Tentamen Kanstat
WI1275TA
27 maart 2009, 14 – 17 uur

Bij dit examen is het gebruik van een (evt. grafische) rekenmachine toegestaan.
Een formuleblad wordt bijgeleverd.
Normering: **1.** 9 pt, **2.** 7 pt, **3.** 8 pt, **4.** 6 pt, **5.** 12 pt, **6.** 12 pt.
Een antwoord moet voorzien zijn van een berekening, toelichting en/of motivatie.
Dit alles goed leesbaar en in goed Nederlands (of Engels).

---

1. Een zak bevat drie munten: twee zuivere munten en een munt met aan beide zijden kop.

a. Als je een willekeurige munt uit de zak neemt en die opgooit, wat is dan de kans op kop?

b. Als je een willekeurige munt uit de zak neemt, die opgooit, kop ziet, wat is dan de (voorwaardelijke) kans dat je de valse munt hebt getrokken?

c. Stel je hebt een munt getrokken en kop gegooid. Als je met dezelfde munt nog een keer gooit, wat is dan de (voorwaardelijke) kans op kop?

2. Een stochast X heeft verdelingsfunctie $F(x) = \begin{cases} 0\ , & \text{als } x \leq 0 \\ 1\ , & \text{als } x \geq 4 \\ \sqrt{x}/2, & \text{als } 0 \leq x \leq 4 \end{cases}$
Bereken achtereenvolgens

a. $\mathrm{P}\,(1 \leq X \leq 4)$ en $\mathrm{P}\,(3 \leq X \leq 9)$.

b. De dichtheid van X en de verwachting van X.

3. X en Y hebben een gezamenlijke verdeling volgens de volgende tabel, waarin $p(a,b) = \mathrm{P}(X = a, Y = b)$:

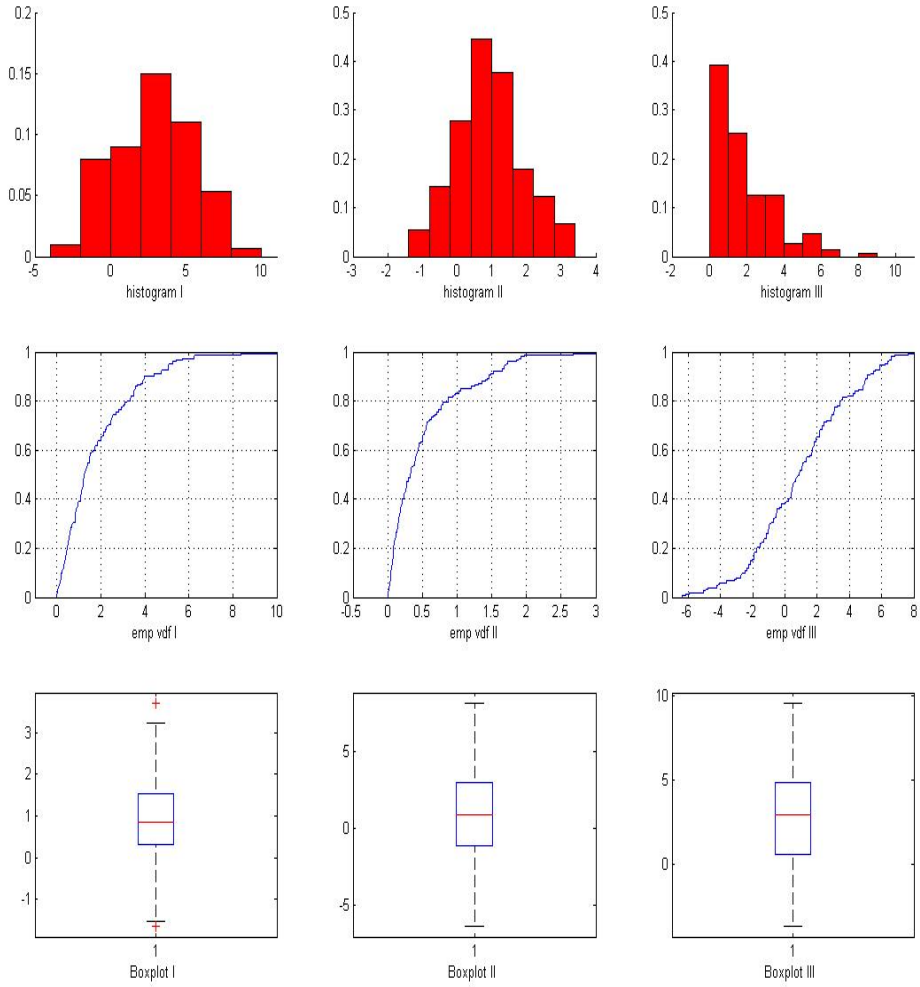
$p(a,b)$		a			
		0	1	2	3
b	-1	0.10	0.10	0.05	0.05
	0	0.10	0.05	0.05	0.05
	1	0.05	0.10	0.15	0.15
$p_X(a)$		0.25	0.25	0.25	0.25

a. Bereken $\mathrm{P}\,(Y = 0)$ en $\mathrm{P}(X = 1 \,|\, Y = 0)$.

b. Ga na of X en Y afhankelijk of onafhankelijk zijn. (Argument!)

c. Bereken $\mathrm{Cov}\,(X, Y)$.

