Tentamen Statistische methoden
4052STAMEY
11 juli 2013, 9:00–12:00

Studienummers: Vult u alstublieft op het meerkeuzevragenformulier uw *Delftse* studienummer in (tbv automatische verwerking); 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. Gegeven zijn $\mathrm{P}(A) = \frac{1}{2}$, $\mathrm{P}(A \mid B) = \frac{4}{5}$ en $\mathrm{P}(A \mid B^c) = \frac{2}{5}$. Dan geldt $\mathrm{P}(B) =$

a. $\frac{1}{4}$ **b.** $\frac{3}{10}$ **c.** $\frac{1}{3}$ **d.** $\frac{3}{8}$ **e.** $\frac{2}{5}$ **f.** $\frac{6}{5}$

2. Groep 1 bestaat uit 20 mannen en 30 vrouwen. Groep 2 telt 45 mannen en 15 vrouwen. We kiezen een willekeurig persoon uit groep 1 en vervolgens een willekeurig persoon uit groep 2, in de hoop een paar te vormen van twee personen van hetzelfde geslacht. Wanneer het geslacht van beide personen verschillend is, sturen we ze weg en herhalen de procedure; beide groepen zijn nu elk 1 kleiner geworden. We tellen het aantal malen X dat we de procedure moeten herhalen totdat we een paar van hetzelfde geslacht gekozen hebben. Dan heeft X een

a. $Bin(2, \frac{11}{20})$ verdeling
b. $Geo(\frac{11}{20})$ verdeling
c. $Bin(45, \frac{11}{20})$ verdeling
d. discrete verdeling met uitkomsten $1, 2, \ldots, 46$
e. $Bin(50, \frac{11}{20})$ verdeling
f. discrete verdeling met uitkomsten $1, 2, \ldots, 50$

3. X en Y zijn onafhankelijke stochasten met kansverdelingen

x	0	1
$\mathrm{P}(X = x)$	1/2	1/2

resp.

y	-1	0	1
$\mathrm{P}(Y = y)$	1/4	1/4	1/2

De kansverdeling van $M = \max\{X, Y\}$ wordt dan gegeven door

a.

m	0	1
$\mathrm{P}(M = m)$	1/4	3/4

b.

m	-1	0	1
$\mathrm{P}(M = m)$	1/4	1/2	1/4

c.

m	-1	0	1
$\mathrm{P}(M = m)$	1/4	1/4	1/2

d.

m	-1	0	1
$\mathrm{P}(M = m)$	1/2	1/4	1/4

e.

m	0	1
$\mathrm{P}(M = m)$	1/3	2/3

f.

m	0	1
$\mathrm{P}(M = m)$	2/3	1/3

4. Stel W, X, Y en Z zijn onafhankelijke standaard normaal verdeelde stochasten. Dan is

$$\mathrm{P}(W + X + Y + Z \geq 1)$$

gelijk aan:

a. 0.5000 **b.** 0.4013 **c.** 0.3446 **d.** 0.3085 **e.** 0.1587 **f.** 0.0228

5. Laat X en Y twee stochasten zijn met $\mathrm{Var}(X) = 5$ en $\mathrm{Var}(Y) = 3$ en correlatie $+1$. Dan is de variantie van de stochast $3X - 2Y - 1$ ongeveer:

a. 8 **b.** 9 **c.** 11 **d.** 33 **e.** 57 **f.** 103

6. De stochasten $X_1, X_2, \ldots, X_{50}$ zijn onafhankelijke $Exp(5)$-verdeelde stochasten. Met behulp van de *Centrale Limietstelling* zien we dat de kans $\mathrm{P}(X_1 + X_2 + \cdots + X_{50} > 12)$ bij benadering gelijk is aan:

a. 0.01 **b.** 0.25 **c.** 0.04 **d.** 0.08 **e.** 0.11 **f.** 0.20

7. Beschouw de dataset: 1 2 2 3 3 4 4 4 4 4.
We trekken onafhankelijk van elkaar en met teruglegging 10 keer uit deze dataset. Wat is de kans dat de dataset die u zo krijgt precies zes keer een 4 bevat?

