Tentamen Statistische methoden
MST-STM
27 juni 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. Voor de gebeurtenissen C en D geldt: C en D^c zijn disjunct (Engels: disjoint), $\mathrm{P}(C) = 0.5$ en $\mathrm{P}(C \cup D) = 0.7$. Dan is $\mathrm{P}(C \cap D)$ gelijk aan

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

2. Een communicatiesysteem bestaat uit twee verbindingen en werkt alleen als beide verbindingen functioneren. Verbinding 1 heeft een faalkans van 1/10, verbinding 2 faalkans 4/10 en het falen gebeurt onafhankelijk van elkaar. Gegeven dat het systeem faalt, wat is de kans dat verbinding 1 gefaald heeft?

a. 0.10 **b.** 0.20 **c.** 0.22 **d.** 0.40 **e.** 0.46 **f.** 0.5

3. Een tennisspeler heeft een kans van 7% dat hij een dubbele fout slaat, en zo het punt verliest, onafhankelijk van de daarvoor gespeelde punten. Wat is de kans dat hij bij 15 punten, precies 5 dubbele fouten slaat?

a. 0.244% **b.** 1.05% **c.** 3.33% **d.** 5.21% **e.** 18.1% **f.** 33.3%

4. Het zuurstofgehalte van de lucht na afloop van een groot college wordt 16 keer gemeten. De fabrikant van de klimaatinstallatie garandeert dat de standaardafwijking van het gemeten zuurstofgehalte gelijk is aan 1%, en dat het zuurstofgehalte goed gemodelleerd wordt door een normale verdeling. Het gemiddelde van de 16 metingen is gelijk aan 19.5%. Geef een 95% betrouwbaarheidsinterval voor het verwachte zuurstof gehalte aan het eind van een groot college.

a. $19.5 \pm z_{0.025} \cdot 1/\sqrt{15}$ **b.** $19.5 \pm t_{0.05,15} \cdot 1/\sqrt{15}$ **c.** $19.5 \pm t_{0.05,15} \cdot 0.25$
d. $19.5 \pm t_{0.025,15} \cdot 0.25$ **e.** $19.5 \pm z_{0.05} \cdot 0.25$ **f.** $19.5 \pm z_{0.025} \cdot 0.25$

5. Een laborante bekijkt 100 kweekjes, en telt het aantal besmettingshaarden per kweekje. Zij vindt gemiddeld 3.64 besmettingshaarden per kweekje, met een steekproefstandaarddeviatie van 2. Noem μ het verwachte aantal besmettingshaarden per kweekje. De laborante voert een t-toets uit voor de nulhypothese $H_0 : \mu = 4$, tegen de alternatieve hypothese $H_1 : \mu \neq 4$, bij significantieniveau 0.05. Geef de waarde van de toetsingsgrootheid en vermeldt uw conclusie.

a. $t = -0.18$; H_0 verwerpen **b.** $t = -0.18$; H_0 niet verwerpen
c. $t = -1.22$; H_0 verwerpen **d.** $t = -1.22$; H_0 niet verwerpen
e. $t = -1.80$; H_0 verwerpen **f.** $t = -1.80$; H_0 niet verwerpen

6. Als X een exponentiële verdeling heeft met parameter 2, dan is de verdelingsfunctie van $Y = 1 - 3X$ gelijk aan

a. $F(y) = \begin{cases} 1 - e^{-2(1-y)/3} & y < 1 \\ 1 & y \geq 1 \end{cases}$ **b.** $F(y) = \begin{cases} 1 - \frac{1}{3}e^{-2(1-y)} & y < 1 \\ 1 & y \geq 1 \end{cases}$

c. $F(y) = \begin{cases} 0 & y < 0 \\ 3e^{-2y} & y \geq 0 \end{cases}$ **d.** $F(y) = \begin{cases} e^{-2(1-y)/3} & y < 1 \\ 1 & y \geq 1 \end{cases}$

e. $F(y) = \begin{cases} 0 & y < 0 \\ 1 - 3e^{-2y} & y \geq 0 \end{cases}$ **f.** $F(y) = \begin{cases} 1 - \frac{2}{3}e^{-2(1-y)/3} & y < 1 \\ 1 & y \geq 1 \end{cases}$

