\cdot}x_i^2/n - m^2 \\ &= (2{\cdot}42^2 + \cdots + 3{\cdot}72^2)/50 - 57{,}7^2 = 7{,}94^2 \end{aligned}$$

Merk op, dat een verschuiving van de klasse met 0,5 geen invloed heeft op de indeling, maar wel op het gemiddelde en de standaardafwijking (vergelijk boek antwoorden).

f



De *cumulatieve* frequentietabel wordt:

| klasse | onder | aantal % |
|---|---|---|
| 40–44 | 45 | 4 |
| 45–49 | 50 | 18 |
| 50–54 | 55 | 38 |
| 55–59 | 60 | 54 |
| 60–64 | 65 | 78 |
| 65–69 | 70 | 94 |
| 70–74 | 75 | 100 |

a Uit de *ongegroepeerde* gegevens ('klassebreedte is 1') volgen iets nauwkeuriger ('andere') gemiddelden dan uit gegroepeerde gegevens (als in b tot f). Deze kunnen gemakkelijk met een rekenmachine in de statistische mode 'in één keer door' worden bepaald. Gemiddelde $\mu$ van de ongegroepeerde gegevens:

$$\mu = (45 + 60 + \cdots + 64 + 63)/50 = 57{,}8$$

Standaardafwijking $\sigma$ van de ongegroepeerde gegevens:

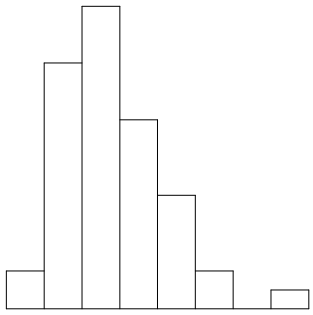
$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{50} x_i^2/50 - \mu^2 \\ &= \frac{45^2 + \cdots + 63^2}{50} - 57{,}8^2 = 7{,}57^2 \end{aligned}$$

2 De *op volgorde* gezette waarnemingen zijn:
34, 39,| 42, 42, 43, 44, 44, 44, 45, 45, 45, 45, 45, 48, 48, | 50, 50, 51, 51, 51, 52, 54, 55, 55, 56, 56, 56, 56, 57, 57, 58, | 60, 60, 60, 61, 62, 65, 67, 67, 68, 68, | 70, 70, 71, 75, 76, 79, | 80, 85, | | , 105.

b De *range* $R$ is de spreidingsbreedte: $105 - 34 = 71$
De *quasi-range* $Q$ is de spreidingsbreedte na wegname van de eerste en laatste $n/20 = 3$: $79 - 42 = 37$
De *kwartielendeviatie* $KR$ is de helft van de spreidingsbreedte na wegname van het eerste en laatste $n/4 = 12$: $(67 - 45)/2 = 11$

c Het *aantal klassen* $k$ is in eerste instantie: $k = R/\sqrt{n} = 71/7 = 10$.
De bijbehorende *klassebreedte* $b$ is $b = R/k = 71/10 = 7$. Omdat 7 een 'onhandig getal' is, nemen we liever klassebreedte $k = 10$, met $k = R/b = 71/10 = 8$ klassen (net 1 teveel voor 7 klassen).
De *klassegrenzen* kiezen we decimaal: 30, 40, 50, 60, 70, 80, 90, 100, 110.
De *klassemiddens* worden dan: 35, 45, 55, 65, 75, 85, 95, 105.

| klasse | midden | aantal |
|---|---|---|
| 30–39 | 35 | 2 |
| 40–49 | 45 | 13 |
| 50–59 | 55 | 16 |
| 60–69 | 65 | 10 |
| 70–79 | 75 | 6 |
| 80–89 | 85 | 2 |
| 90–99 | 95 | 0 |
| 100–109 | 105 | 1 |



d De *modus mo* is het midden van de grootste klasse: $mo = 55$. De *mediaan me* is de waarde van de middelste (nummer 25/26). Deze valt in de derde klasse (van 16 t/m 31), en is daar van de 16 nummer $25 - 16 + 1 = 10$. De waarde wordt geschat vanaf de linkergrens van de klasse: $me = 50 + (10{,}5/16)10 = 56{,}6$.
Het *gemiddelde* $m$ is het gewogen gemiddelde:

$$m = \frac{1}{50}(2{\cdot}35 + 13{\cdot}45 + \cdots + 1{\cdot}105) = 58{,}2$$

a Voor de *ongegroepeerde* waarnemingen geldt:
De *modus mo* is de meest voorkomende: 45 (5X).
De *mediaan me* is de middelste waarneming: 56 (nr 25–26).
Het *gemiddelde* $m$ is het rekenkundig gemiddelde: 56,9. De vuistregel voor $m$: $(3me - mo)/2$ geeft 61,5, wat nogal afwijkt.

e De met de klassen bepaalde of berekende waarden zijn minder nauwkeurig dan de waarden bepaald zonder groepering. De waarde van de modus wijkt daardoor nogal af (55 ipv 45); door een andere indeling kan de modus gemakkelijk verschuiven. Dankzij het middelen zijn de waarden van mediaan en gemiddelde echter niet zo veraf (56,6 ipv 56 en 58,2 ipv 56,9).

**Gemiddelde**

7 Minuten omzetten in uren door conversiefactor: 1uur/60minuten =1.

a
1. Den Haag–Utrecht: $v = 58/(35/60) = 99{,}4$
2. Utrecht–Eindhoven: $v = 72/(60/60) = 72{,}0$
3. Eindhoven–Venlo: $v = 45/(25/60) = 108{,}0$

b Gemiddelden:
- rekenkundig: $\overline{v} = (99{,}4 + 72{,}0 + 108{,}0)/3 = 93{,}1$
- meetkundig: $\overline{v} = \sqrt[3]{99{,}4{\cdot}72{,}0{\cdot}108{,}0} = 91{,}8$
- harmonisch: $\overline{v} = ((1/99{,}4 + 1/72{,}0 + 1/108{,}0)/3)^{-1} = 90{,}3$

c Gemiddelde (werkelijk):
$\overline{v} = (58 + 72 + 45)/(35 + 60 + 25)/60 = 87{,}5$

**Variantie**

11 a

| levensduur | cumulatief | levensduur | frequentie |
|---|---|---|---|
| $< 100$ | 1 | $000 < T < 100$ | 1 |
| $< 200$ | 4 | $100 < T < 200$ | 3 |
| $< 300$ | 12 | $200 < T < 300$ | 8 |
| $< 400$ | 22 | $300 < T < 400$ | 10 |
| $< 500$ | 41 | $400 < T < 500$ | 19 |
| $< 600$ | 58 | $500 < T < 600$ | 17 |
| $< 700$ | 77 | $600 < T < 700$ | 19 |
| $< 800$ | 88 | $700 < T < 800$ | 11 |
| $< 900$ | 97 | $800 < T < 900$ | 9 |
| $< 1000$ | 99 | $900 < T < 1000$ | 2 |
| $< 1100$ | 100 | $1000 < T < 1100$ | 1 |

b Voor het bepalen van de gemiddelde levensduur $m$ en standaardafwijking $s$ nemen we de klassemiddens. Allereerst het gemiddelde:

$$m = (1{\cdot}50 + 3{\cdot}150 + \cdots + 2{\cdot}950 + 1{\cdot}1050)/100 = 551$$

Voor het vinden van de standaardafwijking gebruiken we de uitdrukking:

$$s^2 = (\sum f_i x_i^2)/(n-1) - m^2 n/(n-1)$$

Gebruikmakend van de gevonden $m$:

$$s^2 = (1{\cdot}50^2 + 3{\cdot}150^2 + \cdots + 1{\cdot}1050^2)/99 - 100551^2/99 = 202{,}3^2$$

d De cumulatieve tabel bevat, doordat er totaal 100 lampen zijn gebruikt, precies de 'procent' aantallen (anders hadden we de aantallen nog moeten herschalen naar het totaal 100). Zet de cumulatieve tabel in een grafiek, waarbij de klassemiddens worden gebruikt. Trek een lijn die zo goed mogelijk bij de tabelpunten aansluit. Voilà.

13 De gegevens kunnen worden samengevat in de volgende tabel voor hetmaandsalaris $X$:

| klasse | frequentie |
|---|---|
| $1500 < X < 2000$ | 0,25 |
| $2000 < X < 2500$ | 0,35 |
| $2500 < X < 4000$ | 0,25 |
| $4000 < X < 5000$ | 0,15 |

a De *mediaan* $me$ van het maandsalaris ligt in de tweede klasse, omdat de eerste 25% bevat, en wel 25%, op de 35% totaal, vanaf de linkerrand: $me = 2000 + (2500 - 2000)(25/35) = 2357$. Het gemiddelde $m$ vinden we uit de met de frequenties gewogen klassemiddens:

$$m = 0{,}25{\cdot}1750 + 0{,}35{\cdot}2250 + 0{,}25{\cdot}3250 + 0{,}15{\cdot}4500 = 2713$$

b Voor de standaardafwijking maken we gebruik van de gewogen salarisafwijkingen $X - m$: $1750 - 2713 = -963 =$,
$2250 - 2713 = -463$, $3250 - 2713 = 537$, $4500 - 2713 = 1787$.
Volgt:

$$s^2 = 0{,}25\cdot(-963)^2 + 0{,}35\cdot(-463)^2 + 0{,}25\cdot 537^2 + 0{,}15\cdot 1787^2 = 926^2$$

**Superponeren**

4 Met steekproefgrootte $n = 100$ is $m = 12{,}4$ en $s = 2{,}5$. Een van de waarnemingen is geweest 14,1, maar moest zijn 11,4. Oude en nieuwe steekproef-gemiddelden voldoen aan:

$$\begin{aligned} 12{,}4 &= 14{,}1/100 + (\Sigma x_i)/100 \\ m &= 11{,}4/100 + (\Sigma x_i)/100) \end{aligned}$$

Door aftrekken van deze vergelijkingen vinden we de correctie:
$m - 12{,}4 = (11{,}4 - 14{,}1)/100 = 0{,}027$. Daaruit volgt het gemiddelde:
$m = 12{,}373$. Het gemiddelde wordt dus gecorrigeerd door de (gewogen) nieuwe waarde erbij te doen en de oude eraf te halen.
Voor het vinden van de steekproef-standaardafwijking gebruiken we de uitdrukking:

$$s^2 = (\sum x_i^2)/(n-1) - nm^2/(n-1)$$

In dit geval (gevonden $m$ invullen):

$$\begin{aligned} 2{,}5^2 &= 14{,}1^2/99 + (\sum x_i^2)/99 - 100\cdot 12{,}4^2/99 \\ s^2 &= 11{,}4^2/99 + (\sum x_i^2)/99 - 100\cdot 12{,}373^2/99 \end{aligned}$$

Door aftrekken blijkt de correctie:

$$s^2 - 2{,}5^2 = (11{,}4^2 - 14{,}1^2)/99 - 100(12{,}373^2 - 12{,}4^2)/99$$

waaruit voor de steekproef-standaardafwijking volgt:
$s^2 = 2{,}5^2 - 0{,}6955 + 0{,}6756 = 2{,}496^2$.
Merk op, dat beide correcties (nml van $x^2/(n-1)$ en van $m^2 n/(n-1)$) ongeveer evengroot maar tegengesteld zijn, zodat de standaardafwijking slechts weinig verandert. De verandering in $m$ is 0,22%, de verandering in $s$ is 0,16%.

16 a In de tweede steekproef: $n_2 = 8$, $m_2 = 35$, $s_2 = 12$, moet 20 worden vervangen door 44. Voor het juiste gemiddelde corrigeren we de oude:

$$m_2 = 35 + (44 - 20)/8 = 37{,}8$$

Voor de juiste standaardafwijking corrigeren we de oude variantie:

$$s_2^2 = 12^2 + (44^2 - 20^2)/7 - 8(37{,}8^2 - 35^2)/7 = 11{,}4^2$$

De laatste uitkomst is heel gevoelig voor het afronden van de gemiddelde waarde $m_2$.

b Gemiddelde wordt gewogen gemiddeld bij samenstellen:

$$m = \frac{12}{20}50 + \frac{8}{20}37{,}8 = 45{,}1$$

c Variantie wordt gewogen gemiddeld bij samenstellen:

$$s^2 = \frac{12}{20}s_1^2 + \frac{8}{20}11{,}4^2 = 12{,}7^2$$

Daaruit vinden we $s_1 = 13{,}5$. Ook nu is het antwoord heel gevoelig voor de waarde van de berekende $s_2$.