a. 0 **b.** 0.205 **c.** 0.324 **d.** 0.315 **e.** 0.4 **f.** 0.5

8. De MAD van de onderstaande *Challenger* dataset

53	57	58	63	66	67	67	67	68	69	70	70
70	70	72	73	75	75	76	76	78	79	81	

wordt gegeven door

a. 3 **b.** 4 **c.** 7 **d.** 13 **e.** 28 **f.** 70

9. Jan koopt hooi van een bepaald merk voor zijn konijnen. Hij denkt echter dat er gemiddeld te weinig hooi in de verpakking zit. Hij meet het gewicht van het hooi in 9 zakken precies op, en hij vindt $\bar{x}_n = 0.977$ kg en een steekproefstandaarddeviatie $s_n = 0.030$ kg. Op de verpakking staat dat er gemiddeld 1 kg in een zak zit. We modelleren het gewicht van het hooi als een $N(\mu, \sigma^2)$ verdeling. Als nul-hypothese nemen we $H_0 : \mu = 1$, en als alternatief $H_1 : \mu < 1$. Verder kiezen we een significantieniveau van 5%. Welke van de volgende is de uitkomst voor p en welke conclusie hoort daarbij?

a. $0.01 < p < 0.025$, verwerp H_0.
b. $0.01 < p < 0.025$, verwerp H_0 niet.
c. $0.025 < p < 0.05$, verwerp H_0.
d. $0.025 < p < 0.05$, verwerp H_0 niet.
e. $0.05 < p < 0.10$, verwerp H_0.
f. $0.05 < p < 0.10$, verwerp H_0 niet.

10. Een steekproef van omvang 5 uit een normale verdeling met onbekende μ en onbekende σ resulteert in de volgende dataset:

129.0 129.4 130.8 131.1 132.2

Het steekproefgemiddelde is 130.50 en de steekproefstandaarddeviatie is 1.3038.

Een symmetrisch tweezijdig 90%-betrouwbaarheidsinterval voor μ wordt gegeven door:

a. $129.11 < \mu < 131.89$
b. $129.26 < \mu < 131.74$
c. $129.33 < \mu < 131.67$
d. $129.54 < \mu < 131.46$
e. $129.61 < \mu < 131.39$
f. $129.75 < \mu < 131.25$

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. Er zijn uitstekende tests voor zwangere vrouwen op het syndroom van Down, maar onfeilbaar zijn ze niet. Ten eerste is bekend dat 1 procent van de embryo's lijdt aan het syndroom. Als een baby het syndroom heeft, dan is er 90 procent kans dat de testuitslag positief is, maar bij een gezonde baby is er 1 procent kans dat de testuitslag toch positief is. Bereken de kans dat een vrouw die positief test een baby met het Downsyndroom in zich draagt.

N.B. Een positieve testuitslag betekent: volgens de test is er sprake van het Downsyndroom.

2. De continue stochast X heeft als verdelingsfunctie F, met

$$F(x) = \begin{cases} 0 & \text{als } x \leq 0 \\ \sqrt{x} & \text{als } 0 < x < 1 \\ 1 & x \geq 1. \end{cases}$$

a. Bepaal $\mathrm{P}\left(\frac{1}{4} \leq X \leq \frac{9}{16}\right)$.

b. Bepaal $\mathrm{E}\left[3\sqrt{X} + 5\right]$.

c. Bepaal de verdeling van $\sqrt{X}$.

3. De hoeken van een gelijkbenige driehoek zijn θ, θ en γ. Ze worden elk apart gemeten (in radialen); dus $\theta + \theta + \gamma = \pi$ (rad). De metingen van de drie hoeken zijn X_1, X_2 en X_3. Gegeven is dat X_1, X_2 en X_3 onafhankelijk zijn en zuiver zijn voor respectievelijk θ, θ en γ, met variantie σ^2. We definiëren de volgende schatters voor θ:

$$S = \frac{1}{2}(X_1 + X_2) \quad \text{en} \quad T = \frac{1}{6}(2\pi + X_1 + X_2 - 2X_3)$$

a. Ga voor elk van de schatters na of hij zuiver is voor θ.

b. Bereken de variantie voor beide schatters en geef aan welke van de twee schatters u zou prefereren.

