Tentamen Statistische methoden
MST-STM
8 april 2010, 9:00–12:00

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. Je doet mee aan een quiz-show en als winnaar mag je kiezen uit dozen A, B, C, D en E. In één ervan zit een cheque van 25 000 Euro, de andere zijn leeg. Je kiest doos A. De quizmaster (die weet wat in de dozen zit) opent vervolgens doos C en E, die leeg zijn. Dan vraagt hij of je liever doos B of D wilt hebben. Je kans om te winnen, wanneer je doos D kiest, is:

a. $\frac{1}{5}$ **b.** $\frac{1}{3}$ **c.** $\frac{2}{5}$ **d.** $\frac{2}{3}$ **e.** $\frac{4}{5}$ **f.** niet te berekenen

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. Het 95ste percentiel van de $Exp(0.25)$-verdeling is ongeveer

a. 3.46 **b.** 4.79 **c.** 5.62 **d.** 7.64 **e.** 9.23 **f.** 11.98

4. Voor het uitvoeren van een simulatiestudie is het nodig te simuleren aan de hand van de volgende verdelingsfunctie: $F(x) = 1 - \mathrm{e}^{-0.05x^2}$ 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. $X = 1 - \mathrm{e}^{-0.05U^2}$ **b.** $X = \mathrm{e}^{-0.05U^2}$ **c.** $X = -20 \ln U$
d. $X = -0.05 \ln U$ **e.** $X = \sqrt{-20 \ln U}$ **f.** $X = \sqrt{-\ln(20U)}$

5. Een boer vult zakken met aardappelen. Stel het vulgewicht heeft een verdeling met verwachting 30 kilogram en standaardafwijking 500 gram. Hoe groot is de kans dat een partij van 100 zakken minder dan 2990 kilo aardappelen bevat?

a. ≤ 0.001 **b.** 0.0062 **c.** 0.0228 **d.** 0.0668 **e.** 0.1336 **f.** 0.4207

6. Jongens hebben ongeveer kans 1 op 10 om kleurenblind te zijn, meisjes ongeveer kans 1 op 200. De variantie van het aantal kleurenblinde leerlingen in een klas met 20 jongens en 20 meisjes is ongeveer gelijk aan

a. 0.90 **b.** 1.90 **c.** 2.34 **d.** 2.80 **e.** 4.24 **f.** 5.48

7. Om de slaagkans van een experiment te onderzoeken wordt door elk van honderd personen een serie experimenten uitgevoerd tot ze voor het eerst succes hebben. Een model om dit te beschrijven is: het aantal pogingen benodigde N heeft een $Geo(p)$-verdeling. We willen p schatten met de maximum likelihood methode. De gegevens:

aantal pogingen	1	2	3	≥ 4
frequentie	40	30	15	15

De likelihoodfunctie wordt dan

a. $p^{85}(1-p)^{15}$ **b.** $p^{85}(1-p)^{105}$ **c.** $p^{85}(1-p)^{60}$
d. $p^{100}(1-p)^{105}$ **e.** $p^{100}(1-p)^{60}$ **f.** $p^{100}(1-p)^{15}$

8. We beschikken over een dataverzameling $x_1, x_2, \ldots, x_n$ met gemiddelde $\bar{x}_n$. De dataverzameling is een realisatie van een steekproef $X_1, X_2, \ldots, X_n$ uit een $Exp(\lambda)$ verdeling. De steekproefgrootheid $1/\bar{X}_n$ is een schatter voor λ. Men wil de kansverdeling van de stochast $T_n = 1/\bar{X}_n - \lambda$ benaderen door middel van een bootstrap simulatie. Met wat voor soort bootstrap procedure hebben we hier te maken en wat is de bootstrapversie van T_n?

a. een empirische bootstrap met $T_n^* = 1/\bar{X}_n^* - 1/\bar{x}_n$.
b. een empirische bootstrap met $T_n^* = 1/\bar{X}_n^* - \bar{x}_n$.
c. een empirische bootstrap met $T_n^* = \bar{X}_n^* - 1/\bar{x}_n$.
d. een parametrische bootstrap met $T_n^* = 1/\bar{X}_n^* - 1/\bar{x}_n$.
e. een parametrische bootstrap met $T_n^* = 1/\bar{X}_n^* - \bar{x}_n$.
f. een parametrische bootstrap met $T_n^* = \bar{X}_n^* - 1/\bar{x}_n$.

