Tentamen Statistische methoden
4052STAMEY
20 april 2012, 9:00–12:00

Studienummers: Vult u alstublieft op het MC formulier uw *Delftse* studienummer in; en op het open vragen formulier graag beide, naar volgend voorbeeld: 1234567(D), 7654321(L).

Bij dit examen is het gebruik van een (evt. grafische) rekenmachine toegestaan. Tevens krijgt u een formuleblad uitgereikt—na afloop inleveren alstublieft. Normering: De meerkeuzevragen tellen voor één derde en de open vragen voor twee derde van het cijfer. Bij de open vragen telt elk (vraag)onderdeel even zwaar.

Meerkeuzevragen

Toelichting: In het algemeen zijn niet altijd vijf van de zes alternatieven 100% fout, het juiste antwoord is het meest volledige antwoord. Maak op het bijgeleverde antwoordformulier het hokje behorende bij het door u gekozen alternatief zwart of blauw. Doorstrepen van een fout antwoord heeft geen zin: u moet het òf uitgummen, òf verwijderen met correctievloeistof òf een nieuw formulier invullen. Vergeet niet uw studienummer in te vullen èn aan te strepen.

---

1. Van de gebeurtenissen A, B en C is gegeven dat $B \cap C = \emptyset$, $\frac{1}{2} = \mathrm{P}(B) = 2\mathrm{P}(C)$, $\mathrm{P}(A \mid B) = \frac{1}{2}$ en $\mathrm{P}(A \mid C) = \frac{2}{5}$. Bereken $\mathrm{P}(A \mid B \cup C)$.

a. $\frac{11}{60}$ **b.** $\frac{23}{60}$ **c.** $\frac{28}{60}$ **d.** $\frac{31}{60}$ **e.** $\frac{36}{60}$ **f.** $\frac{54}{60}$

2. Een uitvinder heeft een apparaat bedacht dat olie kan detecteren. Als olie aanwezig is, dan geeft dit apparaat dit in 75% van de gevallen aan. Als er géén olie aanwezig is geeft het apparaat toch in 20% van de gevallen aan dat er wèl olie aanwezig is. Een oliemaatschappij test het apparaat op locaties waarvan men denkt dat de kans op aanwezigheid van olie $\frac{1}{500}$ is. Als het apparaat aanwezigheid van olie aangeeft, wat is dan de kans dat er ook echt olie zit?

a. 0.0006 **b.** 0.0020 **c.** 0.0075 **d.** 0.9925 **e.** 0.9980 **f.** 0.9994

3. Laat X uniform over $[0, 1]$ verdeeld zijn, en laat $Y = X^2$. Bereken $\mathrm{Cov}(X, Y)$. Deze is gelijk aan

a. $-\frac{1}{4}$ **b.** $\frac{3}{16}$ **c.** $\frac{3}{8}$ **d.** $\frac{2}{9}$ **e.** $\frac{1}{12}$ **f.** $\frac{1}{4}$

4. Voor het uitvoeren van een simulatiestudie is het nodig te simuleren aan de hand van de volgende verdelingsfunctie: $F(x) = 1 - \mathrm{e}^{-20x^2}$ voor $x > 0$ (en $F(x) = 0$ voor $x < 0$). Als U een $U(0, 1)$ verdeelde stochast is dan heeft X verdelingsfunctie F indien

a. $X = 1 - \mathrm{e}^{-20U^2}$ **b.** $X = \mathrm{e}^{-20U^2}$ **c.** $X = \sqrt{-0.05 \ln U}$
d. $X = \sqrt{-\ln(0.05U)}$ **e.** $X = -0.05 \ln U$ **f.** $X = -20 \ln U$

5. De bedrijven A en B besluiten via de brievenbus gezamenlijk een reclamecampagne te starten. Zij X de fractie adressen die reageren bij bedrijf A en Y de fractie, die reageert bij bedrijf B. De gezamenlijke dichtheid van het paar (X, Y) wordt gegeven door:

$$f_{X,Y}(x, y) = \begin{cases} \frac{2}{5}(x + 4y), & 0 < x < 1, 0 < y < 1, \\ 0, & \text{elders.} \end{cases}$$