4. Men beschikt over data die een realisatie vormen van een steekproef $X_1, X_2, \ldots, X_{25}$ uit een $N(\mu, 1)$ verdeling. Men ontwerpt een toets voor $H_0 : \mu = 0$ tegen $H_1 : \mu > 0$ met toetsingsgrootheid $\bar{X}_n$, waar als beslissingsregel uitkomt: verwerp H_0 ten gunste van H_1 als voor de data geldt dat $\bar{x}_n \geq 0.3$.

a. Wat is het significantieniveau van de toets?

b. Wat is de kans op een type II fout als in werkelijkheid $\mu = 0.55$?

c. Is het mogelijk de steekproefgrootte n (in plaats van 25) zo te kiezen dat zowel het significantieniveau als de onder 4b bedoelde type II fout beide kleiner zijn dan 0.01? Indien ja, bepaalde kleinst mogelijke n; zo nee, leg uit waarom het niet mogelijk is.

5. Schets het principe van de bootstrap en leg heel precies het verschil uit tussen de parametrische en de empirische bootstrap.

Antwoorden multiple choice:

1 a. Via $\mathrm{P}(A) = \mathrm{P}(B)\,\mathrm{P}(A\,|\,B) + \mathrm{P}(B^c)\,\mathrm{P}(A\,|\,B^c) = \mathrm{P}(B)\,\mathrm{P}(A\,|\,B) + (1-\mathrm{P}(B))\mathrm{P}(A\,|\,B^c)$: $\frac{1}{2} = \frac{4}{5}\mathrm{P}(B) + \frac{2}{5}(1-\mathrm{P}(B)) = \frac{2}{5} + \frac{2}{5}\mathrm{P}(B)$, dus $\mathrm{P}(B) = (\frac{1}{2} - \frac{2}{5})/\frac{2}{5} = \frac{1}{4}$.

2 d. Nadat we de eerste keer een tweetal personen uit de groepen gekozen hebben is de samenstelling veranderd, want ze gaan niet weer terug. De kans op een succesvolle keuze is de tweede keer anders dan de eerste keer en hangt bovendien ook nog af van de eerste uitkomst. Om deze reden is zowel de binomiale als de geometrische verdeling uit te sluiten, want die zijn gebaseerd op een rij *onafhankelijke* 'experimenten' met 'constante' succeskans. Zo blijven d. en f. over en de vraag is dus hoe lang we maximaal bezig kunnen zijn. Daarvoor moet telkens bij een man uit groep 1 een vrouw uit groep 2 worden gekozen en omgekeerd. Voor de man-vrouw combinaties kan dit hoogstens 15 keer, want dan zijn de vrouwen in groep 2 op. Bij het omgekeerde is het aantal vrouwen in groep 1 de limiet: 30. Het kan dus maximaal $15 + 30 = 45$ maal fout gaan. Er zijn dan in beide groepen alleen mannen over, dus de 46ste keer is het dan raak.

3 a. Er geldt $\mathrm{P}(M=0) = \mathrm{P}(X=0, Y=-1) + \mathrm{P}(X=0, Y=0) = \mathrm{P}(X=0)\,\mathrm{P}(Y=-1) + \mathrm{P}(X=0)\,\mathrm{P}(Y=0) = 1/8 + 1/8 = 1/4$. Omdat M alleen de waarde 0 of 1 kan aannemen volgt direct dat $\mathrm{P}(M=1) = 1 - \mathrm{P}(M=0) = \frac{3}{4}$.

4 d. $S = W + X + Y + Z$ is normaal verdeeld met parameters $\mu = 0$ en $\sigma^2 = 1+1+1+1 = 4$. Dus $\mathrm{P}(S \geq 1) = \mathrm{P}(S/2 \geq 0.5) = 0.3085$.