# 3 Grondbeginselen van de waarschijnlijkheidsrekening

## 3.9 Uitwerkingen.

**Relatieve frequentie kans**

1 Noem $p$ de kans om 'geen kruis' (geen $K = \overline{K}$ ) te gooien bij éénmaal opgooien; dan is $p = \frac{1}{2}$. Na een keer gooien hebben we 50 % kans op $\overline{K}$, dat is veel meer dan gevraagd. Gooien we twee keer op, dan is de kans om beide keren geen $K = \overline{K}$ te krijgen een produktkans van onafhankelijke gebeurtenissen:

$$P(\overline{K}_1 \cdot \overline{K}_2) = P(\overline{K}_1) \cdot P(\overline{K}_2) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}$$

De kans op tweemaal $\overline{K}$ is dus al kleiner, maar nog niet klein genoeg. Merkop, dat 12,5 % gelijk is aan 1/8. Dat is de kans bij driemaal achtereen 'geen kruis', vanwege de produktregel:

$$P(\overline{K}_1 \cdot \overline{K}_2 \cdot \overline{K}_3) = P(\overline{K}_1) \cdot P(\overline{K}_2) \cdot P(\overline{K}_3) = (\tfrac{1}{2})^3 = \tfrac{1}{8}$$

2 We gooien tegelijk 3 geldstukken op. Ieder geldstuk kan 'kruis' ($K$) of 'geen kruis' ($\overline{K}$) opleveren. Dat zijn $2^3 = 8$ mogelijke uitkomsten. Minstens eenmaal kruis is het tegengestelde van: 'geen enkel kruis'. Er is maar één manier om geen enkel kruis te gooien (alle drie $\overline{K}$), dus $8 - 1 = 7$ manieren om minstens eenmaal kruis te gooien. De kans daarop is dus 7/8.

3 Een dobbelsteen heeft 6 mogelijke uitkomsten bij het gooien. Hoeveel daarvan zijn '6' ∪ 'even'? Welnu, '6' ∈ 'even', dus blijft alleen de vraag naar het aantal 'even' uitkomsten. Dat zijn er 3, namelijk 'even' = {2,4,6}. De kans $P = 3/6 = 1/2$.

6 We gooien drie keer met een dobbelsteen—of, wat hetzelfde is, met drie gelijke dobbelstenen tegelijk—en tellen het aantal ogen. De kansvariabele $s = k_1 + k_2 + k_3$ is de som van de drie gegooide getallen $k$, die van 1 tot en met 6 kunnen lopen. Nu is volgens de complementregel:

$$P(s \geq 5) = 1 - P(s \leq 4)$$

Volgens de speciale optelregel (van elkaar uitsluitende gebeurtenissen):

$$P(s \leq 4) = P(s = 3) + P(s = 4)$$

Maar één van de $6^3$ mogelijke uitkomsten levert $s = 3$ op, namelijk $k_1 = k_2 = k_3 = 1$; dus $P(s = 3) = 1/216$. Er zijn 3 manieren om $s = 4$ te verkrijgen, namelijk de kombinaties
$s = 2 + 1 + 1 = 1 + 2 + 1 = 1 + 1 + 2$, met kans $P(s = 4) = 3/216$.
Conclusie: $P(s \geq 5) = 1 - (1/216) - (3/216)$.

8 Noem $Z = \{6\}$ 'een 6 gegooid', dan is $\overline{Z} = \{1,2,3,4,5\}$. De gevraagde gebeurtenis $A$ is: '3 maal zes in 4 worpen'. Dan is kennelijk één van de worpen 'niet zes'. De uitkomsten in $A$ zijn:

$$A = \{\overline{Z}ZZZ,Z\overline{Z}ZZ,ZZ\overline{Z}Z,ZZZ\overline{Z}\}$$

Het aantal elementaire uitkomsten in $A$ is $(5{\cdot}1{\cdot}1{\cdot}1)4 = 20$. Het totaal aantal mogelijke uitkomsten is $6^4 = 1296$, zodat de kans is $P(A) = \frac{20}{1296}$.

**Voorwaardelijke kans en productregel**

19 a A en B trekken om beurten een knikker uit de pot met 5 witte en 5 zwarte knikkers. Iedere trekking is afhankelijk van de vorige, omdat de knikker niet wordt teruggelegd. In dat geval gebruiken we de algemene productregel.
Noem W het trekken van een witte knikker, Z het trekken van een zwarte knikker. Een trekking is dan een reeks van Z afgesloten door een W. Bij welke trekkingen wint A? Bij trekkingen die starten met een *even* aantal Z (A en B trekken om beurten).
Omdat er maar 5 zwarte knikkers zijn, worden dat de trekkingen:
1. W, met $P(W) = \frac{5}{10}$.
2. ZZW, met $P(ZZW) = \frac{5}{10}\frac{4}{9}\frac{5}{8}$.
3. ZZZZW, met $P(ZZZZW) = \frac{5}{10}\frac{4}{9}\frac{3}{8}\frac{2}{7}\frac{5}{6}$.

$$P(\text{A wint}) = P(W) + P(ZZW) + P(ZZZZW)$$

20 Het kaartspel bestaat uit 13 harten (♡) en 39 andere kaarten, totaal 52. We trekken nu achtereenvolgens 4 kaarten, zonder terugleggging. Allemaal harten betekent volgens de algemene produktregel:

$$P(\heartsuit_1{\cdot}\heartsuit_2{\cdot}\heartsuit_3{\cdot}\heartsuit_4) = P(\heartsuit_1){\cdot}P(\heartsuit_2{\cdot}\heartsuit_3{\cdot}\heartsuit_4|\heartsuit_1)$$

met $P(\heartsuit_1) = 13/52$. Verdergaande op deze manier is:

$$P(\heartsuit_2{\cdot}\heartsuit_3{\cdot}\heartsuit_4|\heartsuit_1) = P(\heartsuit_2|\heartsuit_1){\cdot}P(\heartsuit_3{\cdot}\heartsuit_4|\heartsuit_1\heartsuit_2)$$

met $P(\heartsuit_2|\heartsuit_1) = (13-1)/(52-1) = 12/51$ omdat al een hartekaart is verdwenen. Totaal levert dat op:

$$P(\heartsuit_1{\cdot}\heartsuit_2{\cdot}\heartsuit_3{\cdot}\heartsuit_4) = \frac{13}{52}{\cdot}\frac{12}{51}{\cdot}\frac{11}{50}{\cdot}\frac{10}{49} = \frac{13{\cdot}12{\cdot}11{\cdot}10}{52{\cdot}51{\cdot}50{\cdot}49}$$

23 Noem $G$ 'de sleutel past', dan is $P(G) = 1/6$ en $P(\overline{G}) = 5/6$. Wanneer pas bij de achtste poging de sleutel past, dan is die gebeurtenis de volgende productgebeurtenis: $\overline{G}\,\overline{G}\,\overline{G}\,\overline{G}\,\overline{G}\,\overline{G}\,\overline{G}G$. De afzonderlijke pogingen zijn *onafhankelijk*, zodat de *speciale* productregel geldt:

$$P(\overline{G}\,\overline{G}\,\overline{G}\,\overline{G}\,\overline{G}\,\overline{G}\,\overline{G}\,\overline{G}G) = P(\overline{G})^7 P(G) = \left(\frac{5}{6}\right)^7 \frac{1}{6}$$

24 Deze opgave leent zich voor een 'kanstabel', waarin alle gegevens overzichtelijk zijn verwerkt (zie voorbeeld 3.15). Noem 'voldoet niet' $N$ en 'voldoet wel' $W$.

| $P$ | $A$ | $B$ | $C$ | subtotaal |
|---|---|---|---|---|
| $N$ | 0,03·0,20 | 0,05·0,30 | 0,10·0,50 | **0,071** |
| $W$ | | | | |
| subtotaal | 0,20 | 0,30 | 0,50 | 1,00 |

De uitgenomen schroef behoort tot de 'voldoet niet' klasse $N$. De (voorwaardelijke) kans dat de schroef uit C komt is:

$$P(C|N) = \frac{0{,}10 \cdot 0{,}50}{0{,}071} = \frac{50}{71}$$

**Somregel, complemenregel en productregel**

27 Noem 'een jongen' $J$ en 'een meisje' $M$. Een gezin van 2 kinderen kan dus als samenstelling hebben: $JJ$,$JM$,$MJ$,$MM$.

a $P(JJ) = 1/4$.

b 'In elk geval een jongen' is het complement van 'geen jongen', dus het complement van $MM$. De kans is $P(JJ|\overline{MM}) = 1/3$.

c In dit geval is er 'in elk geval een jongen', zoals bij de vorige vraag. Is de jongen de oudste, dan zijn de mogelijkheden $JJ$,$JM$ en in 1/2 gevallen heeft hij een broer. Is de jongen de jongste, dan resteren: $JJ$,$MJ$ en is de kans evenzo 1/2. De totale kans is 1/2: 1/2.1/2+1/2.1/2= 1/4+1/4=1/2.
Nemen we 'broer' letterlijk (de andere jongen is ouder), dan blijft slechts de tweede helft van die kans over: 1/4

28 $P(A) = 0{,}3$, $P(B) = 0{,}6$ en $P(A \cup B) = 0{,}7$.

a $A$ en $B$ zijn onafhankelijk als de produktregel geldt. We berekenen daartoe enerzijds de kans op het produkt:

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0{,}3 + 0{,}6 - 0{,}7 = 0{,}2$$

Anderzijds berekenen we het produkt van de kansen:

$$P(A) \cdot P(B) = 0{,}3 \cdot 0{,}6 = 0{,}18$$

In dit geval is $P(A \cup B) > P(A) \cdot P(B)$, zodat $A$ en $B$ positief afhankelijk zijn.

b Volgens de regel van voorwaardelijke kansen geldt:

$$P(B_2|A) = \frac{P(B_2 \cap A)}{P(A)} = \frac{P(A \cap B_2)}{0{,}3}$$

Wat weten we van $A \cap B_2$? Het is een deel van $A \cap B$, omdat $B = B_1 + B_2 + B_3$, dus ook $A \cap B = A \cap B_1 + A \cap B_2 + A \cap B_3$ en

$$P(A \cap B) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

Het linkerlid is reeds berekend, $P(A \cap B) = 0{,}2$. De laatste term in het rechterlid is leeg. Voor de eerste geldt dat de $A$ en $B_1$ onafhankelijk zijn, dus dat de speciale produktregel geldt $P(A \cap B_1) = P(A)P(B_1) = 0{,}3P(B_1)$. Omdat $B_1$ een deel is van $B$ volgt uit $P(B_1|B) = 0{,}4$ dat $P(B_1) = 0{,}4P(B) = 0{,}4{\cdot}0{,}6 = 0{,}24$. Daardoor wordt $P(A \cap B_1) = 0{,}3{\cdot}0{,}24 = 0{,}072$. Alles invullend in de vergelijking:

$$0{,}2 = 0{,}072 + P(A \cap B_2)$$

lossen we op $P(A \cap B_2) = 0{,}128$. Invullen in de bovenste vergelijking levert:

$$P(B_2|A) = \frac{0{,}128}{0{,}3} = 0{,}42$$

29 $A =$ 'er is een jongen'·'er is een meisje', en
$B =$ 'er is hoogstens één meisje' = 'geen meisje' $\vee$ 'één meisje',
dus: $A{\cdot}B =$ 'er is een jongen'·'één meisje' = 'één meisje' als er minstens twee kinderen zijn.
Merkop, dat $\overline{A} =$ 'geen jongen' $\cup$ 'geen meisje' twee mogelijke samenstellingen bevat.
Noem $J$ 'een jongen' en $M$ 'een meisje'.

a Bij een gezin met 3 kinderen zijn er $2^3 = 8$ elementaire samenstellingen. Omdat $\overline{A}$ er 2 heeft, zijn er in $A$ nog $8 - 2 = 6$ (complementregel); dus $P(A) = 6/8$. $B$ is verdeeld in 'geen meisje' (1 samenstelling) en 'één meisje' (3 samenstellingen, omdat het ene meisje de eerste, tweede of derde kan zijn). Totaal heeft $A$ $1 + 3 = 4$ samenstellingen, dus $P(B) = 4/8$. Voor $A{\cdot}B$ ('één meisje') zijn dat er 3, dus $P(A{\cdot}B) = 3/8$. Het product van de kansen $P(A)P(B) = 6/8{\cdot}4/8 = 3/8$ is gelijk aan de productkans $P(A{\cdot}B)$: $A$ en $B$ zijn onafhankelijk.

b Bij een gezin met 4 kinderen zijn er $2^4 = 16$ elementaire samenstellingen. Voor $A$ leidt dat tot een kans $P(A) = 1 - 2/16 = 7/8$. Voor $B$ wordt $P(B) = (1 + 4)/16 = 5/16$. Voor $A{\cdot}B$ is $P(A{\cdot}B) = 4/16 = 1/4$, verschillend van $P(A)P(B) = 7/8{\cdot}5/16$. Dus $A$ en $B$ zijn nu wel afhankelijk.

30 $P(A) = 1/2$ en $P(B) = 1/3$; $A$ en $B$ zijn onafhankelijk.

a Volgens de speciale productregel: $P(AB) = P(A)P(B) = 1/6$.

b Volgens de algemene somregel:$P(A \cup B) = P(A) + P(B) - P(AB) = 1/2 + 1/3 - 1/6$.

c De gevraagde gebeurtenis: 'hoogstens een van de gebeurtenissen' behandelen via het complement: 'geen van beide gebeurtenissen', dus $A \cap B$. $P = 1 - P(A \cdot B) = 1 - 1/6$.

