Technische Universiteit Delft
Fac. der Technische Wiskunde en Informatica

Mekelweg 4
2628 CD Delft

Tentamen wi1604 Kansrekening en Statistiek I
21 juni 2004, 14.00–17.00 uur (deeltentamen tot 16.00 uur)

Toelichting. Er wordt verwacht dat je bij elke vraag een uitwerking geeft voorzien van toelichting en motivering. Alleen het antwoord levert niets op. Alle onderdelen hebben alle hetzelfde gewicht. Bij dit examen is het gebruik van een rekenmachine en het standaard Kanstat formuleblad toegestaan.

Deeltentamen (2 uur): opgave 1–3, 5.

1. Gegeven twee onafhankelijke Poisson verdeelde stochastische variabelen X en Y, met parameter $\lambda > 0$, respectievelijk $\mu > 0$.

   a. Laat zien dat de kansgenererende functie van X gelijk is aan $G(s) = \mathrm{e}^{\lambda(s-1)}$.

   b. Toon aan (met behulp van kansgenererende functies) dat $X+Y$ een $Pois(\lambda+\mu)$ verdeling heeft.

   c. Bepaal (voor $n \in \{1, 2, \ldots\}$) de voorwaardelijke kansen $\mathrm{P}(X = k \,|\, X + Y = n)$, $k = 0, 1, \ldots, n$.

2. a. In 1929 onderzocht de astronoom Hubble de relatie tussen de afstand x tussen sterrenclusters (in eenheden van 1 miljoen lichtjaren), en de snelheid y waarmee de clusters zich van elkaar verwijderen (in eenheden van 1000 mijlen per seconde). Men fit het model $Y = \beta_0 + \beta_1 x + \epsilon$. Gegeven is

$\sum x_i$	$\sum y_i$	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$	n
4185	148.65	95161.2	2685141	3376.313	10

   Bovendien is de residuele kwadraatsom van deze analyse gelijk aan 3.782 en de kleinste kwadratenschatter voor het intercept $\hat{\beta}_0 = 0.096$. Toets de nulhypothese $H_0 : \beta_0 = 0$ bij $\alpha = 0.05$.

   b. Wat is heteroscedasticiteit en hoe kun je aan de residuen van een regressie zien of daar sprake van is? Hoe verschaffen de residuen informatie over de vraag of het terecht is een lineair model aan de data te fitten?

3. Hieronder is een dataset met levensduren gegeven. Een misschien toepasselijke verdeling is de Weibull-verdeling. We willen een betrouwbaarheidsinterval voor de verwachte levensduur 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 en bepaal het 90% betrouwbaarheidsinterval.

c. De gestudentiseerde bootstrap met 1000 herhalingen is uitgevoerd; hieronder staat een deel van de geordende bootstrap-uitkomsten. Gebruik ze om een 90% bootstrap betrouwbaarheidsinterval te maken voor de verwachting.

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

d. Leg uit welke van de twee methoden hier de betere is. Geef hierbij zo goed mogelijk aan welke factoren je beschouwt bij je keuze.

4. Een manuscript bevat een aantal drukfouten, zeg: n. Er zijn twee correctoren die onafhankelijk van elkaar een kopie van het manuscript corrigeren. De eerste corrector vindt in totaal 81 drukfouten; de tweede 56. Als model nemen we dat het aantal door corrector i gevonden fouten gezien kan worden als een binomiaal experiment met parameters n en p_i, $i = 1, 2$.

   a. Leid de formules af voor de maximum-likelihood schatters voor p_1 en p_2, en geef de schattingen die uit de data volgen.

   b. Het aantal drukfouten n is natuurlijk niet bekend en zou geschat moeten worden. Er werden in totaal 37 drukfouten gevonden door beide correctoren. Geef een onderbouwde schatting voor n. *Hint*: Bereken eerst de kans dat een gegeven drukfout door beide correctoren wordt gevonden.

   c. Uit de gegevens blijkt dat de n fouten in vier categorieën vallen (gevonden door beide correctoren, alleen door de eerste, alleen door de tweede, door geen van beide). Druk het aantal manieren waarop dit kan uit in n.

   d. In feite hebben we hier te maken met een probleem met drie onbekende parameters: n, p_1 en p_2. Wat is de likelihoodfunctie voor deze parameters?

5. Laat X een stochastische variabele zijn met dichtheid $f(x) = 0$ voor $x \leq 0$ en

$$f(x) = c\left(\mathrm{e}^{-2x} + \mathrm{e}^{-3x}\right) \qquad \text{voor } x > 0.$$

   a. Bepaal de constante c.

   b. Bepaal de verdelingsfunctie $F(x)$ van X.

   c. Bepaal de kansdichtheid van $Y = \mathrm{e}^{-X}$.

   d. Bereken de covariantie tussen Y en $Z = \mathrm{e}^{X}$.

Beknopte uitwerkingen

1a $\mathrm{E}\big[s^X\big] = \sum_{k=0}^{\infty} s^k \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda s)^k}{k!} = \mathrm{e}^{-\lambda}\cdot\mathrm{e}^{\lambda s} = \mathrm{e}^{\lambda(s-1)}$.

