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

Studienummers: Vult u alstublieft op het meerkeuzevragenformulier 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. In het gezin van Fiona wordt er aan het eind van de maaltijd altijd geloot wie van de vijf zonen moet afwassen en wie er moet afdrogen. De procedure is als volgt: Fiona doet 5 letters van het scrabblespel in een opgevouwen theedoek, drie O's, een D en een W. De theedoek gaat de tafel rond en wie de D pakt moet drogen, wie de W pakt moet afwassen. Bereken de kans dat de derde persoon een van de twee taken krijgt toebedeeld.

a. $\frac{1}{20}$ **b.** $\frac{2}{5}\cdot\frac{1}{4}$ **c.** $1-(3/5)^2$ **d.** $\frac{2}{5}$ **e.** $(2/5)^2$ **f.** $1-\frac{3}{5}\cdot\frac{2}{4}$

2. Op een avond klagen Albert en Bernard dat zij—doordat zij altijd als eerste en tweede moeten trekken (dat komt door de vaste tafelindeling, die alfabetisch is)—dat zij vaker de klos zijn dan Douwe en Elbert, die als vierde en vijfde trekken. Laat AB de gebeurtenis zijn: A en B moeten beiden iets doen.

Er geldt

a. $\mathrm{P}(AB)=\frac{1}{5}$ en $\mathrm{P}(AB)=\mathrm{P}(DE)$ **b.** $\mathrm{P}(AB)=\frac{1}{5}$ en $\mathrm{P}(AB)>\mathrm{P}(DE)$
c. $\mathrm{P}(AB)=\frac{1}{10}$ en $\mathrm{P}(AB)=\mathrm{P}(DE)$ **d.** $\mathrm{P}(AB)=\frac{1}{10}$ en $\mathrm{P}(AB)>\mathrm{P}(DE)$
e. $\mathrm{P}(AB)=\frac{2}{25}$ en $\mathrm{P}(AB)=\mathrm{P}(DE)$ **f.** $\mathrm{P}(AB)=\frac{2}{25}$ en $\mathrm{P}(AB)>\mathrm{P}(DE)$

3. Stel X heeft de dichtheid

$$f(x)=\begin{cases}0 & \text{als } x<-1 \text{ of } x>1;\\ \frac{1}{2}x+\frac{1}{2} & \text{als } -1\le x\le 1\end{cases}$$

De verdelingsfunctie van X is voor $-1\le x\le 1$ gelijk aan

a. 1 **b.** $\frac{1}{2}$ **c.** $\frac{1}{2}x+\frac{1}{2}$
d. $\frac{1}{2}x+\frac{1}{4}x^2$ **e.** $\frac{1}{2}x+\frac{1}{4}x^2+\frac{1}{2}$ **f.** $\frac{1}{2}x+\frac{1}{4}x^2+\frac{1}{4}$

4. Voor het uitvoeren van een simulatiestudie is het nodig te simuleren aan de hand van de volgende verdelingsfunctie: $F(x)=1-\mathrm{e}^{-20\sqrt{x}}$ 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. $X=-20\ln U$ **b.** $X=-0.1\ln U$ **c.** $X=[-\ln(0.05U)]^2$
d. $X=[\ln U]^2/400$ **e.** $X=1-\mathrm{e}^{-20\sqrt{U}}$ **f.** $X=\mathrm{e}^{-20\sqrt{U}}$

5. Stel T_1 en T_2 zijn onafhankelijke zuivere schatters voor een parameter θ met varianties σ^2 resp. $2\sigma^2$. Bekijk nu de schatters $S_1 = \frac{1}{2}(T_1 + T_2)$ en $S_2 = \frac{2}{3}T_1 + \frac{1}{3}T_2$. Dan geldt:

a. beide schatters zijn zuiver en hebben dezelfde variantie
b. beide schatters zijn zuiver en $\text{var}(S_1) < \text{var}(S_2)$
c. beide schatters zijn zuiver en $\text{var}(S_2) < \text{var}(S_1)$
d. $\text{var}(S_1) < \text{var}(S_2)$, maar over de zuiverheid is niets te zeggen
e. $\text{var}(S_1) = \text{var}(S_2)$, maar over de zuiverheid is niets te zeggen
f. $\text{var}(S_2) < \text{var}(S_1)$, maar over de zuiverheid is niets te zeggen