9. Voor een toets doen we 10 metingen, die normaal verdeeld zijn, met onbekende verwachting μ en variantie $\sigma^2 = 10$. De nulhypothese stelt: $\mu = 0$; de alternatieve: $\mu < 0$. Als toetsingsgrootheid gebruiken we het gemiddelde van de dataset en als kritieke gebied het interval $(-\infty, -2]$. Als in werkelijkheid geldt $\mu = -1$, dan is de kans op een fout van de tweede soort ongeveer:

a. 0.15 **b.** 0.35 **c.** 0.65 **d.** 0.75 **e.** 0.85 **f.** 0.95

10. Een bioloog onderzoekt het effect van een magnesiumrijk dieet op de sekse van het nageslacht van bepaalde zoogdieren. De nulhypothese is $H_0 : p = 0.5$, de alternatieve $H_1 : p > 0.5$, waarbij p de kans op een vrouwtje is bij het volgen van het dieet. In een steekproef van 83 werden 48 vrouwtjes geboren. De p-waarde die hierbij hoort is ongeveer (je mag de normale benadering gebruiken)

a. 0.02 **b.** 0.05 **c.** 0.09 **d.** 0.12 **e.** 0.15 **f.** 0.25

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. Gegeven zijn de onafhankelijke stochasten $X_1, \ldots, X_4$, met $Exp(3)$ verdeling.

a. Geef de verdelingsfunctie van $Z = \max(X_1, \ldots, X_4)$.
b. Bepaal de kansdichtheid van $Y = \mathrm{e}^{2X_1}$.

2. De stochasten X en Y hebben de volgende kansdichtheid:

$$f(x,y) = \frac{4}{5}(x + xy + y), \quad 0 \leq x \leq 1, 0 \leq y \leq 1$$

en buiten dit gebied geldt $f(x, y) = 0$.

a. Bepaal de marginale kansdichtheid van X.

b. Gegeven is $\mathrm{E}[X] = \frac{3}{5}$. Bepaal $\mathrm{E}[3X - 1]$ en $\mathrm{Var}(3X - 1)$.

c. Gegeven is bovendien dat $\mathrm{E}[Y] = \mathrm{E}[X]$. Bepaal $\mathrm{Cov}(X, Y)$.

3. De hoeken van een gelijkbenige driehoek zijn θ, θ en γ. (Dus $\theta + \theta + \gamma = \pi$ (rad).) Ze worden elk apart gemeten (in radialen); 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. Hieronder is een dataset met levensduren gegeven. We willen een betrouwbaarheidsinterval voor de verwachte levensduur μ maken. Enkele kentallen van de dataset: $\bar{x} = 40.79$, $s = 22.09$, $n = 43$, eerste en derde kwartiel: respectievelijk 27.77 en 51.13.

4.91	6.15	7.75	16.91	23.51	23.55	24.10	24.15	24.61
24.90	27.77	28.54	28.63	28.99	29.58	30.39	30.63	30.76
32.04	33.84	34.26	34.34	38.25	38.82	40.92	41.19	41.61
41.76	44.70	46.17	49.53	50.43	51.13	52.11	52.45	53.84
60.23	61.52	67.58	79.66	89.92	99.08	102.60		

a. Maak (een schets van) de boxplot van deze dataset. Leg uit hoe de verschillende posities/afmetingen van de boxplot bepaald worden, en geef de bijbehorende numerieke waarden voor de dataset.

b. Doe even alsof de data afkomstig is van een normale verdeling (beide parameters onbekend) en bepaal het 95% betrouwbaarheidsinterval voor μ.

c. Voer met behulp van onderstaande bootstrapresultaten de t-toets uit voor $H_0 : \mu = 48$ versus $H_1 : \mu < 48$, bij significantieniveau 0.04, geef ook de p-waarde van de data. Er is een gestudentiseerde bootstrap met 1000 herhalingen is uitgevoerd; hieronder een deel van de naar grootte geordende bootstrap-uitkomsten.