De kans dat tenminste 30 procent reageert bij bedrijf A ligt in het volgende interval:

a. $[0, \frac{1}{6})$ **b.** $[\frac{1}{6}, \frac{1}{3})$ **c.** $[\frac{1}{3}, \frac{1}{2})$ **d.** $[\frac{1}{2}, \frac{2}{3})$ **e.** $[\frac{2}{3}, \frac{5}{6})$ **f.** $[\frac{5}{6}, 1]$

6. Laat X een normaal verdeelde stochast zijn met kansdichtheid

$$f(x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-1}{2}\right)^2}.$$

Dan is $\mathrm{E}\left[(X-3)^2\right]$ gelijk aan:

a. 1 **b.** 2 **c.** 4 **d.** 6 **e.** 8 **f.** 9

7. Stochasten X en Y zijn onafhankelijk, X heeft een $N(0,1)$ verdeling en Y een $N(1,3)$. Dan is $\mathrm{P}(0 \le X+Y \le 3)$ ongeveer

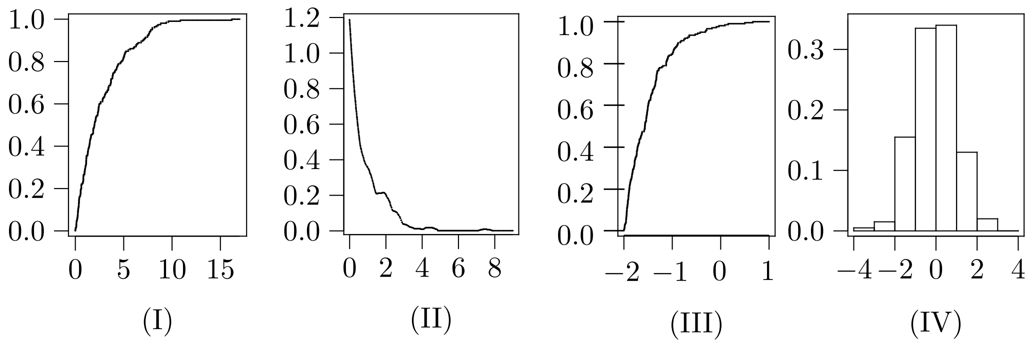
a. 0.533 **b.** 0.467 **c.** 0.631 **d.** 0.718 **e.** 0.841 **f.** 0.933

8. Bij een nieuwe PABX worden de ISDN toestellen en terminals gevoed vanuit de peripheral circuit boards. De voeding levert alleen vermogen aan een toestel als dat actief is. Er is opgegeven dat op elk moment van het drukke uur elk toestel met kans 0.1 actief is. Het aantal toestellen aan de PABX is 200. Om te bepalen of het vermogen van de voeding voldoende is, wil men weten hoeveel toestellen tegelijk actief zijn op een bepaald moment in het drukke uur. Ter vereenvoudiging van het probleem veronderstellen we dat het wel of niet actief zijn van de verschillende toestellen op een bepaald moment onafhankelijk van elkaar plaatsvindt. Zij X het aantal actieve toestellen op een bepaald moment in het drukke uur. Dan is de variantie van X gelijk aan

a. 14 **b.** 16 **c.** 18 **d.** 20 **e.** 22 **f.** 24

9. Onderstaand ziet u vier plaatjes van ofwel een histogram, ofwel een kernschatting, ofwel een empirische verdelingsfunctie. We weten dat één van de vier plaatjes correspondeert met een $Exp(1/3)$ óf een $Exp(15)$ verdeling. Welke is deze correspondentie?

a. $Exp(1/3)$ met (I) **b.** $Exp(1/3)$ met (II) **c.** $Exp(1/3)$ met (IV)
d. $Exp(15)$ met (I) **e.** $Exp(15)$ met (II) **f.** $Exp(15)$ met (III)



10. Voor het schatten van een succeskans p kunnen we kiezen uit twee zuivere schatters: $T_1 = X_1/50$ en $T_2 = X_2/100$, waarbij X_1 binomiaal verdeeld is met parameters $n = 100$ en $p/2$, en X_2 is $Bin(150, 2p/3)$ verdeeld. De relatieve efficiëntie (relative efficiency) van T_2 ten opzichte van T_1 is

a. $\frac{(2-p)}{(9-6p)}$ **b.** $\frac{(9-6p)}{(2-p)}$ **c.** $\frac{(2-p)}{(2-6p)}$ **d.** $\frac{(6-3p)}{(3-2p)}$ **e.** 2 **f.** 3

Open vragen

Toelichting: Een antwoord alleen is *niet* voldoende: er dient een berekening, toelichting en/of motivatie aanwezig te zijn. Dit alles goed leesbaar en in goed Nederlands.