5 c. Voor de variantie maakt de verschuiving over 1 niet uit: $\mathrm{Var}(3X - 2Y - 1) = \mathrm{Var}(3X - 2Y)$. Verder is

$$\mathrm{Var}(3X - 2Y) = \mathrm{Var}(3X) + \mathrm{Var}(2Y) - 2\mathrm{Cov}(3X, 2Y)\,.$$

Omdat X en Y correlatie 1 hebben, geldt er dat $\mathrm{Cov}(X,Y) = \rho \cdot \sigma_x \cdot \sigma_Y = \sqrt{5 \cdot 3}$, en dus dat $\mathrm{Cov}(3X, 2Y) = 6\mathrm{Cov}(X,Y) = 6\sqrt{15} = 23.24$. Alles tezamen vinden we:

$$\mathrm{Var}(3X - 2Y - 1) = 3^2\mathrm{Var}(X) + 2^2\mathrm{Var}(Y) + 12\mathrm{Cov}(X,Y) = 9{\cdot}5 + 4{\cdot}3 + 12\sqrt{15} \approx 45 + 12 - 46.48 = 10.52.$$

6 d. Uit het gegeven dat elke X_i een $Exp(5)$-verdeling heeft volgt direct dat $\mu = \mathrm{E}[X_i] = 1/5$ en $\sigma^2 = \mathrm{Var}(X_i) = 1/25$. Voor de som $S = X_1 + X_2 + \cdots + X_{50}$ geldt dus $\mathrm{E}[S] = 50 \cdot 1/5 = 10$ en $\mathrm{Var}(S) = 50 \cdot 1/25 = 2$. Er geldt dan bij benadering

$$\mathrm{P}(X_1 + X_2 + \cdots + X_{50} > 12) = \mathrm{P}(S > 12) \approx \mathrm{P}\Big(Z > \frac{12-10}{\sqrt{2}}\Big) = \mathrm{P}\Big(Z > \sqrt{2}\Big) \approx 0.0793.$$

7 b. Deze kans is $\binom{10}{6}(1/2)^{10} = 210/1024 = 0.205$.

8 b. De steekproefmediaan is het 12de getal in volgorde naar grootte: 70. De absolute afwijkingen t.o.v. 70 zijn

17	13	12	7	4	3	3	3	2	1	0	0
0	0	2	3	5	5	6	6	8	9	11	

of in volgorde naar grootte

0	0	0	0	1	2	2	3	3	3	3	4	5	5	6	6	7	8	9	11	12	13	17

De steekproefmediaan van deze absolute verschillen is 4.

9 c. De toetsingsgrootheid wordt gegeven door

$$t = \frac{\bar{x}_n - 1}{s_n/\sqrt{n}} = -2.3.$$

Er geldt dat $t_{8,0.05} < |t| < t_{8,0.025}$ (merk op dat T een t-verdeling heeft met $n-1=8$ vrijheidsgraden!). Aangezien alleen nog negatievere waarden van T wijzen op de alternatieve hypothese H_1, geldt dus dat $0.025 < p < 0.05$. Aangezien p kleiner is dan het significantieniveau, verwerpen we H_0.

10 b. De grenzen van het betrouwbaarheidsinterval worden gegeven door $\overline{x}_n \pm t_{n-1,a/2} s_n/\sqrt{n}$, met $\alpha = 0.1$, en $n = 5$. Dit geeft $130.50 \pm 2.132 \cdot 1.3038/\sqrt{5}$.