21–25	−2.367	−2.363	−2.359	−2.327	−2.324
26–30	−2.281	−2.263	−2.216	−2.200	−2.199
31–35	−2.180	−2.173	−2.170	−2.161	−2.133
36–40	−2.114	−2.071	−2.056	−2.043	−2.026
41–45	−2.021	−2.010	−2.002	−1.995	−1.985
46–50	−1.981	−1.979	−1.965	−1.923	−1.903
51–55	−1.899	−1.896	−1.885	−1.863	−1.847
56–60	−1.842	−1.838	−1.836	−1.780	−1.765
941–945	1.406	1.442	1.495	1.507	1.519
946–950	1.527	1.534	1.537	1.538	1.544
951–955	1.564	1.580	1.584	1.599	1.604
956–960	1.615	1.615	1.618	1.620	1.640
961–965	1.657	1.660	1.660	1.666	1.684
966–970	1.690	1.698	1.711	1.720	1.739
971–975	1.746	1.765	1.784	1.786	1.803
976–980	1.822	1.828	1.839	1.892	1.896

Antwoorden multiple choice:

1 c. De kans dat het doos A is, is $\frac{1}{5}$. De overige dozen hebben samen kans $\frac{4}{5}$, ook nadat de quizmaster er twee lege van geopend heeft. Op grond van symmetrie volgt dan voor beide dozen $\frac{2}{5}$. Zie ook paragraaf 1.3.

2 c. Benoem de gebeurtenissen als volgt. O: "er is olie"; A: "apparaat zegt dat er olie is." We weten $\mathrm{P}(O) = 1/500$, $\mathrm{P}(A \,|\, O) = 3/4$, en $\mathrm{P}(A \,|\, O^c) = 1/5$. Vervolgens:

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

3 f. Op te lossen: $F(q) = 1 - \mathrm{e}^{-0.25q} = 0.95$, waaruit $q = 4 \ln 20 \approx 11.98$ volgt.

4 e. Los op naar x: $F(x) = u$ voor $0 \leq u \leq 1$, dan vind je $x = \sqrt{-20 \ln u}$. Zie verder paragraaf 6.2.

5 c. Laat G het totale gewicht zijn. Op grond van de centrale limietstelling is $G = X_1 + X_2 + \ldots + X_{100}$ ongeveer normaal verdeeld met $\mu = 100 \cdot 30 = 3000$ en $\sigma^2 = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_{100}) = 100 \cdot 0.5^2 = 5^2$. Dus geldt

$$\mathrm{P}(G \leq 2990) = \mathrm{P}\left(\frac{G - 3000}{5} \leq \frac{2990 - 3000}{5}\right) = \mathrm{P}(Z \leq -2) = \mathrm{P}(Z \geq 2) \approx 0.0228.$$

6 b. Zowel het aantal kleurenblinde jongens X als het aantal kleurenblinde meisjes Y heeft een binomiale verdeling met $n = 20$, met $p = 0.1$ resp. $p = 0.005$. Dus $\mathrm{Var}(X) = 1.8$ en $\mathrm{Var}(Y) = 0.0995$. We vinden dus $\mathrm{Var}(X + Y) = 1.8 + 0.0995 \approx 1.9$, waarbij we aannemen dat X en Y onafhankelijk zijn.

7 b. $\mathrm{P}(N = k) = (1-p)^{k-1}p$, en $\mathrm{P}(X \geq 4) = (1-p)^3$ (nl. als er de eerste drie keer geen succes is).

$$L(p) = p^{40} \cdot [(1-p)p]^{30} \cdot [(1-p)^2 p]^{15} \cdot [(1-p)^3]^{15} = \ldots = p^{85}(1-p)^{105}$$

8 d. De kansverdeling van de X_i's is van een parametrisch type: een $Exp(\lambda)$ verdeling. Zodoende hebben we te maken met een parametrische bootstrap. De parameter λ schatten we door $\hat{\lambda} = 1/\bar{x}_n$, en dus is de bootstrap versie van T_n gelijk aan de stochast $T_n^* = 1/\bar{X}_n^* - \hat{\lambda} = 1/\bar{X}_n^* - 1/\bar{x}_n$.