7. Laat $X_1, X_2, \ldots, X_{100}$ een rij van onafhankelijke, identiek verdeelde stochasten zijn met verwachting 1 en variantie 4. Laat $Y = X_1 + X_2 + \ldots + X_{100}$. De benadering van $\mathrm{P}(Y < 110)$ volgens de centrale limietstelling is:

a. 0.0228 **b.** 0.1587 **c.** 0.3085 **d.** 0.6915 **e.** 0.8413 **f.** 0.9872

8. Op een bingo-avond speel je telkens aan het volgende spel mee totdat je een keer de hoofdprijs wint. Welk type kansverdeling beschrijft het aantal keer dat je meespeelt het beste?

a. Pareto **b.** Binomiaal **c.** Geometrisch
d. Poisson **e.** Exponentieel **f.** Ander type verdeling

9. Stel X heeft een uniforme verdeling op het interval $[-\mu, \mu]$ waarbij μ onbekend is. De nulhypothese is dat $\mu = 5$, de alternatieve hypothese dat $\mu < 5$. Iemand voert een toets uit door drie onafhankelijke trekkingen X_1, X_2 en X_3 te doen en neemt als toetsingsgrootheid T het maximum van X_1, X_2 en X_3. Hij besluit H_0 te verwerpen ten gunste van H_1 als $T \leq 3$. Als de werkelijke waarde van μ gelijk is aan 4 dan is de kans op een fout van de tweede soort ongeveer

a. 0.12 **b.** 0.44 **c.** 0.28 **d.** 0.33 **e.** 0.53 **f.** 0.21

10. X en Y zijn onafhankelijk en beide normaal verdeeld: $X \sim N(3, 16)$ en $Y \sim N(-3, 23)$. Dan is $\mathrm{P}(X - 2Y \leq 5)$ ongeveer gelijk aan:

a. 0.15 **b.** 0.25 **c.** 0.35 **d.** 0.45 **e.** 0.55 **f.** 0.65

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 hoofdprijs van de populaire televisiequiz *Ik baal van ballen* wordt bepaald door blind te trekken uit een bak met vijftien genummerde ballen: 5,4,4,3,3,3,2,2,2,2 en 1,1,1,1,1. Elke bal is zijn nummer in duizenden euro's waard. De winnaar trekt (zonder teruglegging) zoveel ballen als hij/zij wil. Er is een *maar*: als twee ballen met hetzelfde nummer getrokken zijn, dan en is het spel uit en vervallen alle inkomsten. Dus: trek je 3 en 4 en stop je, dan win je 7000 euro, trek je 2, 4 en 4, dan win je niks.

a. Als iemand van tevoren besluit te stoppen na de tweede bal, wat is de kans dat hij toch met lege handen naar huis gaat?

b. Noem de prijs bij laatstgenoemde strategie X. Bepaal de kansverdeling van X.
Hint: Maak een tabel gebaseerd op de mogelijke combinaties.

Een kandidaat kan het al of niet doorgaan natuurlijk af laten hangen van de nummers die reeds getrokken zijn. Bijvoorbeeld door na elke getrokken bal kijken wat de verwachte 'meeropbrengst' is bij het trekken van een volgende bal, en stoppen zodra deze negatief is.

c. Bereken de verwachte meeropbrengst na het trekken van een bal 4 en een bal 5.

2. In een bepaalde reactie worden twee stoffen gevormd, X en Y, met opbrengsten in molen van θ en ρ. Uit het mengsel kunnen we met behulp van een analyse techniek θ bij benadering bepalen; geef de uitkomst hiervan aan met (de schatter) S. Analoog vinden we T, die ρ schat. De ("random") meetfouten die bij deze bepalingen optreden zijn onafhankelijk van elkaar, en T is twee maal zo nauwkeurig als S, dat wil zeggen, $\mathrm{Var}(S) = 4\sigma^2$ en $\mathrm{Var}(T) = \sigma^2$, waarbij σ een vast getal is.

a. Is er een verschil tussen de uitspraken "S en T zijn onafhankelijk" en "$\mathrm{Cov}(S, T) = 0$"? Leg uit.