1. De stochast X heeft de volgende kansdichtheid:

$$f(x) = \frac{1}{3}x + \frac{1}{6} \quad \text{als } 0 \leq x \leq 2,$$

en buiten dit interval geldt $f(x) = 0$.

a. Bepaal de verdelingsfunctie van X.

b. Bepaal de verwachting en de variantie van X.

c. Bepaal de mediaan van X.

2. Een fabriek maakt schakels voor zware metalen kettingen. De fabrikant laat 20 schakels opmeten en vindt de volgende lengtes in centimeters:

4.82	4.85	4.86	4.87	4.87
4.90	4.92	4.96	4.97	4.99
5.00	5.02	5.02	5.04	5.07
5.11	5.13	5.14	5.18	5.22

Het gemiddelde van deze data is 4.997 cm, de standaarddeviatie 0.118 cm.

a. Stel, er mag aangenomen worden dat de lengte van de schakels $N(\mu, \sigma^2)$ verdeeld is, met onbekende μ en σ. Geef een 95% betrouwbaarheidsinterval voor μ.

b. De fabrikant wil niet uitgaan van normaliteit en besluit om in plaats hiervan een bootstrapbetrouwbaarheidsinterval voor μ te construeren. Beschrijf nauwkeurig het bijbehorende bootstrapexperiment.

c. Het bootstrapexperiment is uitgevoerd met duizend runs. Een deel van de bootstrapuitkomsten is in de tabel weergegeven. Van de geordende lijst van uitkomsten zijn de nummers 21 t/m 60 en 941 t/m 980 gegeven. Bepaal hiermee een 95% bootstrapbetrouwbaarheidsinterval voor μ.

21–25	−2.202	−2.164	−2.111	−2.109	−2.101
26–30	−2.099	−2.006	−1.985	−1.967	−1.929
31–35	−1.917	−1.898	−1.864	−1.830	−1.808
36–40	−1.800	−1.799	−1.774	−1.773	−1.756
41–45	−1.736	−1.732	−1.731	−1.717	−1.716
46–50	−1.699	−1.692	−1.691	−1.683	−1.666
51–55	−1.661	−1.644	−1.638	−1.637	−1.620
56–60	−1.611	−1.611	−1.601	−1.600	−1.593
941–945	1.648	1.667	1.669	1.689	1.696
946–950	1.708	1.722	1.726	1.735	1.814
951–955	1.816	1.825	1.856	1.862	1.864
956–960	1.875	1.877	1.897	1.905	1.917
961–965	1.923	1.948	1.961	1.987	2.001
966–970	2.015	2.015	2.017	2.018	2.034
971–975	2.035	2.037	2.039	2.053	2.060
976–980	2.088	2.092	2.101	2.129	2.143

3. U probeert het smeltpunt van een nieuw materiaal te bepalen, waarbij u over een groot aantal monsters beschikt. Voor elk monster dat u meet vindt u een waarde die dicht ligt bij het werkelijke smeltpunt, c, maar verstoord is met een meetfout. We modelleren dit met de stochastische variabelen:

$$M_i = c + U_i$$

waar M_i is de gemeten waarde in graden Kelvin, en U_i is de optredende toevallige fout. Het is bekend dat $\mathrm{E}[U_i] = 0$ en $\mathrm{Var}(U_i) = 3$ voor alle i, en dat de stochastische variabelen $M_1, M_2, \ldots$ onderling onafhankelijk zijn. Hoeveel monsters heeft u nodig op basis van Chebyshev's ongelijkheid, opdat men 90% zeker is dat het gemiddelde van de metingen binnen een halve graad van c ligt?