9 e. Onder de aanname heeft de toetsingsgrootheid T een $N(-1, 1)$ verdeling ($\sigma^2/10 = 1$). We maken een fout van de tweede soort als we nu de nulhypothese *niet* verwerpen, dus als $T > -2$. $\mathrm{P}(T > -2) = \mathrm{P}(Z > -1)$ voor een standaardnormale Z, en dus gelijk aan $1 - \mathrm{P}(Z > 1) = 1 - 0.1587 = 0.8413$.

10 c. Te bepalen $\mathrm{P}(X \geq 48)$ voor X met een $Bin(83, 0.5)$ verdeling. Het exacte antwoord is 0.0937, met de normale benadering: 0.0768.

Antwoorden open vragen:

1a Zie §8.4: $F_Z(z) = 0$ voor $z < 0$; $F_Z(z) = (1 - \mathrm{e}^{-3z})^4$, voor $z \geq 0$.

1b Omdat $X_1 \geq 0$ geldt $Y \geq 1$. Dus $f_Y(y) = 0$ voor $y < 1$. Voor $y \geq 1$ vinden we $F_Y(y) = \mathrm{P}\big(\mathrm{e}^{2X_1} \leq 1\big) = \mathrm{P}\big(X_1 \leq \frac{1}{2}\ln y\big) = 1 - y^{-3/2}$. Dus voor deze waarden geldt $f_Y(y) = \frac{3}{2}y^{-5/2}$.

2a We krijgen de marginale kansdichtheid door de andere variabele 'eruit te integreren':

$$f_X(x) = \int_0^1 \frac{4}{5}(x + xy + y)\,\mathrm{d}y = \left[\frac{4}{5}\left(xy + \frac{1}{2}xy^2 + \frac{1}{2}y^2\right)\right]_0^1 = \frac{2}{5}(3x+1), \quad 0 \le x \le 1;$$

buiten dit gebied geldt $f_X(x) = 0$.
Een alternatieve weg naar dit antwoord loopt via de simultane verdelingsfunctie (nogal omslachtig, maar niet fout): voor a en b tussen 0 en 1 geldt

$$F(a,b) = \int_{x=0}^{a}\int_{y=0}^{b} \frac{4}{5}(x + xy + y)\,\mathrm{d}y\,\mathrm{d}x = \int_0^a \frac{2}{5}\left(2xb + xb^2 + b^2\right)\,\mathrm{d}x = \frac{2}{5}\left(a^2b + \frac{1}{2}a^2b^2 + b^2a\right).$$

We vinden nu de marginale verdelingsfunctie van X gemakkelijk, namelijk $F_X(a) = F(a,\infty) = F(a,1)$, en differentieren die om f_X te vinden:

$$f_X(a) = \frac{\mathrm{d}}{\mathrm{d}a}F(a,1) = \frac{2}{5}(3a+1), \quad 0 \le a \le 1;$$

buiten dit gebied geldt nog steeds $f_X(a) = 0$.

2b We weten dat $\mathrm{E}[3X-1] = 3\mathrm{E}[X] - 1$ en $\mathrm{E}[X]$ is uit de definitie te bepalen:

$$\mathrm{E}[X] = \int_0^1 x\frac{2}{5}(3x+1)\,\mathrm{d}x = \left[\frac{2}{5}\left(x^3 + \frac{1}{2}x^2\right)\right]_0^1 = \frac{3}{5}.$$