Antwoorden open vragen:

1 Voer in: D: een baby heeft Down, en T: de testuitslag is positief. De gegevens zijn: $\mathrm{P}(D) = 0.01$, $\mathrm{P}(T \,|\, D) = 0.9$ en $\mathrm{P}(T \,|\, D^c) = 0.01$. Gevraagd wordt $\mathrm{P}(D \,|\, T) = \dfrac{\mathrm{P}(D \cap T)}{\mathrm{P}(T)}$. $\mathrm{P}(D \cap T) = \mathrm{P}(D)\,\mathrm{P}(T \,|\, D) = 0.009$, en $\mathrm{P}(T) = \mathrm{P}(D \cap T) + \mathrm{P}(D^c \cap T) =$
$0.009 + \mathrm{P}(D^c)\,\mathrm{P}(T \,|\, D^c) = 0.009 + 0.99 \cdot 0.01 = 0.0189$.
Op elkaar delen geeft het antwoord: $\mathrm{P}(D \,|\, T) = 0.009/0.0189 = 0.4762$.

2a Omdat X een continue stochast is geldt dat

$$\mathrm{P}\left(\tfrac{1}{4} \le X \le \frac{9}{16}\right) = F(\tfrac{9}{16}) - F(\tfrac{1}{4}) = \sqrt{\tfrac{9}{16}} - \sqrt{\tfrac{1}{4}} = \tfrac{3}{4} - \tfrac{1}{2} = \frac{1}{4}.$$

2b Omdat de kansdichtheid $f(x)$ van X gelijk is aan $F'(x)$, geldt er dat:

$$f(x) = \begin{cases} \frac{1}{2\sqrt{x}} & \text{als } 0 < x < 1 \\ 0 & \text{als } x \le 0 \text{ of } x \ge 1. \end{cases}$$

Maar dan geldt er wegens de "Change of variable formula" dat

$$\mathrm{E}\left[\sqrt{X}\right] = \int_{-\infty}^{\infty} \sqrt{x} f(x)\,\mathrm{d}x = \int_0^1 \sqrt{x}\frac{1}{2\sqrt{x}}\,\mathrm{d}x = \int_0^1 \frac{1}{2}\,\mathrm{d}x = \frac{1}{2}.$$

2c Schrijf $Y = \sqrt{X}$. Omdat $0 \le X \le 1$ gelden dezelfde grenzen voor Y. Voor $0 \le y \le 1$ geldt dus:

$$\mathrm{P}(Y \le y) = \mathrm{P}\Big(\sqrt{X} \le y\Big) = \mathrm{P}\big(X \le y^2\big) = \sqrt{y^2} = y.$$

Dit betekent dat Y een $U(0,1)$-verdeling heeft.

3a Omdat X_1 en X_2 beide zuiver zijn voor θ, is $\mathrm{E}[X_1] = \theta$ en $\mathrm{E}[X_2] = \theta$. Dit betekent dat

$$\mathrm{E}[S] = \frac{1}{2}(\mathrm{E}[X_1] + \mathrm{E}[X_2]) = \frac{1}{2}(\theta + \theta) = \theta.$$

Dus S is zuiver voor θ. Op een zelfde manier, gebruikmakend van het feit dat $\pi = 2\theta + \gamma$, is

$$\begin{aligned} \mathrm{E}[T] &= \frac{1}{6}(2\pi + \mathrm{E}[X_1] + \mathrm{E}[X_2] - 2\mathrm{E}[X_3]) \\ &= \frac{1}{6}(2\pi + \theta + \theta - 2\gamma) \\ &= \frac{1}{6}(2(2\theta + \gamma) + \theta + \theta - 2\gamma) \\ &= \theta. \end{aligned}$$

Dus ook T is zuiver voor θ.

3b Voor S krijgen we (vanwege de onafhankelijkheid van X_1 en X_2)

$$\mathrm{Var}(S) = \mathrm{Var}\left(\frac{1}{2}(X_1 + X_2)\right) = \frac{1}{4}\big[\mathrm{Var}(X_1) + \mathrm{Var}(X_2)\big] = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}.$$

Voor T krijgen we

$$\mathrm{Var}(T) = \mathrm{Var}\left(\frac{1}{6}(2\pi + X_1 + X_2 - 2X_3)\right) = \frac{1}{36}\big[\mathrm{Var}(X_1)+\mathrm{Var}(X_2)+4\mathrm{Var}(X_3)\big] = \frac{1}{36}(\sigma^2+\sigma^2+4\sigma^2) = \frac{\sigma^2}{6}.$$

Omdat beide schatters zuiver zijn geven we de voorkeur aan degene met de kleinste variantie, dus aan T.