# 4 Continue kansverdelingen

## 4.6 Uitwerkingen.

**Normale verdeling**

1 a De gestandaardizeerde variabele is $U = (X - \mu)/\sigma = (X - 100)/10$. Zet de grens voor $X$ om naar die voor $U$:

$$P(X > 107) = P((X - \mu)/\sigma > (107 - 100)/10) = P(U > 0{,}7)$$

Volgens tabel 1, pag. 277: $P(U > 0{,}7) = 0{,}2420$.

b $U = (X - \mu)/\sigma = (X - 80)/6$, dus $P(X < 78) = :$

$$P(U < (78 - 80)/6) = P(U < -0{,}333) = P(U > 0{,}333)$$

In de laatste stap is de *spiegelsymmetrie* van de normale verdeling rond het gemiddelde gebruikt. Het getal 0,333 komt niet in tabel 1 voor. We interpoleren tussen de waarden 0,33 en 0,34:
$P(U > 0{,}333) = 0{,}3707 - 0{,}3(0{,}3707 - 0{,}3669) = 0{,}3694$.

c $U = (X - \mu)/\sigma = (X - 115)/12$, dus

$$P(X < 130) = P(U < (130 - 115)/12) = P(U < 1{,}25)$$

Volgens tabel 1: $P(U > 1{,}25) = 0{,}1056$. Verder is volgens de *complementregel* $P(U < 1{,}25) = 1 - P(U > 1{,}25)$, zodat $P(X < 130) = 0{,}8944$.

d $U = (X - \mu)/\sigma = (X - 75)/5$, dus

$$P(X > 62) = P(U > (62 - 75)/5)$$
$$P(U > -2{,}6) = P(U < 2{,}6) = 1 - P(U > 2{,}6)$$

Volgens tabel 1: $P(U > 2{,}6) = 0{,}0047$, dus $P(X > 62) = 0{,}9953$.

e $U = (X - \mu)/\sigma = (X - 210)/16$, dus

$$P(225 < X < 240) = P\left(\frac{225 - 210}{16} < U < \frac{240 - 210}{16}\right)$$

De kans voor een *interval* is het verschil van de kansen voor de grenzen:

$$P(0{,}938 < U < 1{,}875) = P(U > 0{,}938) - P(U > 1{,}875)$$

Volgens tabel 1: $P(U > 0{,}938) = 0{,}1736 + 0{,}2 \cdot 0{,}0026 = 0{,}1741$, $P(U > 1{,}875) = 0{,}0304$, dus
$P(225 < X < 240) = 0{,}1741 - 0{,}0304 = 0{,}1437$.

f $U = (X - \mu)/\sigma = (X - 97{,}5)/4$, dus

$$P(93 < X < 95) = P\left(\frac{93 - 97{,}5}{4} < U < \frac{95 - 97{,}5}{4}\right)$$
$$P(-1{,}125 < U < -0{,}625) = P(U > 0{,}625) - P(U > 1{,}125)$$

In de laatste stap gebruiken we de $-/+$ symmetrie. Volgens tabel 1: $P(U > 0{,}625) = 0{,}2654$, $P(U > 1{,}125) = 0{,}1303$, dus $P(93 < X < 95) = 0{,}2654 - 0{,}1303 = 0{,}1351$.

g Met $\mu = 50$ en $\sigma = 8$ wordt de gevraagde kans:

$$P(45 < X < 65) = P\left(\frac{45 - 50}{8} < U < \frac{65 - 50}{8}\right)$$
$$P(-0{,}625 < U < 1{,}875) = P(U < 0{,}625) - P(U > 1{,}875)$$

Uit tabel 1: $P(U > 0{,}625) = 0{,}2654$, $P(U < 0{,}625) = 0{,}7346$, en $P(U > 1{,}875) = 0{,}0304$. Dus $P(45 < X < 65) = 0{,}7346 - 0{,}0304 = 0{,}7042$.

2 a We zetten eerst de kansen om naar gestandaardiseerde variabelen $U$. Beide kansen zijn $> 0{,}5$, welke niet in fde tabel voorkomen. Dus nemen we eerst complementen:

$$P(X > 85) = 0{,}9332 \quad P(X < 85) = 0{,}0668$$
$$P(X < 122{,}5) = 0{,}9878) \quad P(X > 122{,}5) = 0{,}0122$$

Uit tabel 1 zoeken we de gestandaardiseerde variabelen terug:

$$P(X < 85) = 0{,}0668 = P(U > 1{,}50) = P(U < -1{,}50)$$

(laatste stap is nodig om het kleiner teken te verkrijgen). Omdat $U = (X - \mu)/\sigma$ volgt:

$$P\left(\frac{X - \mu}{\sigma} < \frac{85 - \mu}{\sigma}\right) = P(U < -1{,}50)$$

waaruit de eerste grensvoorwaarde:

$$\frac{85 - \mu}{\sigma} = -1{,}50$$

Uit de andere grens halen we op die manier, via $0{,}0122 = P(U > 2{,}25)$:

$$P(X > 122{,}5) = P(U > 2{,}25) = P\left(U > \frac{122{,}5 - \mu}{\sigma}\right)$$

de tweede grensvoorwaarde :

$$\frac{122{,}5 - \mu}{\sigma} = 2{,}25$$

We elimineren het gemiddelde $mu$ door de eerste grensvoorwaarde van de tweede af te trekken; zodoende wordt:

$$\frac{37{,}5}{\sigma} = 3{,}75$$

Dus de standaard afwijking is $\sigma = 10$, en, dat invullend, het gemiddelde is $\mu = 100$.

b Gebruik de in a berekende $\mu = 100$ en $\sigma = 10$. Dan is

$$P(114{,}0 < X < 114{,}7) = P\left(\frac{114{,}0 - 100}{10} < U < \frac{114{,}7 - 100}{10}\right)$$
$$P(1{,}40 < U < 1{,}47) = P(U > 1{,}40) - P(U > 1{,}47)$$

Volgens tabel 1: $P(U > 1{,}40) = 0{,}0808$ en $P(U > 1{,}47) = 0{,}0708$, dus $P(114{,}0 < X < 114{,}7) = 0{,}0808 - 0{,}0708 = 0{,}0100$.

c We zoeken $x$ waarvoor geldt:

$$P(x < X < 95{,}3) = 19{\cdot}P(114{,}0 < X < 114{,}7)$$

Welnu, gebruik makend van de resultaten van a en b:

$$P\left(\frac{x - 100}{10} < U < \frac{95{,}3 - 100}{10}\right) = 19{\cdot}0{,}01$$
$$P(\frac{x - 100}{10} < U < -0{,}47) = 0{,}19$$

Voor het interval geldt
$P(\frac{x-100}{10} < U < -0{,}47) = P(U > \frac{x-100}{10}) - P(U > -0{,}47) = 0{,}19$.
Volgens tabel 1: $P(U > 0{,}47) = 0{,}3192$, dus $P(U > -0{,}47) = P(U < 0{,}47) = 1 - P(U > 0{,}47) = 1 - 0{,}3192 = 0{,}6818$. Volgt $P(U > \frac{x-100}{10}) = 0{,}19 + 0{,}6818 = 0{,}8718$. Via het complement, om getallen uit de tabel te krijgen:
$P(U < \frac{x-100}{10}) = P(U > -\frac{x-100}{10}) = 0{,}1282$. Terugzoeken levert:
$-\frac{x-100}{10} = 1{,}135$ en $x = 100 - 10{\cdot}1{,}135 = 88{,}6$.

3 Weigeren als $X > 1200$, waarbij het bruto gewicht $X$ normaal verdeeld is met $\mu = 1175$ en $\sigma = 15$ (alles in grammen). Kans op weigeren:

$$P(X > 1200) = P((X - \mu)/\sigma > (1200 - 1175)/15) = P(U > 1{,}667)$$

Volgens tabel 1: $P(U > 1{,}667) = 0{,}0485 + 0{,}67(-0{,}0010) = 0{,}0478$.

4 Bekend is uit het onderzoek: $P(X < 230) = 0{\cdot}034$. Verder zijn de gewichten $X$ normaald verdeeld met $\sigma = 10$. Dus:

$$P(\frac{X - \mu}{\sigma} < \frac{230 - \mu}{10} = P(U < \frac{230 - \mu}{10}) = 0{,}034$$

Uit tabel 1 zoeken we terug: $0{,}034 = P(U > 1{,}825) = P(U < -1{,}825)$. Daaruit concluderen we dat $(230 - \mu)/10 = -1{,}825$, dus
$\mu = 230 + 10{\cdot}1{,}825 = 248{,}3$.

5 De stochast 'lengte buisje' $X$ heeft een kansverdeling N($\mu = 75$, $\sigma = 1$).

a A neemt af 10·000 exemplaren met $73 < X < 75$. Nu is $P(73 < X < 75)$ na standaardisatie te berekenen:

$$P(73 < X < 75) = P(\frac{73-75}{1} < U < \frac{75-75}{1}) = P(-2 < U < 0)$$

Vanwege de symmetrie is die kans
$P(0 < U < 2) = P(U > 0) - P(U > 2) = 0{,}5 - 0{,}0228 = 0{,}4772$
volgens tabel 1. Bij een productie $n$ zijn dus $0{,}4772n = 10{\cdot}000$ goede exemplaren. De productie moet 20956 exemplaren beslaan (helaas veel afval!).

b B eist $75{,}7 < X < 77{,}3$, te halen uit de resterende 'afvalberg' van de productie voor A. Die partij is daarin nog geheel aanwezig. Als tevoren berekenen we eerst de fractie van totale partij die aan de eis van B voldoet:

$$P(75{,}7 < X < 77{,}3) = P(0{,}7 < U < 2{,}3) = P(U{>}0{,}7) - P(U{>}2{,}3)$$

Uit tabel 1: $P(U > 0{,}7) - P(U > 2{,}3) = 0{,}2420 - 0{,}0107 = 0{,}2313$.
Het aantal aan B te leveren wordt $0{,}2313{\cdot}20956 = 4847$.

6 De dikte $X$ is normaal verdeeld N($\mu = 69{,}8$, $\sigma = 0{,}4$). Er zijn 3 categorieën dikte:

- als $69{,}4 < X < 70{,}6$, dan is winst f 2,0.
- als $X > 70{,}6$, dan is winst f 1,5.
- als $X < 69{,}4$, dan is 'winst' f $-5{,}0$

De winstverwachting $w$ per plaatje kan met de kansen worden berekend:

$$w = 2{,}0{\cdot}P(69{,}4 < X < 70{,}6) + 1{,}5{\cdot}P(X > 70{,}6) + -5{,}0{\cdot}P(X < 69{,}4)$$

De twee grenzen worden na standaardizatie: $(69{,}4 - 69{,}8)/0{,}4 = -1$ en $(70{,}6 - 69{,}8)/0{,}4 = 2$. De kansen worden:
$P(X > 70{,}6) = P(U > 2) = 0{,}0228$,
$P(X < 69{,}4) = P(U < -1) = P(U > 1) = 0{,}1587$ en de resterende kans voor de eerste categorie
$P(69{,}4 < X < 70{,}6) = 1 - 0{,}1587 - 0{,}0228 = 0{,}8185$. Ingevuld:
$w = 2{,}0{\cdot}0{,}8185 + 1{,}5{\cdot}0{,}0228 - 5{,}0{\cdot}0{,}1587 = 0{,}8777$
Winst op een partij van 10000 plaatjes wordt dan f 8777.