4. Er wordt geklaagd over een bepaalde vierkeuze-vraag bij een tentamen. Volgens de studenten was deze vraag te moeilijk. De docent vindt een vraag te moeilijk als minder dan 25% van de studenten het antwoord echt wist. In totaal hebben 160 studenten de vraag gemaakt, waarvan 55 de vraag goed hebben beantwoord. Deze 55 zijn enerzijds studenten die het antwoord *wisten* en anderzijds studenten die het niet wisten en het goede antwoord *gegokt* hebben. Noem het aantal dat het juiste antwoord wist: w. We willen de hypothese $H_0 : w = 40$ toetsen tegen $H_1 : w < 40$ met als toetsingsgrootheid het aantal studenten X dat het juiste antwoord gaf.

a. Leg uit waarom onder de nulhypothese geldt: $X = 40+Y$, waarbij Y een $Bin(120, 0.25)$ verdeling heeft. Gebruik vervolgens de normale benadering om de p-waarde van de data (55 goede antwoorden) te bepalen. Bij een significantieniveau α van 5%, wat is uw conclusie over de nulhypothese?

b. Bepaal het kritieke gebied behorende bij deze toets met $\alpha = 0.05$ (gebruik weer de normale benadering!).

c. Stel dat in werkelijkheid 35 studenten het antwoord echt wisten (dus $w = 35$). Wat is de kans op een fout van type II? Mocht u vraag **b** niet hebben kunnen beantwoorden, gebruik dan als kritiek gebied $\{0, 1, \ldots, 60\}$.

Antwoorden multiple choice:

1 c.

$$\begin{aligned}\mathrm{P}(A \mid B \cup C) &= \frac{\mathrm{P}((A\cap B)\cup(A\cap C))}{\mathrm{P}(B\cup C)} = \frac{\mathrm{P}(A\cap B)+\mathrm{P}(A\cap C)}{\mathrm{P}(B)+\mathrm{P}(C)} \\ &= \frac{\mathrm{P}(A \mid B)\mathrm{P}(B)+\mathrm{P}(A \mid C)\mathrm{P}(C)}{\mathrm{P}(B)+\mathrm{P}(C)} = \frac{\frac{1}{2}\frac{1}{2}+\frac{2}{5}\frac{1}{4}}{\frac{1}{2}+\frac{1}{4}} = \frac{7}{15}\end{aligned}$$

2 c.

$$\mathrm{P}(O \mid A) = \frac{\mathrm{P}(A \mid O)\mathrm{P}(O)}{\mathrm{P}(A \mid O)\mathrm{P}(O)+\mathrm{P}(A \mid O^c)\mathrm{P}(O^c)} = \frac{\frac{3}{4}\frac{1}{500}}{\frac{3}{4}\frac{1}{500}+\frac{2}{10}\frac{499}{500}} = 0.0075$$

3 e. $\mathrm{E}[X]=\frac{1}{2}, \mathrm{E}[Y]=\mathrm{E}\left[X^2\right]=\frac{1}{3}, \mathrm{E}[XY]=\mathrm{E}\left[X^3\right]=\frac{1}{4}$. Dus $\mathrm{Cov}(X,Y)=\frac{1}{12}$.

4 c. Los op naar x: $F(x)=u$ voor $0\le u\le 1$, dan vind je $x=\sqrt{-0.05\ln(1-u)}$. Zie verder paragraaf 6.2.

5 e. Gevraagd wordt de kans dat $X>0.3$, en deze kan als volgt worden berekend:

$$\mathrm{P}(X>0.3)=\int_{0.3}^{1}\int_{0}^{1}\frac{2}{5}(x+4y)\,dy\,dx=\int_{0.3}^{1}\frac{2}{5}(x+2)\,dx=\frac{2}{5}(\frac{1}{2}x^2+2x)|_{x=0.3}^{x=1}=0.742$$

6 e. We lezen af dat $\mu=\mathrm{E}[X]=1$ en $\mathrm{Var}(X)=2^2=4$, zodat $\mathrm{E}\left[X^2\right]=4+1^2=5$

$$\mathrm{E}\left[(X-3)^2\right]=\mathrm{E}\left[X^2\right]-6\mu+9=5-6+9=8.$$

7 a. $S=X+Y$ heeft een $N(1,4)$ verdeling, dus $Z=\frac{S-1}{2}$ is $N(0,1)$, en we vinden

$$\mathrm{P}(0\le S\le 3)=\mathrm{P}\left(-\frac{1}{2}\le\frac{S-1}{2}\le\frac{3-1}{2}\right)=\mathrm{P}\left(-\frac{1}{2}\le Z\le 1\right)$$

en

$$\mathrm{P}\left(-\frac{1}{2}\le Z\le 1\right)=\mathrm{P}(Z\le 1)-\mathrm{P}\left(Z\le-\frac{1}{2}\right)=1-0.1587-0.3085=0.533$$

8 c. Na goed lezen zien we dat we moeten aannemen dat X een $Bin(200,0.1)$ verdeling heeft. Dus $\mathrm{Var}(X)=n\cdot p\cdot(1-p)=18$.

