Tentamen Statistische methoden
MST-STM
14 april 2011, 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. Stel C en D zijn onafhankelijke gebeurtenissen met $P(C) = 0.5$ en $P(C \cup D) = 0.7$. Dan is $P(D)$ gelijk aan

a. 0.2 **b.** 0.5 **c.** 0.4 **d.** 0.6 **e.** 0.3 **f.** 0.35

2. Jongens hebben ongeveer kans 1 op 10 om kleurenblind te zijn, meisjes ongeveer kans 1 op 200. Van een leerlingenlijst van een klas met 50 jongens en 20 meisjes kiezen we willekeurig een naam: F. Galema. De docent meldt dat F. Galema kleurenblind is. Gegeven deze informatie is de kans dat F. Galema een jongen is ongeveer

a. 0.100 **b.** 0.714 **c.** 0.952 **d.** 0.980 **e.** 0.990 **f.** 0.999

3. Het 63ste percentiel van de $Par(0.17)$-verdeling is ongeveer:

a. 2.34 **b.** 6.4 **c.** 15 **d.** 28 **e.** 134 **f.** 347

4. Stel $N_1, N_2, \ldots, N_{100}$ zijn onafhankelijke Poisson-verdeelde stochastische variabelen, met verwachting en variantie beide gelijk aan 2, en Z is een standaard normale variabele. Volgens de centrale limietstelling is $\mathrm{P}(X_1 + X_2 + \ldots + X_{100} \leq 210)$ ongeveer gelijk aan:

a. $\mathrm{P}(Z \leq -1/2)$ **b.** $\mathrm{P}(Z \leq -1/20)$ **c.** $\mathrm{P}(Z \leq 1/200)$
d. $\mathrm{P}(Z \leq 1/20)$ **e.** $\mathrm{P}(Z \leq 1/2)$ **f.** $\mathrm{P}\big(Z \leq 1/\sqrt{2}\big)$

5. Een lijmfabrikant heeft een gestandaardiseerde test voor het bepalen van de uithardingstijd van lijm. Onder de testcondities zijn de uithardingstijden bij benadering normaal verdeeld en onafhankelijk. Er worden 17 tests gedaan voor lijmtype *White Rabbit*, met als uitkomsten een gemiddelde uithardingstijd van 54.0 seconden en een standaarddeviatie van 20.4. Het 95% betrouwbaarheidsinterval voor de verwachte uithardingstijd is:

a. $43.5 < \mu < 64.5$ **b.** $44.3 < \mu < 63.7$ **c.** $43.6 < \mu < 64.4$
d. $45.9 < \mu < 62.1$ **e.** $45.4 < \mu < 62.6$ **f.** $51.5 < \mu < 56.5$

6. Een bank heeft een portfolio van uitstaande leningen, 50 van 10 miljoen Euro en 50 van 2 miljoen. De kans op *default*, dwz niet terugbetalen, is 1% respectievelijk 2% per lening, en defaults worden geacht onafhankelijk van elkaar op te treden (een twijfelachtige aanname). Dan zijn de variantie van L, het te leiden verlies in miljoenen, en N, het aantal wanbetalers, repectievelijk

a. 6.91 en 1.390 **b.** 6.91 en 1.475 **c.** 50.48 en 1.390
d. 50.48 en 1.475 **e.** 53.42 en 1.390 **f.** 53.42 en 1.475

7. Van de stochastische variabele X is gegeven dat $\mathrm{E}[X] = 1$ en $\mathrm{Var}(X) = 4$. Van Y is gegeven dat $\mathrm{E}\left[Y^2\right] = 4$. Dan worden $\mathrm{E}\left[X^2\right]$ en $\mathrm{E}\left[-2X + 3Y^2\right]$ gegeven door

a. $\mathrm{E}\left[X^2\right] = 4$ en $\mathrm{E}\left[-2X + 3Y^2\right] = 6$
b. $\mathrm{E}\left[X^2\right] = 3$ en $\mathrm{E}\left[-2X + 3Y^2\right] = 6$
c. $\mathrm{E}\left[X^2\right] = 5$ en $\mathrm{E}\left[-2X + 3Y^2\right] = 10$
d. $\mathrm{E}\left[X^2\right] = 4$ en $\mathrm{E}\left[-2X + 3Y^2\right] = 10$
e. $\mathrm{E}\left[X^2\right] = 3$ en $\mathrm{E}\left[-2X + 3Y^2\right] = 10$
f. $\mathrm{E}\left[X^2\right] = 5$ en $\mathrm{E}\left[-2X + 3Y^2\right] = 6$

8. We toetsen een hypothese H_0 op significantieniveau 0.05. Dit betekent dat:

a. als H_0 onjuist is, de kans op verwerpen van H_0 minstens 0.95 is.
b. als H_1 onjuist is, de kans op verwerpen van H_1 minstens 0.95 is.
c. als H_0 juist is, de kans op niet verwerpen van H_0 minstens 0.95 is.
d. als H_1 juist is, de kans op niet verwerpen van H_1 minstens 0.95 is.
e. als H_0 onjuist is, de kans op verwerpen van H_0 minstens 0.05 is.
f. als H_0 juist is, de kans op niet verwerpen van H_0 minstens 0.05 is.