9 Noem het gewicht $X$. Gevraagd: $P(65 < X < 75)$. De complicatie is, dat er twee soorten gewichten zijn: van de mannelijke en van de vrouwelijke studenten, die ieder een andere verdeling hebben. Dus maken we gebruik van voorwaardelijke kansen:

$$P(X) = P(X{\cdot}M) + P(X{\cdot}V) = P(X|M)P(M) + P(X|V)P(V)$$
$$P(X) = 0{,}9P(X|M) + 0{,}1P(X|V)$$

Voor de mannelijke kansen geldt: $\mu = 75$ en $\sigma = 15$, zodat

$$P(65 < X < 75|M) = P(-0{,}667 < U < 0) = P(U > 0) - P(U > 0{,}667)$$

Met tabel 1 vinden we $P(65 < X < 75|M) = 0{,}5 - 0{,}2525 = 0{,}2475$.
Voor de vrouwelijke kansen geldt: $\mu = 58$ en $\sigma = 10$, zodat

$$P(65 < X < 75|V) = P(0{,}7 < U < 1{,}7) = P(U > 0{,}7) - P(U > 1{,}7)$$

Met tabel 1 vinden we $P(65 < X < 75|V) = 0{,}2420 - 0{,}0446 = 0{,}1974$.
Tenslotte wordt de totale kans:

$$P(65 < X < 75) = 0{,}9 \cdot 0{,}2475 + 0{,}1 \cdot 0{,}1974 = 0{,}2425$$

**Exponentiële verdeling**

11 De gemiddelde tijd tussen twee gesprekken $\mu = 8/20 = 0{,}4$ uur. Gevraagd wordt de tijd $t$ te vinden, waarvoor de kans op een gesprek na die tijdsduur gelijk wordt aan $1/4$. Die kans hangt exponentiëel af van $t$, met de parameter $\lambda = 1/\mu$. Dus voor de stochastische tijdsduur $T$ waarin nog niet is gebeld geldt:

$$P(T < t) = 1 - e^{-t/\mu} = 0{,}25$$

waaruit volgt $\exp(-t/0{,}4) = 0{,}75$, of $t = 0{,}4 \ln(1/0{,}75) = 0{,}115$. Bel niet langer dan 0,115 uur $= 6{,}9$ minuten

12 De gemiddelde tijd tussen waarnemingen is $\mu = 30$ minuten. Gevraagd wordt de tijd $t$ te bepalen waarop de kans op de volgende waarneming $T$ 0,10 is geworden. De tijdsduur $T$ waarin nog geen volgende waarneming valt volgt een exponentiële verdeling met parameter $\lambda = 1/\mu$, zodat de kans 0,10 wordt als:

$$P(T < t) = 1 - e^{-t/\mu} = 0{,}10$$

Daaruit volgt $\exp(-t/30) = 0{,}90$, of $t = 30 \ln(1/0{,}9) = 3{,}2$ minuten.

# 5 Lineaire regressie en correlatierekening

**Correlatie en regressielijn**

Voor het bepalen van de regressielijn, of de correlatie tussen $X$ en $Y$, hebben we allereerst nodig de steekproef-gemiddelden $m$ en steekproef-standaard-afwijkingen $s$ van $X$ en $Y$ apart. Met de rekenmachine in de statistische mode vinden we na invoeren van alle $X$ het steekproef-gemiddelde als $m$ of $\overline{x}$ en de steekproef-standaard-afwijking als $s$ of $\sigma_{N-1}$. Overigens kunnen heel goede rekenmachines ook getallenparen behandelen en rechtstreeks de correlatie bepalen.
Daarnaast bepalen we het steekproef-gemiddelde van het product $m_{xy} = \overline{XY}$. Dat kan met de gewone rekenmachine door herhaald optellen van vermenigvuldigingen, gevolgd door delen door het aantal waarnemingsparen $n$.
Met behulp van $m_x = \overline{X}$, $s_x$, $m_y = \overline{Y}$, $s_y$ en $m_{xy} = \overline{XY}$ zijn nu alle vragen over de regressielijn of de correlatie te beantwoorden. Allereerst geldt voor de steekproef-covariantie tussen $X$ en $Y$:

$$\text{cov}(X,Y) = \frac{n}{n-1}(\overline{XY} - \overline{X}\cdot\overline{Y}) = \frac{n}{n-1}(m_{xy} - m_x \cdot m_y)$$

Daaruit volgt de steekproef-correlatie $r = r(X,Y)$ tussen $X$ en $Y$:

$$r = \frac{\text{cov}(X,Y)}{s_x \cdot s_y}$$

Uitgedrukt in gestandaardiseerde waarden wordt de vergelijking van de regressielijn (van de eerste soort):

$$\frac{y - m_y}{s_y} = r\frac{x - m_x}{s_x}$$

Uitgedrukt in de waarden $x$ en $y$ zelf wordt de regressielijn:

$$y = (m_y - a m_x) + ax \quad a = \frac{s_y}{s_x} r$$

met richtingscoëfficiënt $a$ bepaald door de correlatie.
Voor de volledigheid geven we ook de residuele steekproef-standaardafwijking $s_r$, de (verticale) afwijking van $Y$ vanaf de regressielijnwaarde $y$ (van de eerste soort):

$$s_r^2 = \frac{n-1}{n-2}\left(1 - r^2(X,Y)\right) s_y^2$$

## 5.8 Uitwerkingen

3 Voor de reeks waarnemingen van $X$ apart vinden we:

$$m_x = 0{,}55 \quad s_x = 0{,}30$$

Voor de reeks waarnemingen van $Y$ apart vinden we:

$$m_y = 15{,}5 \quad s_y = 6{,}55$$

Voor de reeks waarnemingen van $X$ en $Y$ samen vinden we:

$$m_{xy} = (0{,}9{\cdot}17 + 0{,}2{\cdot}4 + \cdots + 0{,}5{\cdot}23)/10 = 8{,}67$$

c In dit geval wordt de steekproef-covariantie:

$$\mathrm{cov}(X,Y) = (10/9)(8{,}67 - 0{,}55{\cdot}15{,}5) = 0{,}161$$

In dit geval wordt de steekproef-correlatie:

$$r = \frac{0{,}161}{0{,}30{\cdot}6{,}55} = 0{,}082$$

Dit is een verwaarloosbare correlatie (minder dan 10% gecorreleerd).

a De coëfficiënten $a$ en $b$ van de regressielijn (van de eerste soort) volgen uit de correlatie en de standaarafwijkingen. De richtingscoëfficiënt $a$ is:

$$a = \frac{6{,}55}{0{,}30} 0{,}082 = 1{,}79$$

Het snijpunt met de y-as heeft de hoogte $b$:

$$b = 15{,}5 - 1{,}79{\cdot}0{,}55 = 14{,}5$$

5 Voor de reeks waarnemingen van de diameter $D$ apart vinden we:

$$m_d = 13{,}73 \quad s_d = 1{,}764$$

Voor de reeks waarnemingen van de zaagtijd $Z$ apart vinden we:

$$m_z = 23{,}82 \quad s_z = 3{,}301$$

Voor de reeks waarnemingen van $D$ en $Z$ samen vinden we:

$$m_{dz} = (12{,}1{\cdot}20{,}8 + 12{,}7{\cdot}21{,}2 + \cdots + 15{,}8{\cdot}27{,}8)/9 = 332{,}3$$

b In dit geval wordt de steekproef-covariantie:

$$\mathrm{cov}(D,Z) = (9/8)(332{,}3 - 13{,}73{\cdot}23{,}82) = 5{,}80$$

In dit geval wordt de steekproef-correlatie:

$$r = \frac{5{,}80}{1{,}764{\cdot}3{,}301} = 0{,}996$$

Dit is een perfecte correlatie (minder dan 1% niet gecorreleerd).

d De $10^{\text{e}}$ meting geeftvoor een diameter 11,7 mm een zaagtijd van 21,4 s op. Op grond van de 9 metingen kunnen we voorspellen wat de zaagtijd zou moeten zijn: de waarde op de regressielijn.
De regressielijn (van de eerste soort) heeft de richtingscoëfficiënt $a$:

$$a = \frac{s_z}{s_d} r = \frac{3{,}301}{1{,}764} 0{,}996 = 1{,}864$$

De gemiddelde zaagtijd $Z$ bij de diameter 11,7 moet dus zijn:

$$Z = m_z + a(D - m_d) = 23{,}82 + 1{,}864(11{,}7 - 13{,}73) = 20{,}03$$

De residuele standaard-afwijking $s_r$ hiervan is:

$$s_r^2 = \frac{9-1}{9-2}(1 - 0{,}996^2)3{,}301^2 = 0{,}315^2$$

We vinden echter een residuele-afwijking $21{,}4 - 20{,}03 = 1{,}3$, welke wel 4 standaardafwijkingen is. Dat laatste is bijkans onmogelijk. Denk bijvoorbeeld aan een normale verdeling, waarbij meer dan 3 standaardafwijkingen al een verwaarloosbare kans heeft. We concluderen dat er een meetfout gemaakt moet zijn (of een notatiefout?).

7 De bacteriegroei blijkt exponentiëel te zijn: $y = a \cdot b^x$, dus niet lineair. De theorie van de regressie*lijn* is gebaseerd op een lineaire groei, dus kan niet rechtstreeks worden gebruikt. Met een truc is toch lineariteit te verkrijgen: door naar de logarithmen te kijken in plaats van naar de waarden. Immers, wegens de ln-eigenschappen, is ln evenredig met $x$:

$$\ln(y) = \ln(a) + x \ln(b)$$

De stochastische variabele die we nodig hebben naast $X$ is de logarithmische stochast $Z$:

$$Z = \ln(Y)$$

De waarnemingen worden uitgedrukt in $X$ en $Z$:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| $z$ | 3,40 | 3,81 | 4,14 | 4,51 | 4,88 | 5,25 | 5,63 |

Voor de waarnemingen $X$ apart:

$$m_x = 3{,}00 \quad s_x = 2{,}160$$

Voor de waarnemingen $Z$ apart:

$$m_z = 4{,}518 \quad s_z = 0{,}796$$

Voor het product $XZ$ samen:

$$m_{xz} = 15{,}026$$

Daaruit volgen de steekproef-covariantie

$$\mathrm{cov}(X,Z) = (7/6)(15{,}026 - 3{,}00{\cdot}4{,}518) = 1{,}718$$

en de correlatie

$$r = \frac{1{,}718}{2{,}16{\cdot}0{,}796} = 0{,}9992$$

De correlatie is bijkans perfect! Voor de richtingscoëfficiënt $a$ van de regressielijn $z(x)$ (niet te verwarren met de $a$ in de exponentiële vergelijking) volgt

$$a = \frac{0{,}796}{2{,}16} 0{,}9992 = 0{,}3682$$

De regressielijn heeft de vergelijking:

$$z = 4{,}518 + 0{,}3682(x - 3{,}00) = 3{,}414 + 0{,}3682x$$

Daaruit volgen de meest waarschijnlijke waarden $\ln(a) = 3{,}413$ en $\ln(b) = 0{,}368$, oftewel $a = 30{,}37$ en $b = 1{,}445$. De grote nauwkeurigheid van deze constanten is het gevolg van de uiterst hoge correlatie. Merk op, dat de residuele steekproef-standaard-afwijking $s_r$ een waarde heeft:

$$s_r^2 = \frac{6}{5}(1 - 0{,}9992^2)0{,}796^2 = 0{,}035^2$$

zodat de constanten in de regressievergelijking een nauwkeurigheid in de orde van $0{,}035/\sqrt{7} = 0{,}01$ krijgen.

# 6 Het optellen en aftrekken van stochastische variabelen

## 6.7 Uitwerkingen

**Som en verschil van onafhankelijke stochasten**

1 Laat de stochast $X$ de bedieningstijd van machine A zijn, en $Y$ die van machine B. De gemiddelden en standaard-afwijking van de variabelen zijn, uitgedrukt in seconden:

$$\mu_x = 200, \quad \sigma_x = 24 \qquad \mu_y = 324, \quad \sigma_y = 32$$

Laat $Z$ de totale bedieningstijd van beide machines zijn, dan is $Z$ de som van de twee variabelen:

$$Z = X + Y$$

Voor het gemiddelde van $Z$ geldt *altijd* de someigenschap:

$$\mu_z = \mu_x + \mu_y = 200 + 324 = 524$$

Voor de standaardafwijking van *onafhankelijke variabelen* geldt de somregel (voor varianties):

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 = 24^2 + 32^2 = 40^2$$

We nemen natuurlijk aan dat de variabelen $X$ en $Y$ onafhankelijk zijn. Gevraagd wordt de kans op een bedieningstijd langer dan 10 minuten (600 seconden):

$$P(Z > 600) = P(U_z > \frac{600 - 524}{40}) = P(U_z > 1{,}9)$$

waarbij de variabele is gestandaardiseerd. Is $U_z$ een normaal verdeelde stochast? Jawel, want als $X$ en $Y$ *beiden normaal verdeeld* zijn, dan is de som $Z$ dat ook. Verder is de gestandaardiseerde normaal verdeelde stochast zelf ook normaal verdeeld. Volgens tabel 1 is dan de gevraagde kans 0,0287, dus bijna 3 procent.

2 Wanneer zal het ronde pennetje niet in het cilindervormige buisje passen? Als de diameter van het pennetje $X$ ($\mu_x = 10{,}2$, $\sigma_x = 0{,}375$) groter is dan de diameter van het buisje $Y$ ($\mu_y = 11{,}2$, $\sigma_y = 0{,}500$). De gevraagde kans dat pennetje en buisje weer in het aanvoercircuit

worden opgenomen is daarom te schrijven als
$P(X > Y) = P(X - Y > 0)$. Laat het verschil van de diameters $Z$ zijn:

$$Z = X - Y$$

Voor het gemiddelde van $Z$ geldt *altijd* de 'som'eigenschap (met een $-$-teken):

$$\mu_z = \mu_x - \mu_y = 10{,}2 - 11{,}2 = -1{,}0$$

Voor de standaardafwijking van *onafhankelijke variabelen* geldt de 'som'regel (met een $+$-teken vanwege het kwadrateren):

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 = 0{,}375^2 + 0{,}50^2 = 0{,}625^2$$

We nemen natuurlijk aan dat de variabelen $X$ en $Y$ onafhankelijk zijn. De gevraagde kans kan nu worden berekend:

$$P(Z > 0) = P(U_z > \frac{0 - (-1{,}0)}{0{,}625}) = P(U_z > 1{,}6)$$

Ook het verschil van twee normaal verdeelde stochasten is normaal verdeeld. Volgens tabel 1 wordt $P(U_z > 1{,}6) = 0{,}0548$. Bij ruim 5% passingen zullen pennetje en buisje niet passen.