Uit de massabalans volgt dat moet gelden $2\theta + 3\rho = 5$. Door meetfouten hoeft natuurlijk niet te gelden dat $2S + 3T = 5$. Er is een massadefect $D = 5 - 2S - 3T$. Door het defect te verdelen over de schatters krijgen we (hopelijk) een betere bepaling. Voor de eenvoud concentreren we ons op θ. Definieer de gecorrigeerde schatter $S_c = S + aD$, waarbij a een nog te kiezen constante is.

b. Bepaal $\mathrm{Var}(D)$.

c. Gegeven is dat S en T zuivere schatters zijn. Leg uit wat dit betekent en laat zien dat S_c zuiver is als schatter voor θ, ongeacht de waarde van a.

d. Bepaal $\mathrm{Var}(S_c)$ voor $a = 0.2$. Wat is voor deze S_c de efficientie ten opzichte van S?

e. Voor welke a heeft S_c de grootste efficientie? Leg uit.

3. Bekijk de 5 resultaten van een meting: 12.53, 12.56, 12.47, 12.82 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. Harris[1] stelt een toets voor, met als toetsingsgrootheid $Q = \delta/s$, waarbij δ het verschil is tussen de uitbijter en het gemiddelde (hier $\delta = 12.82 - (12.53 + 12.56 + 12.47 + 12.82 + 12.48)/5$) en s de standaarddeviatie. De gerealiseerde waarde is hier $Q = 1.73$. Geef de precieze definitie van de p-waarde voor deze data en hoe je met de p-waarde de toets uitvoert.

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

c. Bij uitvoering van de vorige deelvraag wordt $p = 0.036$ gevonden. De toets had ook uitgevoerd kunnen worden met de mediaan in plaats van het gemiddelde en de MAD in plaats van de standaardafwijking. We vinden dan $Q = 2.92$ en $p = 0.058$.

Welke redenen zijn er in het algemeen om de mediaan ipv het gemiddelde en de MAD ipv de standaardafwijking te gebruiken? Welke van de twee toetsen zou u gebruiken en waarom?

[1] De data zijn uit: Harris, Quantitative chemical analysis, 2007, p. 65, die deze uitbijter-toets in zijn boek de Q-test noemt.

Antwoorden multiple choice:

1 b. Als C en D^c disjunct zijn, dan geldt $C \subset D$ en $C \cap D = C$.

2 c. Noem F_i de gebeurtenis "verbinding i faalt", dus $\mathrm{P}(F_1) = 0.1$, $\mathrm{P}(F_2) = 0.1$. Er volgt voor F, "het systeem faalt", $F = F_1 \cup F_2$, en dus $\mathrm{P}(F) = 0.1 + 0.4 - 0.1 \cdot 0.4 = 0.46$, waarbij we de algemene somregel gebruiken en het product volgt uit de onafhankelijkheid van F_1 en F_2. Gevraagd is $\mathrm{P}(F_1 \,|\, F)$ en omdat $F_1 \subset F$ is dit precies $\mathrm{P}(F_1)\,/\mathrm{P}(F) = 0.1/0.46 = 5/23 \approx 0.22$.

3 a. Het aantal fouten heeft een $Bin(15, 0.07)$-verdeling, de kans is dus $\binom{15}{5}0.07^5 0.93^{10} = 0.00244$.

4 d. Met 16 trekkingen uit de normale verdeling halen we de kritieke waarde uit de $t(15)$ verdeling. Hier is $s/\sqrt{n} = 1/4 = 0.25$.

5 f. We gebruiken de t-toets, met 100 kweekjes zitten we in het "large sample" regime. We vinden $t = (3.64 - 4)/(2/\sqrt{100}) = -1.80$. De kritieke waarden zijn $-z_{0.025} = -1.96$ en $+1.96$. De gevonden waarde ligt hiertussen, dus niet in het kritieke gebied. Niet verwerpen.

6 d. Als $y \in (-\infty, 1)$,

$$P(Y \leq y) = P(1 - 3X \leq y) = P(3X \geq 1 - y) = P(X \geq (1 - y)/3) = e^{-2(1-y)/3}.$$

7 d.

$$\mathrm{P}(Y < 110) = \mathrm{P}\left(\frac{Y - 100}{\sqrt{100 \cdot 4}} \leq \frac{10}{\sqrt{100 \cdot 4}}\right) \approx \Phi(0.5) \approx 0.69.$$

8 c. Discrete stochast dus b of c of d. Wachten tot een gebeurtenis optreedt, dus c.