1b Uit Stelling 11.2 (aanvulling 11B) volgt wegens onafhankelijkheid van X en Y:

$$G_{X+Y}(s) = G_X(s)\cdot G_Y(s) = \mathrm{e}^{\lambda(s-1)}\mathrm{e}^{\mu(s-1)} = \mathrm{e}^{(\lambda+\mu)(s-1)}.$$

De laatste uitdrukking correspondeert met de $Pois(\lambda+\mu)$ verdeling.

1c Gebruikmakend van het feit dat alle verdelingen Poisson zijn, en dat X en Y onafhankelijk zijn, vinden we voor de gevraagde conditionele kans:

$$\begin{aligned}\mathrm{P}(X=k\,|\,X+Y=n) &= \frac{\mathrm{P}(X=k,Y=n-k)}{\mathrm{P}(X+Y=n)} = \frac{\mathrm{P}(X=k)\cdot\mathrm{P}(Y=n-k)}{\mathrm{P}(X+Y=n)}\\ &= \frac{\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}\cdot\frac{\mu^{n-k}}{(n-k)!}\mathrm{e}^{-\mu}}{\frac{(\lambda+\mu)^n}{n!}\mathrm{e}^{-(\lambda+\mu)}} = \frac{n!}{k!(n-k)!}\cdot\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{n-k}.\end{aligned}$$

We zien dat X, gegeven $X+Y=n$, een $Bin(n,\frac{\lambda}{\lambda+\mu})$ verdeling heeft.

2a Een noodzakelijke, maar niet genoemde, aanname is dat de meetfouten iid zijn en een $N(0,\sigma^2)$ verdeling hebben. Het aantal vrijheidsgraden is $n-2=10-2=8$, de schatting voor de foutvariantie is dus $\hat\sigma^2 = 3.782/8 = 0.4728$. De formule voor de variantie van het intercept (zie het formuleblad) geeft vervolgens:

$$S_a^2 = \frac{\sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2}\hat\sigma^2 = \frac{2685141}{9337185}\cdot 0.4728 = 0.2876\cdot 0.4728 = 0.1360.$$

De waarde van de toetsingsgrootheid wordt dus: $t = (0.096-0)/\sqrt{0.1360} = 0.2603$. De kritieke waarde bij $\alpha=0.05$ en 8 vrijheidsgraden is $t_{8,0.05} = 2.306$, dus we verwerpen de nulhypothese niet. De P-waarde is ongeveer 0.8.

2b Zie Section 22.2: **Residuals**.

3a 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 (want $27.77 - 1.5\cdot\mathrm{IQR}$ is kleiner). 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.

3b 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 $t_{42,0.05} \approx 1.682$ (door interpolatie in de tabel). We vinden dus $40.79 \pm 1.682\cdot 22.09/\sqrt{43} = 40.79 \pm 1.682\cdot 3.369 = 40.79 \pm 5.67 = (35.12, 46.46)$.

3c Uit de bootstrapresultaten vinden we schattingen voor de kritieke waarden: de 900 'middelste' uitkomsten liggen tussen -1.903 en 1.564. Bij benadering geldt dus $\mathrm{P}(-1.903 < T^* < 1.564) = 0.9$, en volgens de bootstrapbenadering geldt hetzelfde (bij benadering) voor T, zodat het betrouwbaarheidsinterval wordt: $(40.79 - 1.564\cdot 22.09/\sqrt{43}, 40.79 - (-1.903)\cdot 22.09/\sqrt{43}) = (40.79 - 5.27, 40.79 + 6.41) = (35.52, 47.20)$.

3d Er zijn twee dingen om te overwegen. Namelijk, in hoeverre het waargenomen fenomeen een normale verdeling volgt, en hoe groot de dataset is. Aan de boxplot is te zien dat de dataset enigzins scheef naar rechts is. (Dit komt bij levensduren vaker voor, en de Weibull-verdeling is over het algemeen scheef—afhankelijk van de parameter waarden zelfs zeer scheef). De omvang van de dataset is met 43 waarden 'middelmatig'. Alles bijeen is er geen goede reden de normale benadering te gebruiken. (We zien dat de bootstrap het betrouwbaarheidsinterval 'corrigeert': het is wat ruimer en 'scheef naar rechts'.)