3 De verschillende deeltijden van het estafetteteam 4x400m zijn de stochasten $A$, $B$, $C$ en $D$. De gemiddelden en standaardafwijkingen zijn achtereenvolgens:

$$\mu_A = 48{,}4 \quad \mu_B = 50{,}6 \quad \mu_C = 49{,}8 \quad \mu_D = 51{,}2$$
$$\sigma_A = 1{,}2 \quad \sigma_B = 2{,}1 \quad \sigma_C = 2{,}4 \quad \sigma_D = 2{,}8$$

a Een recordpoging slaagt als de totaaltijd $X = A + B + C + D < 190{,}5$. Het gemiddelde van de totaaltijd is volgens de somregel de som van de deeltijden:

$$\mu_x = 48{,}4 + 50{,}6 + 49{,}8 + 51{,}2 = 200{,}0$$

De standaardafwijking van de totaaltijd volgt uit de somregel voor varianties, mits de deeltijden *onafhankelijk* zijn:

$$\sigma_x^2 = 1{,}2^2 + 2{,}1^2 + 2{,}4^2 + 2{,}8^2 = 4{,}410^2$$

De kans op een record is:

$$P(X < 190{,}5) = P(U_x < \frac{190{,}5 - 200{,}0}{4{,}41}) = P(U_x < -2{,}15)$$

Nemen we aan dat de deeltijden normaal verdeeld zijn, dan is ook de totaaltijd normaal verdeeld. Dan geldt
$P(U_x < -2{,}15) = P(U_x > 2{,}15) = 0{,}0158$. De kans is nog geen 2%.

b Als B een betere tijd dan A maakt, dan is $B < A$, dus $B - A < 0$. De verschilstochast $Y = B - A$ heeft gemiddelde

$$\mu_y = \mu_B - \mu_A = 50{,}6 - 48{,}4 = 2{,}2$$

Om de standaardafwijking van de verschilstochast te kunnen bepalen veronderstellen we dat $A$ en $B$ onafhankelijk zijn, zodat de somregel voor varianties geldt:

$$\sigma_y^2 == \sigma_B^2 + \sigma_A^2 = 2{,}1^2 + 1{,}2^2 = 2{,}42^2$$

De kans dat B sneller is dan A is:

$$P(Y < 0) = P(U_y < \frac{0 - 2{,}2}{2{,}42}) = P(U_y < -0{,}910)$$

Zijn A en B normaal verdeeld, dan is ook het verschil $Y$ normaal verdeeld zodat $P(U_y < -0{,}910) = P(U_y > 0{,}910) = 0{,}1814$ volgens tabel 1.

c Wil D sneller zijn dan de drie anderen, dan moet aan drie condities vodaan zijn $(D < A)\cdot(D < B)\cdot(D < C)$. Helaas: ook al zijn de verschillende tijden onafhankelijk, die condities zijn positief *afhankelijk*. Volgens de produkteigenschap van positief afhankelijke gebeurtenissen is:

$$P((D < A)\cdot(D < B)\cdot(D < C)) \geq P(D < A)\cdot P(D < B)\cdot P(D < C)$$

Het bepalen van de afzonderlijke kansen is analoog aan het bepalen van de kans dat B sneller is dan A, met dien verstande dat er nu drie verschilstochasten zijn, die ieder een eigen gemiddelde en standaardafwijking hebben. Controleer dat:

| $Y$ | $\mu_y$ | $\sigma_y$ | $U_y$ | $P(Y < 0)$ |
|---|---|---|---|---|
| $D - A$ | 2,8 | 3,05 | $-0{,}919$ | 0,1791 |
| $D - A$ | 0,6 | 3,50 | $-0{,}171$ | 0,4321 |
| $D - A$ | 1,4 | 3,13 | $-0{,}447$ | 0,3275 |

De kans dat D sneller is dan de anderen wordt dus minstens $0{,}1791\cdot0{,}4321\cdot0{,}3275 = 0{,}025$, zeg een kleine kans van 3%.

**Som en verschil van afhankelijke stochasten**

5 Uit de gegevens kunnen we allereerst de standaardafwijking van de variabele $X = W - S$, het verschil tussen wiskunde en statistiek, bepalen. Gegeven is immers dat $P(W > S + 1) = 0{,}33$, dus $P(X > 1) = 0{,}33$. Het gemiddelde van $X$ is

$$\mu_x = \mu_w - \mu_s = 6{,}35 - 5{,}90 = 0{,}45$$

Het gegeven is nu te omschrijven in de gestandaardiseerde stochast:

$$P(U_x > \frac{1 - 0{,}45}{\sigma_x}) = P(U_x > \frac{0{,}55}{\sigma_x}) = 0{,}33$$

Nemen we aan dat $X$ normaal verdeeld is (bijvoorbeeld omdat $W$ en $S$ normaal verdeeld zijn), dan kunnen we terugzoeken in tabel 1 dat $0{,}55/\sigma_x = 0{,}44$, waaruit de standaardafwijking volgt $\sigma_x = \sigma_{w-s} = 1{,}25$. De standaardafwijking van een som of verschil van (afhankelijke) variabelen kan in het algemeen worden bepaald uit de 'som'regel voor de varianties:

$$\sigma^2_{w-s} = \sigma^2_w + \sigma^2_s - 2\,\mathrm{cov}(W,S)$$

(let op het minteken bij de covariantie, maar het plusteken bij de varianties). Ingevuld: $1{,}25^2 = 0{,}9^2 + 1{,}2^2 - 2\,\mathrm{cov}(W,S)$, waaruit volgt dat de covariantie cov tussen $W$ en $S$ is $\mathrm{cov}(W,S) = 0{,}3438$. De correlatie $\rho$ tussen wiskunde en statistiek wordt tenslotte berekend uit de covariantie door te delen door de twee standaardafwijkingen:

$$\rho = \frac{\mathrm{cov}}{\sigma_w \sigma_s} = \frac{0{,}3438}{0{,}9 \cdot 1{,}2} = 0{,}318$$

Een correlatie van 32% tussen wiskunde en statistiek is redelijk, doch aan de zwakke kant.

**Centrale limietstelling bij afhankelijke variabelen**

9 Speculaasjes hebben een gewicht $X$ met $\mu_x = 3$ en $\sigma_x = 0{,}2$. De doosjes hebben een gewicht $Z$ met $\mu_z = 15$ en $\sigma_z = 0{,}5$. De 50 speculaasjes die in een doosje gaan hebben een gewicht $Y$,

$$Y = X_1 + \cdots + X_{50}$$

Voor het gemiddelde van $Y$ geldt de somregel:

$$\mu_y = \mu_x + \cdots + \mu_x = 50\mu_x = 150$$

Voor de standaardafwijking geldt de somregel van de varianties, omdat de speculaasjes onafhankelijke gewichten hebben,:

$$\sigma^2_y = \sigma^2_x + \cdots + \sigma^2_x = 50\sigma^2_x = 1{,}41^2$$

Vanwege het *grote aantal* $n = 50$ speculaasjes mag de centrale limietstelling worden toegepast: $Y$ is *normaal verdeeld*.
Het gewicht van de gevulde doosjes $G = Y + Z$ heeft een gemiddelde volgens de somregel

$$\mu_g = \mu_y + \mu_z = 150 + 15 = 165$$

De standaardafwijking van het totaalgewicht van onafhankelijke stochasten volgt uit de somregel voor varianties:

$$\sigma^2_g = \sigma^2_y + \sigma^2_z = 1{,}41^2 + 0{,}5^2 = 1{,}5^2$$

De kans dat een gevuld doosje minder dan 162 weegt is:

$$P(G < 162) = P(U_g < \frac{162 - 165}{1{,}5}) = P(U_g < -2)$$

Nu is $G$, de som van twee onafhankelijke normale stochasten, zelf ook normaal verdeeld. De kans is dus met tabel 1 te vinden:
$P(U_g < -2) = P(U_g > 2) = 0{,}0228$.

11 Het gewicht $P$ van een pakje, $\mu_p = 20$ en $\sigma_p = 5$, is normaal verdeeld. Ook het gewicht $D$ van een doos, $\mu_d = 520$ en $\sigma_p = 13$, is normaal verdeeld. Het vulgewicht $K$ aan koffie in een pakje, $\mu_k = 250$ en $\sigma_k = 12$, heeft een onbekende verdeling, doch is onafhankelijk van $P$ (pakje) of $D$ (doos). Het totaalgewicht $T = P + K$ van pakje heeft, volgens de somregel van gemiddelden, een gemiddelde

$$\mu_t = \mu_p + \mu_k = 20 + 250 = 270$$

en volgens de somregel van varianties van onafhankelijke stochasten, een standaardafwijking

$$\sigma_t^2 = \sigma_p^2 + \sigma_k^2 = 5^2 + 12^2 = 13^2$$

a De 24 pakjes tezamen met de doos hebben een gewicht $G = D + T_1 + \cdots + T_{24}$. Het gemiddelde gewicht daarvan is volgens de somregel

$$\mu_g = \mu_d + \mu_t + \cdots \mu_t = 520 + 24{\cdot}270 = 7000$$

De standaardafwijking in het gewicht volgt uit de somregel voor varianties van onafhankelijke stochasten:

$$\sigma_g^2 = \sigma_d^2 + \sigma_t^2 + \cdots + \sigma_t^2 = 13^2 + 24{\cdot}13^2 = 65^2$$

b Als doos1 een gewicht $G_1$ heeft, doos2 een gewicht $G_2$, dan wordt gevraagd $P(|G_1 - G_2| > 100)$. Op grond van de *centrale limietstelling* is het gewicht van een doos normaal verdeeld. Verder zijn de twee dozen als onafhankelijk te beschouwen. Dan is het verschil van de twee onafhankelijke normale stochasten $X = G_1 - G_2$ ook een normaal verdeelde stochast, met gemiddelde en standaardafwijking

$$\mu_x = \mu_g - \mu_g = 0 \qquad \sigma_x^2 = \sigma_g^2 + \sigma_g^2 = 91{,}9^2$$

Dan is volgens de spiegelsymmetrie van de verdeling $P(|X| > 100) = 2P(X > 100)$. Na standaardisatie en tabel 1 volgt: $P(X > 100) = P(U_x > 100/91{,}9) = P(U_x > 1{,}088) = 0{,}1383$. De kans op meer dan 100 gram verschil is 27,7%.

c Nadat de koffievulmachine anders is ingesteld, zullen alle gemiddelden gelijk blijven, maar de standaard-afwijkingen veranderen. De doos zal daardoor een gewicht $G$ hebben met gemiddelde $\mu_g = 7000$ en een onbekende standaardafwijking $\sigma_g$. Uit de vereiste $P(G > 7074) = P(U_g > (7074 - 7000)/\sigma_g) = 0{,}0228$ rekenen we de standaardafwijking terug met tabel 1:

$$\frac{74}{\sigma_g} = 2 \qquad \sigma_g = 37$$

Uit de analyse van de standaard-afwijking in a volgt dat:

$$\sigma_g^2 = 13^2 + 24(5^2 + \sigma_k^2)$$

Met $\sigma_g = 37$ volgt daaruit de machineafstelling $\sigma_k = 5$.

14

**Steekproef-gemiddelde en steekproef-standaard-afwijking**

15 De levensduur in uren van een biooscooplampje $X$ is een stochast, $\mu_x = 3250$ en $\sigma_x = 400$, met een normale verdeling.

a De kans om langer dan 3000 uren te branden, $P(X > 3000)$, is vanwege de symmetrie rond het gemiddelde 3250 gelijk aan $P(X < 3500)$. Verder is $P(U_X < (3500-3250)/400) = P(U_X < 0{,}625) = 1-P(U_X > 0{,}625)$. Volgens tabel 1 is $P(U_X > 0{,}625) = 0{,}2660$. Dus 26,6% van de lampen voldoet *niet* aan de eisen.

b Bij veranderde gemiddelde levensduur $\mu_x$ wordt het hiervoor bedoelde percentage voor 3000 branduren gelijk aan 2,5%. De bijbehorende gestandaardiseerde stochastwaarde is $U_t = 1{,}96$. Het gemiddelde ligt nu 1,96 standaard-afwijkingen boven 3000 uur: $\mu_x = 3000 + 1{,}96{\cdot}400 = 3784$.