9. Men trekt een getal X uit een uniforme verdeling op het interval $[0, \theta]$. Men toetst $H_0 : \theta = 2$ tegen $H_1 : \theta \neq 2$ en verwerpt H_0 ten gunste van H_1 als $X \leq 0.1$ of als $X \geq 1.9$. Als $\theta = 4$, dan is de kans op een type II fout gelijk aan

a. 0.2 **b.** 0.45 **c.** 0.475 **d.** 0.525 **e.** 0.55 **f.** 0.8

10. De simultane kansverdeling van twee stochasten X en Y is gegeven door de volgende (onvolledige) tabel:

			X		
		-1	0	1	
Y	2		1/4	0	
	4				
			3/4		1

Als $\mathrm{E}[X] = \frac{1}{8}$ en $\mathrm{E}[Y] = \frac{7}{2}$, dan is $\mathrm{Cov}(X, Y)$ aan:

a. $-\frac{11}{16}$ **b.** $-\frac{1}{2}$ **c.** $-\frac{1}{4}$ **d.** 0 **e.** $\frac{1}{16}$ **f.** $\frac{1}{2}$

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. Vijf vriendinnen (Ada, Bea, Cora, Dora en Eva) gaan lootjes trekken.

a. Bereken de kans dat Eva zichzelf trekt als gegeven is dat Ada zichzelf *niet* trekt.

b. Bereken de kans dat precies drie van de vijf vriendinnen zichzelf trekken.

2. Gegeven is de stochastische variabele X met als dichtheid

$$f(x) = \begin{cases} \frac{3}{4}(1 - x^2) & -1 \leq x \leq 1; \\ 0 & \text{elders.} \end{cases}$$

en als verdelingsfunctie $F(x)$.

a. Bereken $F(1/2)$.

b. Bereken $\mathrm{E}[X]$.

3. Gegevens uit het zwangerschapsonderzoek van Weinberg en Gladen, waaraan in totaal 474 vrouwen meededen.
Regel 1: perioden tot zwanger worden
Regel 2: aantal (rokende) vrouwen die zoveel perioden nodig hadden

perioden	1	2	3	4	5	6	7	8	9	10	11	12
frequentie f_i	198	107	55	38	18	22	7	9	5	3	6	6

We nemen aan: elke vrouw heeft elke maand opnieuw dezelfde kans om zwanger te geraken. Daaruit volgt dat Y, het aantal perioden dat een vrouw nodig heeft om zwanger te geraken, een geometrische verdeling heeft met parameter p.

a. We schatten p via de relatieve frequentie:

$$\hat{p} = \frac{\text{aantal vrouwen dat in 1 maand zwanger was}}{\text{totaal aantal vrouwen}}$$

en we noemen de bijbehorende schatter T_1. Toon aan dat T_1 zuiver is voor p.

b. Hoe kun je met behulp van de gevonden waarde van $\hat{p}$ de kans $\mathrm{P}(Y > 2)$ schatten? Geef de waarde van de schatting en noem de bijbehorende schatter T_2.
Geef de functie g zodat $T_2 = g(T_1)$.

Net als p hierboven kan $\mathrm{P}(Y > 2)$ geschat worden door een relatieve frequentie. Noem de hier relevante schatter S_2.

c. Ga van beide schatters voor $\mathrm{P}(Y > 2)$ na of ze zuiver zijn.

4. Bekijk de 5 resultaten van een meting: 12.53, 12.56, 12.47, 12.67 en 12.48. Hierbij wordt aangenomen dat de onderliggende verdeling normaal is. Het lijkt er op dat de 4^e meting een bijzonder punt, een uitbijter (Engels: *outlier*), kan zijn.

a. Geef aan hoe zo'n uitbijter te identificeren.

b. Harris[1] stelt een Q-test voor, met als toetsingsgrootheid $Q = g/R$, waarbij g de 'gap' is tussen uitbijter en dichtstbijzijnde punt (hier $g = 12.67 - 12.56$) en R de range. De gerealiseerde waarde: $Q = 0.55$. Geef de preciese definitie van de p-waarde voor deze data en hoe je met de p-waarde de toets uitvoert.

c. Beschrijf hoe je een parametrische bootstrap voor Q uitvoert en hoe je met behulp van de uitkomsten de p-waarde hierboven kunt bepalen.