6. Gegeven de stochasten X en Y met correlatie $\rho \neq 0$. Beschouw de volgende twee beweringen

A $\quad \text{Cov}(X, X+Y) < \text{Cov}(X, Y)$
B $\quad \text{Var}(X+Y) < \text{Var}(X) + \text{Var}(Y)$

a. A en B zijn allebei onwaar
b. A is waar en B is onwaar
c. A is waar als $\rho < 0$ en B is onwaar
d. A is onwaar en B is waar
e. B is waar als $\rho < 0$ en A is onwaar
f. als $\rho < 0$ zijn A en B beide waar

7. Omwonenden van de nieuwe spoorlijn van Bronsvoort naar Gouddrecht klagen over geluidsoverlast in de nacht. De Zilverlandse autoriteiten beweren dat goederentreinen niet meer dan 90 decibel produceren. Een onderzoeksbureau heeft gedurende zeven weken het geluidsniveau van de eerste vier treinen na middernacht gemeten. De 196 metingen gaven een gemiddelde van 93 met een standaardafwijking van 15.
Bereken de p-waarde van deze uitkomst als we $H_0 : \mu = 90$ toetsen tegen $H_0 : \mu > 90$.

a. 0.0808 **b.** 0.0668 **c.** 0.0228 **d.** 0.0139 **e.** 0.0062 **f.** 0.0026

8. [zelfde contekst als vorige opgave] Voor de omwonenden is natuurlijk meer van belang hoe vaak er een 'grove overschrijding' is. We noemen een geluidsniveau boven de 105 decibel onacceptabel. Geef, onder de aanname dat het geluidsniveau van goederentreinen normaal verdeeld is met de uit de data geschatte waarden als parameters, een schatting van het percentage treinen dat een onacceptabele hoeveelheid geluid voortbrengt.

a. 38 % **b.** 31 % **c.** 21 % **d.** 16 % **e.** 8 % **f.** 5 %

9. Over een histogram is het volgende gegeven:

cel	hoogte
$[0, 2]$	0.05
$(2, 6]$	0.10
$(6, 9]$	0.10
$(9, 13]$	0.05

De waarden van de empirische verdelingsfunctie in de punten 2 en 4 zijn dan respectievelijk

a. 0.05 en 0.10
b. 0.10 en 0.20
c. 0.10 en 0.30
d. 0.05 en niet te bepalen
e. 0.10 en niet te bepalen
f. 1/6 en 1/3

10. De inspectie voor de volksgezondheid doet 25 metingen aan de concentratie van een giftige stof in grondwater. De metingen leiden tot een gemiddelde concentratie van 2.25 ppm en een steekproefvariantie van 0.25 ppm^2. Men berekent een 95% betrouwbaarheidsinterval voor de verwachte concentratie μ in het grondwater onder de aanname dat de 25 metingen een realisatie vormen van een steekproef uit een normale verdeling. Het interval is

a. $(2.21, 2.29)$
b. $(2.15, 2.35)$
c. $(2.08, 2.42)$
d. $(2.04, 2.46)$
e. $(1.39, 3.11)$
f. $(1.22, 3.28)$

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. Ajdacic-Gross et al (2012) rapporteerde dat de kans op overlijden op een verjaardag 14% hoger is dan op enig andere dag van het jaar. Dit was gebaseerd op de analyse van 2.5 miljoen Zwitsers die overleden in de periode 1969-2008.

a. Formuleer een toepasselijke toets en bereken de p-waarde behorende bij deze data. Beschrijf en motiveer duidelijk de stappen en aannamen, die genomen zijn om het antwoord te bereiken.

b. Er worden drie verklaringen geopperd: mensen wachten onbewust op de komende verjaardag (de "hang-on" hypothese); nemen meer risico op hun verjaardag (het "jumping the gun effect"); op grote schaal is door administratieve fouten de verjaardag ingevoerd als sterfdag.

Geef voor elk van deze mogelijke verklaringen (afzonderlijk) zo precies mogelijk aan hoe deze statistisch getoetst of anderszins onderzocht kan worden.