Voor de pakketjes met 4 lampjes geldt een totale levensduur $T = X_1 + \cdots + X_4$, met een gemiddelde (somregel)

$$\mu_T = \mu_x + \cdots + \mu_x = 4{\cdot}3250 = 13000$$

De standaardafwijking van $T$ wordt (somregel voor varianties van onafhankelijke stochasten):

$$\sigma_T^2 = \sigma_x^2 + \cdots + \sigma_x^2 = 4{\cdot}400^2 = 800^2 =$$

c Nu is de kans om niet aan de eis te voldoen milder, omdat we met de som werken: $P(T < 12000) = P(T > 14000) = P(U_T > (14000 - 13000)/800) = P(U_T > 1{,}25)$. Daaruit vinden we 10,6% uitval.

d Als tevoren is het gemiddelde 1,96 standaard-afwijkingen boven de grenswaarde: $\mu_T = 12000 + 1{,}96{\cdot}800 = 13568$. Per lampje is dus de levensduur $\mu_x = 13548/4 = 3392$.

e Voor pakketjes met $n$ lampjes is analoog:

$$\mu_T = n\mu_x = 3250n \quad \sigma_T^2 = n\sigma_x^2 = 400^2 n$$

We maken nu $n$ zo groot, dat slechts 0,62% uitvalt bij de totaaleis dat $T > n{\cdot}3000$. Dus:

$$0{,}0062 = P(T < n{\cdot}3000) = P(U_T < \frac{n{\cdot}3000 - n{\cdot}3250}{400\sqrt{n}})$$

Met de symmetrie wordt dat:

$$0{,}0062 = P(U_T > 0{,}625\sqrt{n})$$

Uit tabel 1 blijkt dat $0{,}625\sqrt{n} = 2{,}5$, of $n = 16$.

17 De levensduur $P$ van een TL-buis van merk P is een stochast met $\mu_P = 14000$ en $\sigma_P = 2000$. De levensduur $Q$ van een TL-buis van merk Q is een stochast met $\mu_P = 12000$ en $\sigma_P = 1000$. We nemen een steekproef van 125 stuks. Een steekproef-gemiddelde is zelf een stochast $m$, waarvan het gemiddelde gelijk is aan het populatie-gemiddelde $\mu$:

$$\mu_{m_P} = \mu_P = 14000 \quad \mu_{m_Q} = \mu_Q = 12000$$

Het steekproef-gemiddelde, als stochast, heeft een standaard-afwijking die een factor $\sqrt{n} = 11{,}2$ kleiner is dan de populatie-standaard-afwijking $\sigma$:

$$\sigma_{m_P} = \frac{\sigma_P}{\sqrt{n}} = 178{,}9 \quad \sigma_{m_Q} = \frac{\sigma_Q}{\sqrt{n}} = 89{,}4$$

Hoe groot is de kans dat de steekproef-gemiddelden meer dan 1800 verschillen? Het verschil $X = m_P - m_Q$ is een stochast waarvan het gemiddelde en de standaard-afwijking volgt uit de somregels (P en Q zijn natuurlijk onafhankelijk):

$$\mu_X = \mu_{m_P} - \mu_{m_Q} = 14000 - 12000 = 2000$$
$$\sigma_X^2 = \sigma_{m_P} + \sigma_{m_Q} = 178{,}9^2 + 89{,}4^2 = 200^2$$

Volgens de centrale limietstelling zijn de steekproef-gemiddelden normaal verdeelde stochasten. Verder is het verschil van twee onafhankelijke normaal verdeelde stochasten, zoals $X$, zelf ook normaal verdeeld. De vraag is dus wat de kans is dat $P(|X| > 1800)$ als $X$ een normaal verdeelde stochast is met gemiddelde $\mu_x = 2000$ en standaard-afwijking $\sigma_x = 200$.

$$P(|X|>1800) = P(X>1800) + P(X<-1800) = P(U_x > \frac{1800 - 2000}{200})$$

Gezien de positie van het gemiddelde is de kans op negatieve waarden nihil. Dan is
$P(U_x > -1) = P(U_x < 1) = 1 - P(U_x > 1) = 1 - 0{,}1587 = 0{,}8413$. In meer dan 84,1% van de steekproeven zal dus type Q meer dan 1800 uren langer werken (gemiddeld). Dat was te verwachten op grond van het oorspronkelijke verschil van 2000 uren, en de grootte van de steekproef waardoor de standaardafwijking meer dan een factor 10 kleiner wordt.

18 Het steekproef-gemiddelde $m$, van een steekproef ter grootte $n$ van een normaal verdeelde stochast, is zelf ook een normaal verdeelde stochast, met gemiddelde $\mu$ en standaard-afwijking $\sigma/\sqrt{n}$.
Uit de steekproeven van de Quality Officer (steekproeven ter grootte 100) volgt voor de dikte in mm: $P(m_{\mathrm{QO}} > 25{,}19) = 0{,}1711$ , met $m_{\mathrm{QO}}$ normaal verdeeld, gemiddelde $\mu$ en standaard-afwijking $\sigma/\sqrt{100}$. Na standaardisatie betekent dat:

$$P(U_m > \frac{25{,}19 - \mu}{\sigma/10} = 0{,}1711$$

Uit tabel 1 zoeken we terug:

$$\frac{25{,}19 - \mu}{\sigma/10} = 0{,}95$$

Op dezelfde manier kunnen we de resultaten van de Production Manager (steekproeven ter grootte 25) uitdrukken als:

$$\frac{25{,}76 - \mu}{\sigma/5} = 1{,}90$$

Er zijn nu twee vergelijkingen voor de twee onbekenden $\mu$ en $\sigma$, het gemiddelde en de standaardafwijking van de tegeldikte. Deze kunnen worden opgelost (bijvoorbeeld: eerst vergelijking op elkaar delen om $\sigma$ te elimineren zodat $\mu$ kan worden gevonden) met $\mu = 25$ en $\sigma = 2$.

# 7 Discrete kansverdelingen

## 7.8 Uitwerkingen opgaven

### Kansverdeling

3 $K$ is 'het aantal gooien van een dobbelsteen tot een 6 is verkregen'. De *kansfunctie* is $f(k) = P(\text{gooien } 1 \ldots k-1\text{: niet 6; gooi } k\text{: wel 6})$. De waarde is: $f(k) = (5/6)^{k-1}(1/6)$ voor $k = 1 \ldots$. Dat zijn achtereenvolgens de kansen: 0,167; 0,139; 0,116; 0,096; 0,080; 0,067; 0,056; enz.

### Binomiale-verdeling

8 $K$ is 'in een 4 kinderen gezin zijn er evenveel jongens als meisjes'. In een 4 kinderen gezin betekent evenveel jongens als meisjes: 2 jongens en 2 meisjes. Van de $2^4 = 16$ mogelijke gezinssamenstellingen zijn er 6 met 2 jongens en 2 meisjes. De gevraagde kans is dus 6/16=3/8=0,375.

6 $K$ is 'tenminste 1 van de 4 keer niet in gesprek bij bellen tussen 14.00 en 15.00 in Dobbeldam'. Dus is $\overline{K}$ gelijk aan 'alle 4 in gesprek'. Noem 'in gesprek' $G$, met $P(G) = 1/4$. De gevraagde kans is:

$$P(K) = 1 - P(G^4) = 1 - \tfrac{1}{4}^4 = 0{,}996$$

7 $K$ is 'aantal zieke werknemers'. De kans op 'ziek' $Z$ is $P(Z) = 0{,}04$. Verder is 'hoogstens 1 ziek' gelijk $(K \leq 1) = (K = 0) + (K = 1)$. Bij $n = 5$ werknemers en 'geen zieke':

$$P(K = 0) = P((\overline{Z})^5) = 0{,}96^5 = 0{,}815$$

Bij 'één zieke' zijn er 5 mogelijke werknemers kandidaat. Dus:

$$P(K = 1) = 5 \cdot P((\overline{Z})^4(Z)^1) = 5 \cdot 0{,}96^4 \cdot 0{,}04^1 = 0{,}170$$

Tezamen is de *cumulatieve kans* $P(k \leq 1 | n = 5,\ p = 0{,}04) = 0{,}985$.

9 $R$ is het trekken van een rode knikker. Dan is bij een eerste trekking

$$P(R) = \tfrac{1}{6} \quad P(\overline{R}) = \tfrac{5}{6}$$

Noem $k$ het aantal rode knikkers van de 3 getrokken knikkers.

b Bij teruglegging zijn de kansen bij iedere trekking gelijk. Trekkingen met teruglegging zijn onafhankelijk, zodat de speciale produkteigenschap geldt. Voor $k = 0$:

$$P(k = 0) = P(\overline{R}^3) = \tfrac{5}{6}^3 = 0{,}579$$

Voor $k = 1$ kan de ene rode knikker in drie trekkingen zitten:

$$P(k = 1) = 3P(\overline{R}^2 \cdot R) = 3 \cdot \tfrac{5}{6}^2 \cdot \tfrac{1}{6} = 0{,}347$$

Voor $k = 2$ kan de ene niet rode knikker in 3 trekkingen zitten:

$$P(k = 2) = 3P(\overline{R} \cdot R^2) = 3 \cdot \tfrac{5}{6} \cdot \tfrac{1}{6}^2 = 0{,}069$$
$$P(k = 3) = P(R^3) = \tfrac{1}{6}^3 = 0{,}005$$

a Zonder teruglegging moeten we naar het totaal effect kijken. We kunnen op 6.5.4 manieren 3 knikkers trekken uit 6 knikkers. Bij de eerste hebben we keus uit 6, bij de tweede keus uit 5 en bij de derde keus uit 4. Bij de getrokken knikkers zit niet of wel de rode knikker, dus $k = 0$ of $k = 1$. Voor $k = 0$ zijn er 5.4.3 trekkingen (nu mogen we alleen uit de 5 niet rode knikkers kiezen). Dus

$$P(k = 0) = \frac{5 \cdot 4 \cdot 3}{6 \cdot 5 \cdot 4} = \frac{3}{6} = 0{,}50$$

Met de complementregel volgt dan $P(k = 1) = 0{,}50$.

**Binomiale-verdeling met normale benadering**

16 $K$ is 'aantal personeelsleden dat buiten vestigingsplaats woont'. $K$ heeft een binomiale verdeling met een groot aantal herhalingen $n = 475$ met kleine kans $p = 0{,}05$. We voldoen hiermee aan de conditie voor een normale benadering: $n \geq 9(0{,}95/0{,}05) = 171$. In die benadering is het gemiddelde $\mu = np = 475 \cdot 0{,}05 = 23{,}75$ en de spreiding $\sigma = \sqrt{475 \cdot 0{,}05 \cdot 0{,}95} = 4{,}75$. Verder moeten we in plaats van de discrete $k = 20$ de continu '$X = 20$' berekenen:

$$P(k = 20) = P(19{,}5 < X < 20{,}5)$$

Dat wordt na standaardisatie:

$$P(\frac{19{,}5 - 23{,}75}{4{,}75} < U < \frac{20{,}5 - 23{,}75}{4{,}75}) = P(-0{,}895 < U < -0{,}684)$$

Met de tabel 1 vinden we de kans
$P(U > 0{,}684) - P(U > 0{,}895) = 0{,}2470 - 0{,}1854 = 0{,}062$. De kans is iets te klein (vergeleken met de Poisson benadering, die iets beter is).

22 $K$ is 'het gemiddeld aantal zessen bij het gooien van een dobbelsteen'. $K$ heeft een binomiale verdeling, met een gemiddelde $\mu = \frac{1}{6}$ en een spreiding $\sigma = \sqrt{\frac{1}{6}\cdot\frac{5}{6}/n} = \sqrt{5}/6\sqrt{n}$. Volgens de centrale limietstelling tendeert de verdeling naar een normale verdeling (als aan de voorwaarde $n > 9(\frac{5}{6}/\frac{1}{6}) = 45$ is voldaan). Gevraagd is om $n$ te bepalen zodat:

$$P(9/60 < K < 11/60) \geq 0{,}9544$$

Zonder de continuïteitscorrectie wordt dit benadert met de normaal verdeelde $X$:

$$P(9/60 < X < 11/60) = P(-\frac{1/60}{\sigma} < U < \frac{1/60}{\sigma}) \geq 0{,}9544$$
$$P(U > \frac{1/60}{\sigma}) \leq \tfrac{1}{2}\cdot(1 - 0{,}9544) = 0{,}0228$$

Volgens tabel 1 is dan $1/(60\sigma) \geq 2{,}00$, dus $\sigma \leq 1/120$. Als opgemerkt is $\sigma = \sqrt{5}/6\sqrt{n}$, zodat voor de steekproefgrootte $\sqrt{n} \geq 120\sqrt{5}/6 = 20\sqrt{5}$. Conclusie: $n \geq 2000$.
Merk op, dat na 2000 gooien de kans op 6 nog 5% spreiding heeft. In het algemeen is de relatieve spreiding gelijk aan $\sqrt{5/n}$, zodat we voor 1% moeten doorgaan tot een steekproef van 50000!