9 a. De verdeling die bij (III) neemt duidelijk ook negatieve waarden aan. (II) zou een kernschatting kunnen zijn bij een exponentiële verdeling, maar bij een $Exp(\frac{1}{3})$-verdeling zal de waarde in 0 ongeveer gelijk zijn aan $\frac{1}{3}e^0=\frac{1}{3}=0.333$, en bij een $Exp(15)$-verdeling zal de waarde in 0 ongeveer gelijk zijn aan $15e^0=15$. Verder is de verwachting (1/15) te klein voor (I) bij de $Exp(15)$-verdeling (Nog beter: bereken de mediaan, die is gelijk aan 0.046!). Het is dus $Exp(1/3)$ met (I).

10 d.

$$\mathrm{Var}(T_1)=\frac{\mathrm{Var}(X_1)}{2500}=\frac{100\cdot\frac{1}{2}p\cdot(1-\frac{1}{2}p)}{2500}=\ldots=\frac{p(2-p)}{100}.$$

Analoog:

$$\mathrm{Var}(T_2)=\frac{\mathrm{Var}(X_2)}{10000}=\frac{150\cdot\frac{2}{3}p\cdot(1-\frac{2}{3}p)}{10000}=\frac{p(3-2p)}{300}$$

Dus

$$\frac{\mathrm{Var}(T_1)}{\mathrm{Var}(T_2)}=\frac{p(2-p)/100}{p(3-2p)/300}=\frac{6-3p}{3-2p}.$$

Antwoorden open vragen:

1a Voor $x < 0$ geldt $F(x) = 0$, voor $x > 2$ dat $F(x) = 1$, en voor $0 \le x \le 2$:

$$F(x) = \int_0^x \left(\frac{1}{3}t + \frac{1}{6}\right) \mathrm{d}t = \left[\frac{1}{6}t^2 + \frac{1}{6}t\right]_0^x = \frac{1}{6}x^2 + \frac{1}{6}x$$

N.B. Ten controle vullen we $x = 0$ en $x = 2$ in en vinden respectievelijk 0 en 1, zoals moet.

1b We bepalen eerst de verwachting:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} t\, f(t)\, \mathrm{d}t = \int_0^2 t\, \left(\frac{1}{3}t + \frac{1}{6}\right) \mathrm{d}t = \left[\frac{1}{9}t^3 + \frac{1}{12}t^2\right]_0^2 = \frac{8}{9} + \frac{4}{12} = \frac{11}{9}.$$

Voor de variantie bepalen we eerst $\mathrm{E}\left[X^2\right]$:

$$\mathrm{E}\left[X^2\right] = \int_0^2 t^2\, \left(\frac{1}{3}t + \frac{1}{6}\right) \mathrm{d}t = \left[\frac{1}{12}t^4 + \frac{1}{18}t^3\right]_0^2 = \frac{16}{12} + \frac{8}{18} = \frac{16}{9}.$$

Dus $\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 = \frac{23}{81}$.

1c De mediaan is de oplossing van $F(m) = \frac{1}{2}$, ofwel $m^2 + m - 3 = 0$. De vergelijking heeft de wortels $m_{1,2} = -\frac{1}{2} \pm \frac{1}{2}\sqrt{13}$, de tweede ligt tussen 0 en 2, dus $m = -\frac{1}{2} + \frac{1}{2}\sqrt{13} = 1.303$. Het antwoord klopt met de intuitie: omdat de dichtheid stijgt moet de mediaan dichter bij 2 dan bij 0 liggen.

2a $T = \dfrac{\bar{X}_{20} - \mu}{0.118/\sqrt{20}} = \dfrac{\bar{X}_{20} - \mu}{0.0264}$ heeft een t_{19}-verdeling. De kritieke waarden bij $\alpha = 0.05$ zijn $\pm t_{19,0.025} = \pm 2.093$. Bij de realisatie $\bar{x}_{20} = 4.997$ wordt een 95% betrouwbaarheidsinterval voor μ: $(4.997 - 0.0264 \cdot 2.093,\ 4.997 + 0.0264 \cdot 2.093) = (4.9417,\ 5.0523)$.

2b Zie dictaat, § 23.3.