[1] De data zijn uit: Harris, Quantitative chemical analysis, 2007, p. 65.

Antwoorden multiple choice:

1 c. Gebruik: $\mathrm{P}(C)+\mathrm{P}(D) = \mathrm{P}(C \cup D)+\mathrm{P}(C \cap D)$ en, vanwege de onafhankelijkheid $\mathrm{P}(C \cap D) = \mathrm{P}(C)\,\mathrm{P}(D)$. Invullen van de gegeven waarden levert

$$0.5 + \mathrm{P}(D) = 0.7 + 0.5\mathrm{P}(D)$$

waaruit eenvoudig volgt dat $\mathrm{P}(D) = 0.4$.

2 d. Dit is een Bayes' som. Definieer de volgende gebeurtenissen: K: de leerling is kleurenblind; J: het is een jongen; M: een meisje. Dan geldt $\mathrm{P}(J \cap K) = 50/700$, $\mathrm{P}(M \cap K) = 1/700$ en dus $\mathrm{P}(J \mid K) = 50/51 \approx 0.98$.

3 f. Los op: $1 - t^{-\alpha} = 0.63$ voor $\alpha = 0.17$.

4 f. Verwachting en variantie van de som zijn beide 200, dus met standaardiseren geeft de CLS:

$$\mathrm{P}(X_1 + X_2 + \ldots + X_{100} \leq 210) \approx \mathrm{P}\left(Z \leq \frac{210 - 200}{\sqrt{200}}\right).$$

5 a. De juiste t-waarde: $t(16, 0.025) = 2.12$, zodat het betrouwbaarheidsinterval wordt: $54.0 \pm 2.12\frac{20.4}{\sqrt{17}} = 54.0 \pm 10.5 = (43.5, 64.5)$.

6 f. Als X het aantal defaults is van 10 miljoen en Y dat van 2 miljoen, dan $X \sim \mathit{Bin}(50, 0.01)$ en $Y \sim \mathit{Bin}(50, 0.02)$, onafhankelijk. Er geldt $L = 10X + 2Y$ en $N = X + Y$ zodat $\mathrm{Var}(L) = 100\mathrm{Var}(X) + 4\mathrm{Var}(Y)$ en $\mathrm{Var}(N) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ waaruit via de formules van de binomiale verdeling volgen $\mathrm{Var}(L) = 53.42$ en $\mathrm{Var}(N) = 1.475$.
Op het MST-STM tentamen van 14 april 2011 kiest 68.5% van de deelnemers antwoord **b**, blijkbaar denkt men dat $\mathrm{Var}(aX + bY) = a\mathrm{Var}(X)+b\mathrm{Var}(Y)$, ondanks hameren van de docenten op de regel aan het eind van §7.4...

7 c. We weten dat $\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2$, waaruit volgt dat

$$\mathrm{E}\left[X^2\right] = \mathrm{Var}(X) + (\mathrm{E}[X])^2 = 4 + 1^2 = 5.$$

Verder volgt uit de lineariteit van de verwachting dat

$$\mathrm{E}\left[-2X + 3Y^2\right] = -2\mathrm{E}[X] + 3\mathrm{E}\left[Y^2\right] = 10.$$

8 c. Het significantieniveau is de maximaal toelaatbare kans op een type I fout, waarbij een type I fout overeenkomt met verwerpen van H_0, terwijl deze in werkelijkheid juist is. Als het significantieniveau 0.05 is, dan betekent dit dus dat in het geval dat H_0 juist is, de kans op verwerpen hoogstens 0.05 en dus is de kans op niet verwerpen minstens 0.95.

9 b. De kans op een type II fout bij $\theta = 4$ is

$$\mathrm{P}(\text{niet verwerpen} \mid \theta = 4) = \mathrm{P}(0.1 < X < 1.9 \mid \theta = 4) = \frac{1.9 - 0.1}{4} = 0.45.$$

10 e. $\mathrm{E}[Y] = 3\frac{1}{2}$ impliceert $\mathrm{P}(Y = 2) = \frac{1}{4}$. De marginale kansen van X volgen eveneens uit de verwachting. Daarna is de tabel makkelijk in te vullen:

		X			
		-1	0	1	
Y	2	0	1/4	0	1/4
	4	1/16	1/2	3/16	3/4
		1/16	3/4	3/16	1

$\mathrm{E}[XY] = \frac{1}{2}$ volgt, en $\mathrm{Cov}(X, Y) = \frac{1}{16}$.