**Binomiale-verdeling met Poisson benadering**

16 $K$ is 'aantal personeelsleden dat buiten vestigingsplaats woont'. $K$ heeft een binomiale verdeling met een groot aantal herhalingen $n = 475$ met kleine kans $p = 0{,}05$. Ze voldoet hiermee aan de conditie voor een Poisson benadering, waarbij het gemiddelde $m = np = 475\cdot0{,}05 = 23{,}75$. Daarmee is de gevraagde kans:

$$P(k = 20) \approx e^{-23{,}75}\frac{23{,}75^{20}}{20!} = 0{,}065$$

12 $K$ is 'er is precies één kolom 'goed' (minstens 8 juist) ingevuld'. Ingevuld worden $n = 10\cdot10 = 100$ kolommen. Gevraagd wordt naar de discrete gebeurtenis 1 kolom goed, 99 fout. De goede kolom kan een van de 100 zijn, zodat de gevraagde kans de binomiaalkans is voor $k = 1$, $n = 100$ en $p = P(K)$:

$$P(100\cdot K\cdot\overline{K}^{99}) = 100\cdot P(K)\cdot P(\overline{K})^{99}$$

Om een kolom 'goed' te hebben moeten 9 of 8 uitslagen juist zijn (met kans 1 uit 3). Ook dat zijn binomiaalkansen, dus

$$P(K) = \tfrac{1}{3}^9 + 9\cdot\tfrac{1}{3}^8\cdot\tfrac{2}{3}^1 = 9{,}653\cdot10^{-4}$$

Vanwege de kleine waarde van $p = P(K)$ en de grote van $n$ is de Poisson benadering van de binomiaalkansen van toepassing, met $m = np = 100{\cdot}9{,}653{\cdot}10^{-4} = 0{,}09653$:

$$P(k=1) \approx e^{-0{,}09653} \cdot \frac{0{,}09653^1}{1!} = 0{,}088$$

11 Winst wordt gemaakt als de tweede en derde schijf gelijk zijn aan de eerste. De kans daarop is $p = (1/5)^2 = 1/25 = 0{,}04$.

a Na één spel is de verwachtte uitbetaling gemiddeld $\mu = 0{,}04{\cdot}10 = 0{,}4$. Na 30 *onafhankelijke* spelen is de uitbetaling dus $30{\cdot}0{,}4 = 12$

b 'Na 30 keer winst' betekent dat na 30 keer meer dan 30 is uitbetaald, oftewel dat er minstens 4 maal is gewonnen. We hebben te maken met een binomiale-verdeling met kleine kans $p = 0{,}04$ en een groot aantal herhalingen $n = 30$. Eerst berekenen we $P(0)$, $P(1)$, $P(2)$ en $P(3)$ en gebruiken de binomiaalcoëfficiënten $\binom{30}{0} = 1$, $\binom{30}{1} = 30$, $\binom{30}{2} = 30{\cdot}29/2 = 435$, $\binom{30}{3} = 30{\cdot}29{\cdot}28/2{\cdot}3 = 4060$:

$$\begin{aligned} P(0) &= 0{,}96^3 0 = 0{,}294 \\ P(1) &= 30{\cdot}0{,}04{\cdot}0{,}96^2 9 = 0{,}367 \\ P(2) &= 435{\cdot}0{,}04^2{\cdot}0{,}96^2 8 = 0{,}222 \\ P(3) &= 4060{\cdot}0{,}04^3{\cdot}0{,}96^2 7 = 0{,}086 \end{aligned}$$

De kans op winst blijkt dus te zijn, gebruik makend van de complementsregel:

$$P(k>3) = 1 - 0{,}294 - 0{,}367 - 0{,}222 - 0{,}086 = 0{,}031$$

We kunnen ook rechtstreeks de verdere termen uitrekenen: $P(k=4)$, $P(k=5)$, $P(k=6)$, enz. De algemene term is voor $n = 30$ en $p = 0{,}04$ te benaderen als Poisson-verdeling met $m = np = 30{\cdot}0{,}04 = 1{,}2$:

$$P(k) = \binom{30}{k} 0{,}04^k 0{,}96^{30-k} \approx \frac{1{,}2^k}{k!} e^{-1{,}2}$$

Dat geeft de waarden:

$$\begin{aligned} P(4) &\approx e^{-1{,}2}{\cdot}1{,}2^4/24 = 0{,}026 \\ P(5) &\approx e^{-1{,}2}{\cdot}1{,}2^5/120 = 0{,}006 \\ P(6) &\approx e^{-1{,}2}{\cdot}1{,}2^6/720 = 0{,}001 \\ P(7) &\approx e^{-1{,}2}{\cdot}1{,}2^7/5040 = 0{,}000 \end{aligned}$$

waaruit de benadering $P(k>3) \approx 0{,}033$. Slechts een klein verschil.

Zou ook de normale benadering zinvol zijn? De conditie daarvoor is dat $n \geq 9(0{,}96/0{,}04) = 216$. Nu is $n = 30 << 216$, zodat de normale benadering niet zinvol is. Zouden we toch die benadering gebruiken om een indruk te krijgen, dan zouden we vinden dat $P(k \geq 4) \sim 0{,}016$. Inderdaad is de normale benadering niet zinvol.

**Poisson-verdeling**

14 c $K$ is 'aantal vraagtekens op twee pagina's', een Poisson verdeelde stochast met een gemiddelde $m = 2{\cdot}11/10 = 2{,}2$. Gevraagd wordt

$$P(k = 4|m = 2{,}2) = 0{,}109$$

volgens tabel 3.

a $K$ is 'aantal vraagtekens op een pagina', een Poisson verdeelde stochast met een gemiddelde $m = 11/10 = 1{,}1$. Gevraagd wordt

$$P(k \geq 3|m = 1{,}1) = 1 - P(k \leq 2|m = 1{,}1)$$

De laatste kans is op te zoeken in tabel 4:
$P(k \leq 2|m = 1{,}1) = 0{,}900$, zodat de kans op minstens 3 vraagtekens op 2 pagina's is 0,10.

b $K$ is 'aantal pagina's met elk minstens 3 vraagtekens', een Poisson verdeelde stochast met een gemiddelde $m = np$, waarbij $n$ het (onbekende) aantal pagina's is en $p$ de kans op 'minstens 3 vraagtekens op een pagina'. Volgens 'a' is $p = 0{,}1$, waarmee $m = 0{,}1n$ wordt. Gegeven is dat

$$P(k \leq 5|m = 0{,}1n) \geq 0{,}785$$

Uit tabel 4 zoeken we terug dat $m = 0{,}1n \geq 4{,}0$, dus $n \geq 40$.

23 a $K$ is 'het aantal binnenkomende en uitgaande gesprekken per minuut tussen 9.00 en 10.00', een Poisson verdeelde stochast met een gemiddelde $m = 960/60 = 16$. Het gemiddelde is zo groot dat we de *normale benadering* kunnen toepassen, met een gemiddelde $\mu = m = 16$ en een standaardafwijking $\sigma = \sqrt{m} = 4$. De centrale raakt overbelast als $k > 20$, dus $k \geq 21$. De gevraagde kans is:

$$P(k \geq 21) = P(X > 20{,}5) = P(U > \frac{20{,}5 - 16}{4} = P(U > 1{,}125)$$

Volgens tabel 1 is $P(U > 1{,}125) = 0{,}1303$.

b Noem de capaciteit na uitbreiding $n$. Dan kan in het vorige '20' door '$n$' worden vervangen. In het bijzonder volgt uit het gegeven dat

$$P(U > \frac{n + 0{,}5 - 16}{4}) \leq 0{,}05$$

Uit tabel 1 teuggezocht is $n + 0{,}5 - 16/4 \geq 1{,}645$, dus $n \geq 22{,}08$. Praktisch gesproken kan de capaciteit tot 22 worden uitgebreid, alhoewel voor de zekerheid 23 nodig is.

**Poisson-verdeling met normale benadering**

14 d $K$ is 'aantal pagina's met elk meer dan 2 vraagtekens', een discrete variabele die *binomiaal* verdeeld is, met $n = 400$ en $p$ de kans op 'meer dan 2 vraagtekens op een pagina'. Het laatste is gelijk aan 'minstens 3 vraagtekens op een pagina', dus $p = 0{,}1$ volgens 14a. Verder is $n$ zo groot, $n \geq 9 \cdot 0{,}9/0{,}1 = 81$, dat we de *normale benadering* kunnen gebruiken. De continue stochast $X$ heeft een gemiddelde $\mu = np = 40$ en een standaardafwijking $\sigma = \sqrt{npq} = 6{,}00$. Dan is, met de continïteitscorrectie,

$$P(k \geq 51) \approx P(X > 50{,}5) = P(U > \frac{50{,}5 - 40}{6{,}00})$$

met $P(U > \frac{50{,}5-40}{6{,}0}) = P(U > 1{,}75) = 0{,}040$ volgens tabel 1.

20 $K$ is 'aantal jarigen per dag', een *Poisson* verdeelde discrete variabele. Het gemiddelde aantal $\mu = 36500/365 = 100$ met standaardafwijking $\sigma = \sqrt{\mu} = 10$. Gevraagd wordt $P(K \geq 90)$. Omdat $\mu > 20$ kunnen we de normale benadering toepassen:

$$P(K \geq 90) = P(X > 89{,}5) = P(U > \frac{89{,}5 - 100}{10}$$

Volgens tabel 1 is
$P(U > -1{,}05) = 1 - P(U > 1{,}05) = 1 - 0{,}1469 = 0{,}8531$. De kans op minstens 90 jarigen is 85,31%.

# 9 Het toetsen van hypothesen

## 9.12 Uitwerkingen opgaven

**Normaleverdeling met bekende standaardafwijking**

3 We volgen de toetsingsprocedure als beschreven in 9.5.
1. Nulhypothese $H_0$: $\mu = 500$;
   Alternatieve hypothese $H_1$: $\mu < 500$.
2. De toetsingsvariabele $X$, het gemiddelde netto gewicht, heeft een normale verdeling met standaardafwijking $\sigma = 28$.
3. De onbetrouwbaarheid $\alpha = 0{,}05$ links-eenzijdig ('onvoldoende gewicht'). De bijbehorende $u$-waarde is $u(0{,}05) = 1{,}645$
4. Op grond van de nulhypothese is $\mu = 500$. Bij een steekproefgrootte $n = 16$ wordt de linker kritieke grens $l$:
$$l = \mu - u\frac{\sigma}{\sqrt{n}} = 500 - 1{,}645\frac{28}{\sqrt{4}} = 488{,}5$$
5. In de steekproef wordt gevonden $m = 485$. Dus ligt het steekproefgemiddelde in het kritieke gebied:
$$485 = m < l = 488{,}5$$
6. Aangezien de kans daarop slechts 0,05 is, wordt de nulhypothese verworpen, en de alternatieve hypothese aanvaardt.
7. De machine moet worden bijgesteld teneinde voldoende gemiddelde gewicht te kunnen leveren.

6 Gegeven is $\sigma_1 = 1200$ en $\sigma_2 = 1600$ van normaal verdeelde treksterkten van betonstaal volgens methoden 1 en 2 gewalst. Nulhypothese $H_0$: $\mu_1 = \mu_2$; alternatief $H_1$: $\mu_1 \neq \mu_2$. Te toetsen met een tweezijdige onbetrouwbaarheid $\alpha = 0{,}05$.
De toetsvariabele $X = X_1 - X_2$ is het verschil in treksterkte van de steekproefgemiddelden. Voor de toetsvariabele vinden we (volgens nulhypothese)
$$\mu = 0$$
De steekproefgemiddelden zijn normaal verdeeld met standaardafwijkingen $\sigma/\sqrt{n}$, dus respectievelijk $1200/\sqrt{12}$ en $1600/\sqrt{15}$. Voor de standaardafwijking van $X$ (onafhankelijke steekproeven, dus varianties tellen op) geldt:
$$\sigma^2 = \frac{1200^2}{12} + \frac{1600^2}{15} = 539^2$$

De onbetrouwbaarheid is tweezijdig, bij $\alpha/2 = 0{,}025$ is $u = 1{,}96$, zodat de linker en rechter kritieke grens worden:

$$l = \mu - u\sigma = 0 - 1{,}96 \cdot 539 = -1056 \quad r = +1056$$

Uit de steekproeven komen $\mu_1 = 60000$ en $\mu_2 = 59000$, zodat de gevonden toetsvariabele is:

$$m = \mu_1 - \mu_2 = 60000 - 59000 = 1000$$

Omdat $m$ *niet* in het kritieke gebied $(X < l) \cup (X > r)$ ligt: $-1056 = l < m = 1000 < 1056 = r$, zullen we de nulhypothese niet verwerpen. De twee walsmethoden geven waarschijnlijk (95%) dezelfde gemiddelde treksterkte.