4a Bij een binomiaal experiment wordt de loglikelihood, met k het aantal gevonden drukfouten:

$$l(p) = \ln L(p) = \ln \binom{n}{k} + k \ln p + (n-k) \ln(1-p).$$

Differentieren naar p en nulstellen geeft $0 = \frac{k}{p} - \frac{n-k}{1-p}$, zodat $\hat{p} = \frac{k}{n}$. Via de tweede afgeleide van $l(p) = -k/p^2 - (n-k)/(1-p)^2$, die negatief is, zien we dat dit punt inderdaad een maximum oplevert. De uitdrukkingen zijn dus:

$$\hat{p}_1 = \frac{81}{n} \qquad \text{en} \qquad \hat{p}_2 = \frac{56}{n}.$$

4b De kans dat een gegeven drukfout door beide wordt gezien is vanwege onafhankelijkheid gelijk aan $p_1 p_2$ en kan dus geschat worden door

$$\frac{81 \cdot 56}{n^2}.$$

Anderzijds is ook $\frac{37}{n}$ een schatting voor $p_1 p_2$. De wet van de grote aantallen suggereert dat beide schattingen hetzelfde zouden moeten opleveren:

$$\frac{81 \cdot 56}{n^2} = \frac{37}{n} \quad \text{waaruit volgt:} \quad n \approx 123.$$

Een schatting voor het aantal resterende drukfouten is dus $123 - 81 - 56 + 27 = 23$.

4c De aantallen voor de vier categorieën zijn: 37, 44, 19, $n - 100$. Wanneer we ons voorstellen dat de fouten genummerd zijn, en we ze verdelen over posities 1 tot en met n, en 1–37 correspondeert met categorie 1, 38–81 met 2, enzovoorts, dan is het duidelijk dat het aantal manieren is:

$$\frac{n!}{37!44!19!(n-100)!} = \binom{n}{100} \frac{100!}{37!44!19!}.$$

Een andere manier is als volgt. De fouten voor de eerste categorie kunnen we op $\binom{n}{37}$ manieren kiezen. Ongeacht *hoe* dat gebeurt, voor de tweede categorie zijn er $\binom{n-37}{44}$ manieren mogelijk, en $\binom{n-81}{19}$ voor de derde, en de vierde categorie ligt dan vast. Het produkt van de binomiaalcoëfficienten vereenvoudigt tot de uitdrukking boven.

4d De uitdrukking voor de kans om de n fouten als gegeven te verdelen over de vier categorieën (met kans respectievelijk $p_1 p_2$, $p_1(1-p_2)$, $(1-p_1)p_2$, $(1-p_1)(1-p_2)$) vereenvoudigt tot:

$$L(n, p_1, p_2) = \frac{100!}{37!44!19!} \binom{n}{100} p_1^{81} (1-p_1)^{n-81} p_2^{56} (1-p_2)^{n-56}.$$

In het maximum geldt: $p_1 = \frac{81}{n}$, $p_2 = \frac{56}{n}$. Door dit te substitueren houden we een functie van n over. Met wat moeite en een luie maximalisatie met Splus volgt: $\hat{n} = 121$.

5a Wegens $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = c(\frac{1}{2}+\frac{1}{3}) = 1$ volgt $c = 6/5$.

5b Voor $t < 0$ geldt natuurlijk $F(t) = 0$, en voor $t \geq 0$:

$$F(t) = \frac{6}{5}\int_0^t \left(\mathrm{e}^{-2x} + \mathrm{e}^{-3x}\right)\mathrm{d}t = 1 - \frac{3}{5}\mathrm{e}^{-2t} - \frac{2}{5}\mathrm{e}^{-3t}.$$

5c Uit de definitie van Y en het bereik van X zien we dat Y een stochast in het interval $(0,1)$ is. Wegens $\mathrm{P}(Y \leq t) = \mathrm{P}\left(\mathrm{e}^{-X} \leq t\right) = \mathrm{P}(X \geq -\ln t)$ vinden we voor $0 \leq t \leq 1$:

$$F_Y(t) = 1 - F_X(-\ln t) = \frac{3}{5}t^2 + \frac{2}{5}t^3 \quad \text{en dus} \quad f_Y(t) = \frac{6}{5}t(1+t).$$

Dit is ook af te leiden met behulp van de transformatie-formule uit Stelling 8.1 uit de aanvulling op Chapter 8.

5d We dienen te bepalen: $\mathrm{E}[YZ]$, $\mathrm{E}[Y]$ en $\mathrm{E}[Z]$. De eerste is makkelijk: $YZ \equiv 1$, dus $\mathrm{E}[YZ] = 1$. Voor de andere twee moeten we rekenen. Van Y kennen we de kansdichtheid, dus we doen het als volgt:

$$\mathrm{E}[Y] = \int_{-\infty}^{\infty} t f_Y(t)\mathrm{d}t = \int_0^1 \frac{6}{5}\left(t^2 + t^3\right)\mathrm{d}t = \frac{7}{10}.$$

De stochast Z kennen we alleen als functie van X dus gebruiken we de "change of variable formula:"

$$\mathrm{E}[Z] = \mathrm{E}\left[\mathrm{e}^X\right] = \int_{-\infty}^{\infty} \mathrm{e}^x f(x)\mathrm{d}x = \frac{6}{5}\int_0^{\infty} \mathrm{e}^x\left(\mathrm{e}^{-2x} + \mathrm{e}^{-3x}\right)\mathrm{d}x = \frac{9}{5}.$$

Samenvattend: $\mathrm{Cov}(Y,Z) = \mathrm{E}[YZ] - \mathrm{E}[Y]\,\mathrm{E}[Z] = 1 - \frac{7}{10}\cdot\frac{9}{5} = -\frac{13}{50}$.