9 d. Gezocht is $\mathrm{P}(T > 3 \,|\, \mu = 4) = 1 - \mathrm{P}(X_1 \leq 3, X_2 \leq 3, X_3 \leq 3)\, \mu = 4 = 1 - (7/8)^3 = 169/512 \approx 0.33$.

10 c. Uit de eigenschappen van de normale verdeling volgt dat $W = X - 2Y$ ook een normale verdeling heeft, met $\mathrm{E}[W] = 9$ en $\mathrm{Var}(W) = \mathrm{Var}(X) + 4\mathrm{Var}(Y) = 108$. Er volgt $\mathrm{P}(W \leq 5) = \mathrm{P}\big((W - 9)/\sqrt{108} \leq (5 - 9)/\sqrt{108}\big) = \mathrm{P}(Z \leq -0.3849) \approx 0.35$ (met Z standaardnormaal).

Antwoorden open vragen:

1a Dit is de kans op twee gelijke trekkingen:

$$\frac{2}{15}\frac{1}{14} + \frac{3}{15}\frac{2}{14} + \frac{4}{15}\frac{3}{14} + \frac{5}{15}\frac{4}{14} = \frac{40}{210} = \frac{4}{21}$$

1b De kanstabel voor de te winnen prijs X (eenheid: 1000 Euro)[2]:

k	0	3	4	5	6	7	8	9
$\mathrm{P}(X = k)$	$\frac{40}{210}$	$\frac{40}{210}$	$\frac{30}{210}$	$\frac{44}{210}$	$\frac{26}{210}$	$\frac{20}{210}$	$\frac{6}{210}$	$\frac{4}{210}$

1c De kans dat het bij de derde bal fout gaat is $\frac{1}{13}$, de meeropbrengst is dan -9; met kans $\frac{4}{13}$ wordt er een 2 getrokken wat meeropbrengst 2 geeft; enzovoort. De kanstabel voor de meeropbrengst W (eenheid: 1000 Euro) na een bal 4 en een bal 5:

[2] Voor het opschrijven en .. het nakijken .. is het gemakkelijk als altijd dezelfde noemer wordt gebruikt, in dit geval 210.

k	-9	3	2	1
$\mathrm{P}(W=k)$	$\frac{1}{13}$	$\frac{3}{13}$	$\frac{4}{13}$	$\frac{5}{13}$

De verwachte meeropbrengst:

$$\sum_k k \cdot \mathrm{P}(W=k) = (-9)\cdot\frac{1}{13} + 3\cdot\frac{3}{13} + 4\cdot\frac{2}{13} + 1\cdot\frac{5}{13} = \frac{13}{13} = 1$$

2a Zie §10.2.

2b Omdat S en T onafhankelijk zijn geldt $\mathrm{Var}(D) = \mathrm{Var}(2S) + \mathrm{Var}(3T) = 2^2\mathrm{Var}(S) + 3^2\mathrm{Var}(T) = 25\sigma^2$.

2c Voor de definitie, zie §19.3. Er geldt $\mathrm{E}[S_c] = \mathrm{E}[S] + a\mathrm{E}[D]$, vanwege de lineariteit van de verwachting (§10.1). Voorts $\mathrm{E}[D] = 5 - 2\mathrm{E}[S] - 3\mathrm{E}[T] = 5 - 2\theta - 3\rho = 0$ (massabalans), zodat $\mathrm{E}[S_c] = \mathrm{E}[S] = \theta$ voor alle a, wat de zuiverheid bewijst.

2d Twee mogelijkheden. Eerste: Uitgaande van $\mathrm{Var}(S+aD) = \mathrm{Var}(S) + \mathrm{Var}(aD) + 2\mathrm{Cov}(S, aD)$ bepalen we vervolgens $\mathrm{Cov}(S, aD) = \mathrm{E}[aSD] - \mathrm{E}[S]\,\mathrm{E}[aD] = a\mathrm{E}[SD]$, want $\mathrm{E}[D] = 0$. Door slim in te vullen—of een iets langere rekenpartij—vind je $\mathrm{E}[SD] = -2\mathrm{Var}(S)$, zodat het eindresultaat wordt

$$\mathrm{Var}(S+aD) = 4\sigma^2 + a^2 25\sigma^2 + 2a\cdot(-2)\cdot 4\sigma^2 = (4 - 16a + 25a^2)\sigma^2. \tag{1}$$