Antwoorden open vragen:

1a $\mathrm{P}(E \,|\, A^c) = \dfrac{\mathrm{P}(E \cap A^c)}{\mathrm{P}(A^c)} = \dfrac{\mathrm{P}(E)\,\mathrm{P}(A^c \,|\, E)}{\mathrm{P}(A^c)} = \dfrac{1/5 \cdot 3/4}{4/5} = \dfrac{3}{16}$, want als gegeven is dat E zich voordoet, dan volgt dat Ada met gelijke kansen A,B, C of D trekt.

1b Een situatie met trekken *zonder* teruglegging, waarbij er dus geen onafhankelijkheid is! Voor elke drietal, zeg A, B en E, is de kans dat zij zichzelf trekken en de andere twee niet zichzelf trekken gelijk aan $\frac{1}{5} \cdot \frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{120}$. Er zijn $\binom{5}{3} = 10$ drietallen, dus de kans dat precies een drietal zichzelf trekt is gelijk aan $\frac{10}{120} = \frac{1}{12}$.

2a Volg de definitie van verdelingsfunctie, in dit geval:

$$F(1/2) = \frac{3}{4}\int_{-1}^{1/2} (1-x^2)\mathrm{d}x = \frac{3}{4}\left[x - \frac{1}{3}x^3\right]_{x=-1}^{1/2} = \frac{81}{96}.$$

Bij het MST-STM tentamen van 14 april 2011 maakt slechts 30% van de deelnemers deze vraag voldoende...

2b $\mathrm{E}[X] = \int_{-1}^{1} x f(x)\mathrm{d}x = 0$. Men ziet overigens direct, dat de dichtheid symmetrisch is om 0.

3a Voor de definitie van zuiverheid zie Hoofdstuk 19. De kans om in een maand zwanger te zijn is volgens het model: $\mathrm{P}(Y = 1) = p$, dus

$$T_1 = \frac{\text{aantal } Y_i \text{ gelijk aan } 1}{474}; \quad \text{schatting } \hat{p} = \frac{198}{474} \approx 0.42.$$

Het aantal vrouwen X met $Y_i = 1$ heeft (natuurlijk!) een $\mathrm{Bin}(n, p)$ verdeling. Dus $T_1 = \frac{1}{n}X$, en dan uiteraard $\mathrm{E}[T_1] = \frac{1}{n}\mathrm{E}[X] = \mathrm{P}(Y_i = 1) = p$.

3b Er geldt $\mathrm{P}(Y > 2) = (1-p)^2$, dus dit kan worden geschat met $(1-\hat{p})^2 \approx 0.34$ en $T_2 = g(T_1) = (1-T_1)^2$.

3c Met hetzelfde argument als in **a** volgt dat S_2 zuiver is voor $\mathrm{P}(Y > 2)$. Voor de gevonden g geldt $g'(x) = 2 > 0$, dus g is convex. De ongelijkheid van Jensen geeft dan dat $\mathrm{E}[T_2] = \mathrm{E}[g(T_1)] > g(\mathrm{E}[T_1]) = g(p) = (1-p)^2$, dus $\mathrm{E}[T_2] > (1-p)^2 = \mathrm{P}(Y > 2)$, waarmee is aangetoond dat T_2 een positieve bias heeft t.o.v. $\mathrm{P}(Y > 2)$.

4a Gebruik de boxplot: als het punt buiten de snorharen ligt dan kan men het als zodanig aanmerken. Nadere beschrijving volgens §16.4 van het boek.

4b De p-waarde is hier de kans dat Q groter is dan de gevonden 0.55 indien Q bepaald wordt voor een steekproef van 5 trekkingen uit een normale verdeling met parameters geschat uit de data.[2] We verwerpen de hypothese dat dit punt geen uitbijter is, als de p-waarde te klein is. Zie §25.2 van het boek.

4c Bepaal gemiddelde en standaardafwijking van de 5 metingen en gebruik deze voor μ en σ voor de normale verdeling in de parametrische bootstrap. Genereer een (bootstrap)dataset van 5 onafhankelijke trekkingen uit deze verdeling en evalueer hiervoor de toetsingsgrootheid Q. Hier is een keuze te maken, namelijk of je alleen de grootste waarde als potentiele uitbijter beschouwt, of ook de kleinste. Wij kijken alleen naar de grootste. Herhaal de procedure, zeg, 10000 maal. Het histogram van de bootstrapwaarden $q_1^*, \ldots, q_{10000}^*$ geeft een schatting van de verdeling van de toetsingsgrootheid, als H_0 waar is. De p-waarde is de fractie uitkomsten die groter is dan 0.55. Wij vonden 0.1025 (naar beide kanten kijkend: 0.2072).

[2] De verdeling van Q hangt overigens hier niet van af.