**Normaleverdeling met onbekende standaardafwijking**

5 We volgen de toetsingsprocedure als beschreven in 9.5.

1. Nulhypothese $H_0$: $\mu = 7{,}1$; alternatief $H_1$: $\mu > 7{,}1$.
2. Toetsvariabele $X$ is het gemiddelde benzineverbruik.
3. Onbetrouwbaarheid $\alpha = 0{,}05$ is rechts-eenzijdig.
4. Volgens de nulhypothese is $\mu = 7{,}1$. Volgens de steekproef, $n = 10$, is de steekproefstandaardafwijking $s = 0{,}23$. De rechteroverschrijdingskans moet worden afgeschat met de t-verdeling bij aantal vrijheidsgraden $v = 10 - 1 = 9$. Bij een betrouwbaarheid $\alpha = 0{,}05$ is $t = 1{,}833$ (tabel 5). Daaruit volgt de de (rechter)grens van het kritieke gebied:

   $$r = \mu + t \cdot \frac{s}{\sqrt{n}} = 7{,}1 + 1{,}833 \cdot \frac{0{,}23}{\sqrt{10}} = 7{,}23$$

   Dit betekent: bij een gemiddelde van 7,1 is de kans dat het steekproefgemiddelde $m > r$ kleiner dan 5%.
5. Het gevonden steekproefgemiddelde $m = 7{,}3$ ligt in het kritieke gebied:

   $$r = 7{,}23 < m = 7{,}3$$

   zodat de nulhypothese moet worden verworpen, en de altenatieve aanvaardt.
6. Zeer waarschijnlijk (95%) is het werkelijke benzineverbruik hoger dan de fabrikant beweert.

7 Nulhypothese $H_0$: $\mu = 20$; alternatief $H_1$: $\mu > 20$.
Toetsvariabele is het gemiddelde benzineverbruik.
De betrouwbaarheid $\alpha = 0{,}05$ is rechts-eenzijdig.
Steekproefgrootte $n = 3$ met steekproefgemiddelde $m = 22$ en steekproefstandaardafwijking $s = 1$. Bij een betrouwbaarheid $\alpha = 0{,}01$

hoort, bij $3-1=2$ vrijheidsgraden, een $t$-waarde $t=6{,}965$. Daaruit vinden we de rechtergrens van het kritieke gebied als:

$$r = 20 + 6{,}965 \cdot \frac{1}{\sqrt{3}} = 24{,}0$$

De gevonden $m = 22 < 24 = r$ ligt *niet* in het kritieke gebied. De nulhypothese wordt niet verworpen. Er zijn onvoldoende aanwijzingen dat er significant meer copiën worden afgedrukt per minuut. Misschien dat een grotere steekproef daarin meer duidelijk kan geven.

**Binomialeverdeling met onbekende $p$**

4 Nulhypothese $H_0$: $p = P(J) = 0{,}500$; alternatief $H_1$: $p \neq 0{,}500$.
Toetsingsvariabele is het aantal geboren jongens $K$ in de steekproef van grootte $n = 3000$.
Betrouwbaarheid $\alpha = 0{,}05$ tweezijdig getoetst.
Omdat $n = 3000 > 9(q/p) = 9$ mogen we aannemen dat $K$, die binomiaal verdeeld is, in goede benadering normaal verdeeld is met een gemiddelde $\mu = np = 3000 \cdot 0{,}50 = 1500$ en een standaardafwijking $\sigma = \sqrt{npq} = \sqrt{3000 \cdot 0{,}5 \cdot 0{,}5} = 27{,}39$.
Voor de linker en rechter kritieke grens vinden we, met de bij $\alpha/2 = 0{,}025$ behorende $u = 1{,}96$ in eerste instantie de continue benadering:

$$l = \mu - u\sigma = 1500 - 1{,}96 \cdot 27{,}39 = 1446{,}3 \quad r = 1553{,}7$$

Vanwege de continuïteitscorrectie (respectievelijk $-0{,}5$ en $+0{,}5$) worden de kritieke grenzen:

$$l = 1445 \quad r = 1555$$

De nulhypothese zal *niet* worden verworpen als $1445 < K < 1555$ wordt gevonden.

**Toets frequentieverdeling**

11 Nulhypothese $H_0$: alle frequenties zijn gelijk; alternatief $H_1$: niet $H_0$.
De betrouwbaarheid is $\alpha = 0{,}01$, eenzijdig te toetsen. De toetsvariabele zal zijn de gewogen som $\chi^2$, met het aantal vrijheidsgraden $v = n - 1 = 10 - 1 = 9$ (er zijn $n = 10$ verschillende cijfers $k$-waarden die kunnen worden gegenereerd). Op grond van tabel 6 hoort in geval van 9 vrijheidsgraden bij een betrouwbaarheid $\alpha = 0{,}01$ een kritieke waarde voor $\chi^2_v = 21{,}67$.
Teneinde uit de metingen $\chi^2$ te berekenen bepalen we naast de *gevonden* frequenties $O_k$, op grond van de nulhypothese, de te *verwachten* frequenties $E_k$.

**Frequenties**

| cijfer $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| gevonden $O$ | 35 | 22 | 17 | 20 | 24 | 42 | 37 | 43 | 24 | 36 |
| verwacht $E$ | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 |

Daarmee berekenen we de toetsvariabele $\chi^2$, de som van termen $(O-E)^2/E$:

$$\begin{aligned}\chi^2 &= \frac{(35-30)^2}{30}+\frac{(22-30)^2}{30}+\cdots+\frac{(24-30)^2}{30}+\frac{(36-30)^2}{30}\\ &= \frac{25}{30}+\frac{64}{30}+\cdots+\frac{36}{30}+\frac{36}{30}=0{,}83+2{,}13+5{,}63+\\ &\quad +3{,}33+1{,}20+4{,}80+1{,}63+5{,}63+1{,}20+1{,}20=27{,}58\end{aligned}$$

Omdat de gevonden $\chi^2 = 27{,}58 > 21{,}67 = \chi^2_v$ moet de nulpypothese worden verworpen. De randomgenerator genereert dus zeker *niet random*.

12 Er zijn $n = 400$ 'gezinnen met 4 kinderen' onderzocht op het 'aantal jongens' resp. meisjes (alternatieven). Met succeskans $P(J) = p = 0{,}5$ voldoen de stochasten meestal aan de binomiale verdeling. We zullen de gegevens analyseren en stap voor stap vaststellen hoe het 'werkelijk' zit:

a. is de kans $p = 0{,}5$ (aannemend wel binomiaal)?
b. zo nee, wat is $p$ dan (op grond gegevens)?
c. is de verdeling binomiaal (op grond gevonden $p$)

a De nulhypothese $H_0$: $p = 0{,}5$ in binomiale verdeling; alternatief $H_1$: $p \neq 0{,}5$. De onbetrouwbaarheid $\alpha = 0{,}05$ wordt tweezijdig getoetst. De toetsvariabele is $\chi^2$, met in dit geval $v = (5-1) = 4$ vrijheidsgraden. Bij deze betrouwbaarheid hoort volgens tabel 6 een kritieke $\chi^2_v = 9{,}49$.
Teneinde de werkelijke $\chi^2$ te bepalen berekenen we op grond van de nulhypothese het aantal gezinnen met een gegeven samenstelling. Het te verwachten aantal is $E = n\binom{4}{k}/16 = 25\binom{4}{k}$:

| aantal $O(E)$ | 29(25) | 121(100) | 155(150) | 84(100) | 11(25) |
|---|---|---|---|---|---|

Daarmee wordt

$$\begin{aligned}\chi^2 &= \frac{(29-25)^2}{25}+\frac{(121-100)^2}{100}+\frac{(155-150)^2}{150}+\\ &\quad +\frac{(84-100)^2}{100}+\frac{(11-25)^2}{25}\\ &= 0{,}64+4{,}41+0{,}17+2{,}56+7{,}84=15{,}6\end{aligned}$$

Omdat de gevonden $\chi^2 = 15{,}6 > \chi^2_v = 9{,}49$ moet de nulhypothese worden verworpen. De kans op een jongen is niet gelijk aan de kans op een meisje.

b Uit de gevens volgt de volgende verwachting voor de kans op een jongen $p$:

$$p = (29{\cdot}4 + 121{\cdot}3 + 155{\cdot}2 + 84{\cdot}1 + 11{\cdot}0)/(400{\cdot}4) = 0{,}546$$

Denk eraan dat het aantal kinderen $400{\cdot}4 = 1600$ is.
We wijden even uit over de vraag 'hoeveel decimalen' verantwoord (significant) zijn. Voor het antwoord gaan we uit van de steekproefstandaardafwijking $s = 0{,}235$. Daaruit volgt een standaardafwijking voor het gemiddelde (dus de berekende $p$) van $s/\sqrt{1600} = 0{,}006$, zodat 95% van de waarden binnen 0,012 zal liggen. De *tweede decimaal* is dus al niet meer betrouwbaar in de waarde van de gemiddelde $p = 0{,}546$. Toch geven we nog een extra decimaal, omdat anders de waarden relatief te sterk verspringen (bij verandering in laatste decimaal).

c Nulhypothese $H_0$: binomiale verdeling met $p = 0{,}546$; alternatief $H_1$: andere verdeling. De onbetrouwbaarheid $\alpha = 0{,}05$ wordt rechtszijdig getoetst op de toetsvariabele $\chi^2$. Het aantalvrijheidgraden is $v = 5 - 1 = 4$ en de kritieke $\chi^2_v = 9{,}49$.
Op grond van de binomiale verdeling met $n = 400$ en $p = 0{,}546$ bepalen we opnieuw de te verwachten frequenties $E = nP(k)$:

$$E = n\binom{4}{k}p^k q^{4-k} = 400\binom{4}{k}p^k q^{4-k}$$

Dat leidt tot de volgende gevonden ($O$) en verwachtte ($E$) frequenties:

| $O(E)$ | 29(35,5) | 121(118,2) | 155(147,5) | 84(81,7) | 11(17,0) |
|---|---|---|---|---|---|

Daarmee wordt

$$\begin{aligned}\chi^2 &= \frac{(29-35{,}5)^2}{35{,}5} + \frac{(121-118{,}2)^2}{118{,}2} + \\ &\quad + \frac{(155-147{,}5)^2}{147{,}5} + \frac{(84-81{,}7)^2}{81{,}7} + \frac{(11-17{,}0)^2}{17{,}0} \\ &= 1{,}19 + 0{,}07 + 0{,}38 + 0{,}06 + 2{,}12 = 3{,}82\end{aligned}$$

Omdat de gevonden $\chi^2 = 3{,}82 < \chi^2_v = 9{,}49$ kan de nulhypothese *niet* worden verworpen. Waarschijnlijk hebben we te maken met een binomale verdeling, voor het aantal jongens in vierkind-gezinnen, met een ongelijke kans voor jongen of meisje: $P(J) = 0{,}546$.

**Toets onafhankelijkheid**

18 Nulhypothese $H_0$: oorzaak defect onafhankelijk assemblageplaats; alternatief $H_1$: niet $H_0$. Onbetrouwbaarheid $\alpha = 0{,}05$ rechtszijdig te toetsen op de toetsvariabele $\chi^2$. Deze variabele heeft $v = (2-1)(2-1) = 1$ vrijheidsgraden, omdat de tabel horizontaal uit 2 rijen, en verticaal uit 2 kolommen bestaat; uit tabel 6 volgt de kritieke $\chi^2_v = 3{,}84$.
De onafhankelijkheid in de nulhypothese betekent '$P(A{\cdot}B) = P(A){\cdot}P(B)$', als $A$ op 'plaats' en $B$ op 'fout' duidt. In het bijzonder is $P(\text{München}) = 600/1500 = 0{,}4$ en
$P(\text{Seoel}) = 1 - 0{,}4 = 0{,}6$. Verder is
$P(\text{materiaalfout}) = 1100/1500 = 0{,}733$ en
$P(\text{constructiefout}) = 1 - 0{,}733 = 0{,}267$. Daarmee zijn de te verwachten waarden $E$ van de tabel te berekenen:

| frequentie | München | Seoel | totaal |
|---|---|---|---|
| materiaalfout(E) | 500(440) | 600(660) | 1100 |
| constructiefout(E) | 100(160) | 300(240) | 400 |
| totaal | 600(600) | 900(900) | 1500 |

De toetsvariabele $\chi^2 = \sum(O-E)^2/E$ wordt:

$$\begin{aligned}\chi^2 &= \frac{(500-440)^2}{440} + \frac{(600-660)^2}{660} + \\ &\quad + \frac{(100-160)^2}{160} + \frac{(300-240)^2}{240} \\ &= 8{,}18 + 5{,}45 + 22{,}5 + 15{,}0 = 51{,}1\end{aligned}$$

Omdat de gevonden $\chi^2 = 51{,}1 > \chi^2_v = 3{,}84$ moet de nulhypothese worden verworpen. De fouten zijn plaatsafhankelijk, in het bijzonder is het aantal constructiefouten in München significant laag (en die in Seoel dus significant hoog).

C.M.Fortuin, fortuinc@xs4all.nl