Tweede: Schrijf $S_c = S + aD = S + a(5 - 2S - 3T) = 5a + (1-2a)S - 3aT$. Deze lineaire combinatie heeft het voordeel dat nu eenvoudig (wegens onafhankelijkheid!) volgt: $\mathrm{Var}(S_c) = (1-2a)^2\mathrm{Var}(S) + 9a^2\mathrm{Var}(T) = \ldots$. Invullen van $a = 0.2$ leidt tot $\mathrm{Var}(S_c) = \frac{8}{5}\sigma^2$. De gevraagde efficiëntie is dus $4/\frac{8}{5} = 2.5$, een flinke winst.

2e De uitdrukking (1) is een “dal”parabool, waarvan het minimum eenvoudig is te bepalen. $a_{opt} = 8/25$, met een relatieve efficientie van $25/9 \approx 2.78$ (om dit laatste was niet gevraagd). Opmerkelijk is dat velen concluderen dat $a = 0$ optimaal is (omdat ze hierboven de covariantieterm gemist hebben): uit hun formule volgt dat voor alle $a \neq 0$ geldt $\mathrm{Var}(S_c) > \mathrm{Var}(S)$. Dat laatste is bijzonder vreemd omdat het suggereert dat het “corrigeren” van schattingen zodat ze aan de massabalans voldoen de nauwkeurigheid zou doen *afnemen*, iets zeer tegenintuitiefs.

3 Deze vraag was voor een deel ook gesteld in het tentamen van 14 april 2011.

3a Er wordt hier eerst om de definitie van p-waarde gevraagd. De p-waarde voor deze data is de kans dat Q groter is dan de gevonden 1.73 indien Q bepaald wordt voor een steekproef van 5 trekkingen uit een normale verdeling met parameters geschat uit de data.[3] In het tweede deel van de vraag wordt gevraagd hoe je de toets uitvoert. De hypothese is dat er geen uitbijter is. We verwerpen deze hypothese, als de p-waarde te klein is, en verwerpen de hypothese niet[4], als de p-waarde groot is. Zie §25.2 van het boek.

3b Aangezien aangenomen is dat de metingen normaal verdeeld zijn, gaan we een schatting maken van deze verdeling. Gebruik hiervoor het gemiddelde en de standaardafwijking van de 5 metingen voor μ en σ van deze normale verdeling. Dit zijn de parameters voor de parametrische bootstrap. Genereer een (bootstrap)dataset van 5 onafhankelijke trekkingen uit deze verdeling en evalueer hiervoor de toetsingsgrootheid Q. 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 1.73.[5]

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

[4] We zeggen dus niet dat het bewezen is dat het punt geen uitbijter is. Het is niet bewezen, dat het punt wel een uitbijter is. Een subtiel verschil.

[5] Wij vonden 0.0366. De bootstrap methodiek helpt dus om deze p-waarde te vinden.

3c De reden om mediaan en MAD te gebruiken is dat zij robuuste schatters zijn, die minder gevoelig zijn voor uitbijters[6]. In de vraag is geopperd dat er uitbijters konden zijn. De schatting voor de toetsingsgrootheid Q is daarom ook een robuuste grootheid.
Welke toets te gebruiken is dan nog een vraag. Argumenten kunnen zijn:

1. Bovengenoemde argument van robuustheid is een valide argument.

2. Met de gewone methode en de robuuste methode worden resp p=0.036 en 0.058 gevonden. De verleiding bestaat om op grond van deze getallen een keuze te maken, hetgeen zeer discutabel is. Het is mogelijk dat er een verschil is in onderscheidend vermogen, hier te interpreteren als hoe groot het verschil moet zijn ten opzichte van de nulhypothese om tot verwerpen te besluiten. Mogelijk dat de ene toets hierbij gevoeliger is dan de andere. Dit valt echter niet te concluderen zonder nader onderzoek.

[6] Sommige kandidaten gebruikten de term 'nauwkeuriger'. Dit refereert naar de zuiverheid van de grootheid, niet de robuustheid.