4a Merk op dat $\bar{X}_{25}$ een $N(\mu, 1/25)$ verdeling heeft. De kans op een type I fout is daarom gelijk aan

$$\mathrm{P}\big(\bar{X}_n \geq 0.3 \mid \mu = 0\big) = \mathrm{P}\left(\frac{\bar{X}_n}{1/5} \geq \frac{0.3}{1/5}\right) = \mathrm{P}(Z \geq 1.5) \approx 0.0668.$$

4b Dit is

$$\mathrm{P}\big(\bar{X}_n < 0.3 \mid \mu = 0.55\big) = \mathrm{P}\left(\frac{\bar{X}_n - 0.55}{1/5} < \frac{0.3 - 0.55}{1/5}\right) = \mathrm{P}(Z < -1.25) = \mathrm{P}(Z > 1.25) \approx 0.1056.$$

4c Wanneer we de berekeningen bij de vorige onderdelen overdoen voor steekproefgrootte n, waarvoor geldt dat $\bar{X}_n$ een $N(\mu, 1/n)$ verdeling heeft, dan vinden we voor de type I fout $\mathrm{P}(Z \geq 0.3 \cdot \sqrt{n})$ en voor de type II fout $\mathrm{P}(Z \leq -0.25 \cdot \sqrt{n}) \equiv \mathrm{P}(Z \geq 0.25 \cdot \sqrt{n})$. De tweede kans is altijd groter; het is dus voldoende te zorgen dat die onder de 0.01 blijft. In de tabel vinden we $z_{0.01} = 2.326$, dus er moet gelden: $0.25 \cdot \sqrt{n} \geq 2.326$ ofwel $n \geq 86.56$. Derhalve: $n = 87$ is de kleinste waarde die aan de eisen voldoet.

5 Dit refereert aan hoofdstuk 18 in het boek van Dekking et al (2005). Het principe van de bootstrap (pagina 270) is toe te passen op elke steekproefgrootheid $h(X_1, X_2, \ldots, X_n)$, afgeleid van een *random sample* $X_1, X_2, \ldots, X_n$:

1. Bepaal uit de gegeven data $x_1, x_2, \ldots, x_n$ een schatting $\hat{F}$ van de verdelingsfunctie F.
2. Vervang het random sample $X_1, X_2, \ldots, X_n$ door het (bootstrap) random sample $X_1^*, X_2^*, \ldots, X_n^*$ dat wordt getrokken uit $\hat{F}$.
3. Benader de verdeling van $h(X_1, X_2, \ldots, X_n)$ door die van $h(X_1^*, X_2^*, \ldots, X_n^*)$.

In de praktijk wordt het principe vaak toegepast door een groot aantal trekkingen van $h(X_1^*, X_2^*, \ldots, X_n^*)$ met behulp van simulatie uit te voeren, en hiermee de verdeling van $h(X_1, X_2, \ldots, X_n)$ te schatten. Soms is deze stap overbodig omdat de verdeling van $h(X_1^*, X_2^*, \ldots, X_n^*)$ uit $\hat{F}$ te bepalen is.
Het verschil tussen de parametrische en empirische bootstrap wordt bepaald door de wijze waarop de schatting van de verdelingsfunctie, $\hat{F}$, tot stand komt. Bij de parametrische bootstrap wordt de verdeling of de verdelingsfunctie gegeven met eventueel meerdere parameters. Deze parameters moeten uit de bestaande steekproef geschat worden. Bij de empirische bootstrap is deze informatie niet beschikbaar en maken we gebruik van de empirische verdelingsfunctie direct afgeleid uit de data. Dit laatste geval is equivalent aan het trekken van een bootstrap monster uit de bestaande data, met teruglegging.