2c De bootstrap kritieke waarden zijn $c_\ell = t^*_{(25)} = -2.101$ en $c_r = t^*_{(976)} = 2.088$, Een betrouwbaarheidsinterval voor μ wordt nu $(4.997 - 0.0264 \cdot 2.088, 4.997 + 0.0264 \cdot 2.101) = (4.9419,\ 5.0524)$.

3 Dit is exact vraag 13.5 van de opgaven sectie van hoofdstuk 13 uit het boek.
Gevraagd wordt het aantal experimenten n zo, dat $P(|\bar{M} - c| \le 0.5) > 0.9$.
De oplossing maakt gebruik van Chebychev's ongelijkheid, die in dit geval als volgt kan worden uitgewerkt:

$$P(|\bar{M} - c| > 0.5) \le \frac{1}{0.5^2}\frac{\sigma^2}{n},$$

waarbij we aan de rechterkant reeds hebben ingevuld dat de variantie van het gemiddelde gelijk is aan de variantie van een meting gedeeld door het aantal metingen n.
Deze vergelijking kan nu omgeschreven worden naar de vereiste vorm van de vraag:

$$P(|\bar{M} - c| \le 0.5) > 1 - \frac{1}{0.5^2}\frac{\sigma^2}{n} = 1 - \frac{12}{n} = 0.9.$$

Uit het laatste volgt $n = 120$

4a Als 40 studenten 'het' weten, moeten $160 - 40 = 120$ studenten gokken, elk met kans 0.25 op succes. Het aantal goede antwoorden Y dat zo tot stand komt is dus $Bin(120, 0.25)$-verdeeld, en X is de som van 40 en Y. Onder H_0 geldt dus $\mathrm{P}(X \leq 55) = \mathrm{P}(Y \leq 15)$, met Y $Bin(120, 0.25)$-verdeeld. Normale benadering:

$$\begin{aligned}\mathrm{P}(Y \leq 15) &= \mathrm{P}\left(\frac{Y - 120 \cdot 0.25}{\sqrt{120 \cdot 0.25 \cdot (1 - 0.25)}} \leq \frac{15 - 120 \cdot 0.25}{\sqrt{120 \cdot 0.25 \cdot (1 - 0.25)}}\right) \\ &= \mathrm{P}(Z \leq -3.162) \approx 0.0008.\end{aligned}$$

Deze kans is (flink) kleiner dan α, dus H_0 wordt verworpen.

4b Er wordt eenzijdig getoetst: H_1: $w < 40$, dus alleen bij waarden van X die duiden op kleine w wordt verworpen, en dat zijn in dit geval ook kleine waarden van X. Want hoe minder studenten het weten, hoe minder goede antwoorden we zullen zien (ga na: de verwachting van X is $w + (160 - w) \cdot 0.25 = 40 + 0.75w$). We zoeken dus een linker kritieke waarde c zodat $\mathrm{P}(X \leq c) = \alpha$.
Onder H_0 heeft X bij benadering een normale verdeling (want dat geldt voor Y en $X = 40 + Y$), met verwachting 70 en variantie $120 \cdot 0.25 \cdot 0.75 = 22.5$, dus

$$\mathrm{P}(X \leq c) = \mathrm{P}\left(\frac{X - 70}{\sqrt{22.5}} \leq \frac{c - 70}{\sqrt{22.5}}\right) = \mathrm{P}\left(Z \leq \frac{c - 70}{\sqrt{22.5}}\right) = 0.05.$$

Uit de tabel van de normale verdeling vinden we:

$$\frac{c - 70}{\sqrt{22.5}} = -1.645$$

en dus $c = 70 - 1.645\sqrt{22.5} = 62.19$. De uitkomsten van X zijn geheeltallig, we moeten dus afronden, en wel *naar beneden* omdat anders de kans op een type I fout boven de 0.05 uitkomt. Het kritieke gebied voor X is derhalve: $\{0, 1, 2, \ldots, 62\}$.

4c Gevraagd wordt: $\mathrm{P}(X > 62 \,|\, w = 35) = \mathrm{P}(Y > 27)$, waarbij Y nu binomiaal verdeeld is met $n = 160 - 35$ en $p = 0.25$. Dezelfde aanpak als in onderdeel **a** geeft de normale benadering $1 - \mathrm{P}(Y \leq 27) \approx 1 - \mathrm{P}(Z \leq -0.878) = 1 - \mathrm{P}(Z \geq 0.878) \approx 0.81$.