2. Een stochastische variabele X heeft kansdichtheidsfunctie

$$f(x) = \frac{6}{a^3}(c^2 - x^2) \text{ met } a, c > 0 \text{ voor } |x| \leq c \text{ en } 0 \text{ elders.}$$

a. Laat zien dat moet gelden: $a = 2c$.

b. Beschrijf zo duidelijk mogelijk het *maximum likelihood principle*, dat de basis is voor maximum likelihood methode (ML). Maximum likelihood schatters hebben enkele gunstige eigenschappen. Beschrijf die zo precies mogelijk.

c. Er zijn nu een aantal observaties $x_1, \ldots, x_n$ gedaan. Laat zien dat de ML schatting voor c voldoet aan

$$\sum_{i=1}^{n} \frac{1}{c^2 - x_i^2} = \frac{3n}{2c^2},$$

waarbij $c \geq \max |x_i|$.

3. In de polymeerkunde[1] is de lengte X van een polymeermolecule, het aantal monomeren in de keten, een stochastische grootheid met de volgende verdeling:

$$\mathrm{P}(X = i) = Kp^i \text{ met } i = 1, 2, \ldots,$$

waarbij de keten dus (theoretisch) willekeurig lang kan zijn. Het getal p is te interpreteren als de kans dat een "volgend" monomeer aansluit bij de keten.

a. Bepaal de constante K in de formule.

b. Bepaal de *weight average degree of polymerization*, gegeven door $\frac{\mathrm{E}[X^2]}{\mathrm{E}[X]^2}$ en laat zien dat deze een uitkomst heeft op het interval $(1, 2)$.

4. 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

[1] Young, R.J., *Introduction to polymers*, Chapman and Hall, 1983

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

a. De fabrikant wil niet uitgaan van normaliteit en besluit om een bootstrapbetrouwbaarheidsinterval voor μ te construeren. Beschrijf nauwkeurig het bijbehorende bootstrapexperiment; geef hierbij duidelijk aan hoe een bootstrapsteekproef getrokken wordt en wat er per steekproef wordt berekend.

b. 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

Antwoorden multiple choice:

1 d. $\frac{2}{5}$

2 c.

3 f.

4 d. Los op naar x: $F(x) = u$ voor $0 \leq u \leq 1$, dan vind je $x = [-0.05 \ln(1-u)]^2$. Zie verder paragraaf 6.2.

5 c. Uit de lineariteit van de verwachting volgt $\mathrm{E}[S_1] = \mathrm{E}[S_2] = \theta$. Voor de varianties vinden we: $\mathrm{E}[S_1] = \frac{3}{4}\sigma^2$ en $\mathrm{E}[S_2] = \frac{2}{3}\sigma^2$. De laatste is de kleinste.

6 e. Met de rekenregel voor de covariantie vinden we $\mathrm{Cov}(X, X+Y) = \mathrm{Var}(X) + \mathrm{Cov}(X,Y)$ hetgeen minstens $\mathrm{Cov}(X,Y)$ bedraagt; A is dus onwaar. Het linker- en rechterlid van B verschillen precies twee maal $\mathrm{Cov}(X,Y)$, dus als dit negatief is, is B waar.

7 f.

8 c.

9 e. De waarde van de empirische verdelingsfunctie kan alleen op de celgrenzen bepaald worden. De eerste cel heeft oppervlak 0.05, dus $0.1 = F_n(2)$. Verder kunnen we alleen concluderen $0.1 \leq F_n(4) \leq 0.5$, maar de precieze waarde is niet te bepalen.

10 d. We weten dat voor de echte onbekende verwachting μ geldt:

$$\frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \text{ heeft een } t(n-1) \text{ verdeling.}$$

Hieruit volgt de formule voor het betrouwbaarheidsinterval vor μ:

$$\left(\bar{x}_n - t_{0.025,24} s_n/\sqrt{n}, \bar{x}_n + t_{0.025,24} s_n/\sqrt{n}\right)$$

Dit leidt tot $(2.25 - 2.064 \cdot 0.5/5, 2.25 + 2.064 \cdot 0.5/5) = (2.04, 2.46)$.

Antwoorden open vragen:

1 Voor het volledige artikel zie: http://dx.doi.org/10.1016/j.annepidem.2012.04.016.

1a H_0: overlijdenskans op verjaardag = 1/365. H_1: overlijdenskans op verjaardag ¿ 1/365. Toetsingsgrootheid T is het aantal op zijn/haar verjaardag overledenen. Onder H_0 heeft T een $Bin(2.5 \cdot 10^6, 1/365)$ verdeling, het verwachte aantal is dan $2.5 \cdot 10^6/365 = 6849$. Het werkelijke aantal is 1.14 maal zo groot, dus (ongeveer) 7808. De p-waarde is derhalve $\mathrm{P}(T \geq 7808)$. We gebruiken de centrale limietstelling om de binomiale verdeling van T te benaderen met een normale, hetgeen geoorloofd is omdat n heel erg groot is (ook al is p klein). Bij benadering is T derhalve $N(6849, 6831)$ verdeeld. Zo vinden we: $\mathrm{P}(T \geq 7808) \approx \mathrm{P}(Z \geq 11.6)$, minuscuul en niet in de tabel $(2.4 \cdot 10^{-31})$.