Er volgt: $\mathrm{E}[3X-1] = 3\cdot\frac{3}{5} - 1 = \frac{4}{5}$. Voor $\mathrm{Var}(3X-1)$ bepalen we eerst $\mathrm{Var}(X)$ via $\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2$ en vervolgens gebruiken we $\mathrm{Var}(3X-1) = 9\,\mathrm{Var}(X)$:

$$\mathrm{E}\left[X^2\right] = \int_0^1 x^2\frac{2}{5}(3x+1)\,\mathrm{d}x = \left[\frac{2}{5}\left(\frac{3}{4}x^4 + \frac{1}{3}x^3\right)\right]_0^1 = \frac{13}{30}.$$

$$\mathrm{Var}(X) = \left(\frac{13}{30} - \frac{9}{25}\right) = \frac{11}{150} \quad \text{en} \quad \mathrm{Var}(3X-1) = 9\cdot\mathrm{Var}(X) = 9\cdot\frac{11}{150} = \frac{33}{50}.$$

2c We bepalen

$$\mathrm{E}[XY] = \int_{x=0}^1\int_{y=0}^1 xyf(x,y)\mathrm{d}y\,\mathrm{d}x = \int_{x=0}^1\int_{y=0}^1 xy\frac{4}{5}(x+xy+y)\,\mathrm{d}y\,\mathrm{d}x = \frac{4}{5}\left[\frac{1}{6} + \frac{1}{9} + \frac{1}{6}\right] = \frac{16}{45}.$$

De covariantie vinden we nu als volgt: $\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] = \frac{16}{45} - \left(\frac{3}{5}\right)^2 = -\frac{1}{225} \approx -0.00444$.

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)

$$\begin{aligned}\mathrm{MSE}(S) &= \mathrm{Var}(S) + (\mathrm{E}\,[S] - \theta)^2 \\ &= \mathrm{Var}(S) = \mathrm{Var}\left(\frac{1}{2}(X_1 + X_2)\right) \\ &= \frac{1}{4}\big[\mathrm{Var}(X_1) + \mathrm{Var}(X_2)\,\big] = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}.\end{aligned}$$

Voor T krijgen we

$$\begin{aligned}\mathrm{MSE}(T) &= \mathrm{Var}(T) + (\mathrm{E}\,[T] - \theta)^2 \\ &= \mathrm{Var}(T) = \mathrm{Var}\left(\frac{1}{6}(2\pi + X_1 + X_2 - 2X_3)\right) \\ &= \frac{1}{36}\big[\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 4\mathrm{Var}(X_3)\,\big] \\ &= \frac{1}{36}(\sigma^2 + \sigma^2 + 4\sigma^2) = \frac{\sigma^2}{6}.\end{aligned}$$

We zien dat de mean squared error van T het kleinst is en derhalve is T te prefereren boven S.

4a De box loopt van 27.77 (eerste kwartiel) tot 51.13 (derde kwartiel). Bij 34.34 loopt de streep door de box die de mediaan markeert. De linker snorhaar eindigt bij 4.91, de kleinste data-waarde. De rechter snorhaar zou tot 86.2 mogen lopen en eindigt dus bij 79.66 (de grootste waarde daaronder), de drie grootste waarden worden dus apart gemarkeerd. Zie verder Chapter 16.

4b Omdat ook de variantie van de normale verdeling onbekend is, moet die geschat worden en maken we een betrouwbaarheidsinterval op basis van de t-verdeling (van het gestudentiseerde gemiddelde). Het aantal vrijheidsgraden is 42, we gebruiken dus $t_{42,0.025} \approx 2.019$ (door interpolatie in de tabel). We vinden dus $40.79 \pm 2.019 \cdot 22.09/\sqrt{43} = 40.79 \pm 2.019 \cdot 3.369 = 40.79 \pm 6.80 = (33.99, 47.59)$.

4c We vinden $t = (40.79 - 48)/(22.09/\sqrt{43}) = -2.1403$, hetgeen tussen de 34ste en 35ste bootstrapwaarde in ligt. De p-waarde is dus ongeveer 0.034. We verwerpen, al zou ik de bootstrapsimulatie nog eens doen met, zeg, 10 000 herhalingen, voor een nauwkeuriger bepaling.