1b Toets op een overlijdensdip vòòr de verjaardag; concreet: vergelijk het aantal overledenen in de periode (zeg) 10 dagen voor de verjaardag; corresponderende kans zou 10/365 moeten zijn; het toetsen gaat als bij **a.**.
Toets op een dip op de dagen ná de verjaardag; gaat analoog.
Doe een steekproef uit de als op-verjaardag-overleden geregistreerde personen en controleer de overlijdensdatum; als administratieve fouten de verklaring zijn, dan zou ongeveer 14% fout moeten zijn; ook dit zou je kunnen toetsen. We hebben hier een dataset van (circa) 7808 mensen.

2a Uit de eigenschap dat de totale kansmassa per definitie gelijk is 1:

$$\int_{\infty}^{\infty} f(x)dx = \int_{-c}^{c} \frac{6}{a^3}(c^2 - x^2)dx = 8\frac{c^3}{a^3} \equiv 1$$

volgt dat $a = 2c$. Formeel zijn er uiteraard nog 2 mogelijkheden, maar die voldoen niet aan de voorwaarden voor a en c.

2b Zie §21.1 voor het ML-principe, §21.4 voor de eigenschappen: invariantieprincipe; asymptotisch zonder bias; asymptotisch minimale variantie.

2c De likelihoodfunctie

$$L(c) = \prod_{i=1}^{n} f(x_i)$$

is gelijk aan nul voor $c \leq \max_{i=1,\ldots,n} |x_i|$, omdat minstens één term in het product nul is. Ergo, c is groter dan het maximum van de absolute waarden. Elders is

$$L(c) = \prod_{i=1}^{n} \frac{6}{(2c)^3}(c^2 - x_i^2),$$

dus de log-likelihood is

$$l(c) = n \ln \frac{6}{(2c)^3} + \sum_{i=1}^{n} \ln(c^2 - x_i^2).$$

Differentiëren naar c en 0 stellen levert de gevraagde uitdrukking. Een tekenoverzicht van l' laat zien dat het stationaire punt inderdaad een maximum is.
Dit is overigens een mooi voorbeeld van een situatie waarbij de ML-schatter niet gegeven is door een eenvoudige gesloten uitdrukking.

3a We herkennen de geometrische verdeling, alleen wordt bij de $Geo(p)$ verdeling de kans $\mathrm{P}(Y = i) = p(1-p)^{i-1}$ telkens een factor $1 - p$ kleiner, waar dat in de gegeven formule een factor p is. Blijkbaar heeft de ketenlengte een $Geo(1-p)$ verdeling, geometrisch met parameter $1 - p$. Dus $K = (1 - p)/p$.

3b Van het formuleblad voor $X \sim Geo(1-p)$: $\mathrm{E}[X] = 1/(1-p)$ (in de polymeerkunde heet dit *Carothers equation*) en $\mathrm{Var}(X) = p/(1-p)^2$. Verder geldt $\mathrm{E}\left[X^2\right] = \mathrm{Var}(X) + (\mathrm{E}[X])^2$, zodat

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

Verder zien we uit de formules boven dat $\mathrm{Var}(X)/\mathrm{E}[X]^2 = p$, dus de gevraagde verhouding is $1 + p$. Omdat p een kans is, ergo een getal op het interval $(0, 1)$, ligt de uitkomst op het interval $(1, 2)$.

4 Deze vraag was gelijk aan vraag **2bc** van het tentamen van 20 april 2012!

4a Zie dictaat, § 23.3.

4b We gebruiken de formule uit §23.3. De bootstrap kritieke waarden zijn $c_\ell = t^*_{(25)} = -2.101$ en $c_r = t^*_{(976)} = 2.088$, Een betrouwbaarheidsinterval voor μ wordt nu $(4.997 - 0.0264 \cdot 2.088, 4.997 + 0.0264 \cdot 2.101) = (4.9419,\ 5.0524)$.