4. Hieronder tref je negen plaatjes aan van trekkingen van 150 elementen uit de volgende vijf verdelingen:
$\mathcal{N}(1, 3^2)$, $\mathcal{N}(3, 3^2)$, $\mathcal{N}(1, 1)$, Exp(2) en Exp(1/2).
Geef voor elk van de negen aan welke verdeling het meest waarschijnlijk is **en waarom**.



5. Gegevens uit het zwangerschapsonderzoek van Weinberg en Gladen.
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

In totaal deden er 474 vrouwen mee, en is er gegeven dat

$$f_1 + 2f_2 + 3f_3 + \ldots + 12f_{12} = 1285.$$

Aanname: Elke vrouw heeft elke maand opnieuw dezelfde kans om zwanger te geraken. Dan: Y, het aantal perioden dat een vrouw nodig heeft om zwanger te geraken, heeft een geometrische verdeling met parameter p.

a. Druk de kans $\mathrm{P}(Y = 1)$ uit in p.
Schat p op grond van de data; noem de bijbehorende schatter T_1.
Welke functie van $X_1, X_2, \ldots, X_{474}$ geeft T_1?

b. Wat zou op grond van de door jou gekozen waarde van p de kans $\mathrm{P}(Y > 2)$ zijn? Noem de bijbehorende schatter T_2.
Geef de functie g zodat $T_2 = g(T_1)$
Wat is de 'empirische' kans $\mathrm{P}(Y > 2)$?
Noem de hier relevante schatter S_2.

c. Ga van beide schatters uit onderdeel **b.** na of ze zuiver zijn.

d. Geef de maximum-likelihoodfunctie $L(p)$ bij de gegeven data.

6. Een oliemaatschappij gaat een nieuwe boormethode testen. De ontwikkelaars van de methode claimen dat deze methode in tachtig procent van de gevallen beter is dan de 'conventionele' methode. Deze claim wordt als nulhypothese genomen in een toetsingsprobleem. Hiervoor worden 30 boringen gedaan met de nieuwe en de oude methode. Het aantal boringen waar de nieuwe methode het beter doet noemen we X. De werkelijke kans op een beter resultaat met de nieuwe boormethdoe noemen we p_B.

a. Wat is, onder de nulhypothese, de verdeling van X?

H_0 wordt getoetst tegen de alternatieve hypothese $H_1 : \ p_B < 0.8$.

b. Stel dat de nieuwe methode in 20 van de 30 gevallen beter scoort. Bereken de bijbehorende p-waarde. Bij dit onderdeel moet je daarvoor de normale benadering met continuïteitscorrectie gebruiken.

c. Geef het kritieke gebied bij het significantieniveau $\alpha = 0.10$.

Een andere aanpak van het toetsingsprobleem: neem als alternatieve hypothese: de methoden zijn even goed, oftewel $p_B = 0.5$. De nulhypothese wordt verworpen bij uitkomsten k die onder H_1 een grotere kans hebben dan onder H_0.

d. Wordt de nulhypothese nu verworpen bij de uitkomst $k = 20$?

e. Wat wordt het kritieke gebied bij de tweede methode?

Antwoorden open vragen

1a Snelste aanpak: elk van de zes zijden heeft kans 1/6 om boven te komen; vier van de zes zijden zijn kop, dus de kans op kop is $4/6 = 2/3$.
Met voorwaardelijke kansen: voer in: Z: getrokken munt is zuiver, en K_1: eerste worp levert kop. Dan

$$\mathrm{P}(K_1) = \mathrm{P}(Z)\,\mathrm{P}(K_1 \mid Z) + \mathrm{P}(Z)^c\,\mathrm{P}(K_1 \mid Z^c) = \frac{2}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot 1 = \frac{2}{3}$$

1b

$$\mathrm{P}(Z^c \mid K_1) = \frac{\mathrm{P}(K_1 \cap Z^c)}{\mathrm{P}(K_1)} = \frac{1/3}{2/3} = \frac{1}{2}$$

Andere redenering: twee van de vier kop-zijden, die alle dezelfde kans hebben, zijn zijden van de valse munt.

1c Nu is de situatie als volgt: met kans 1/2 heb je de valse munt, met kans 1/2 de zuivere munt. In het eerste geval is de kans op (weer) kop 1, in het tweede geval is die kans 1/2. Daaruit volgt: $\mathrm{P}(K_2 \mid K_1) = 1/2 \cdot 1 + 1/2 \cdot 1/2 = 3/4$.

2a $\mathrm{P}(1 \le X \le 4) = F(4) - F(1) = \sqrt{4}/2 - \sqrt{1}/2 = 1/2$, en analoog
$\mathrm{P}(3 \le X \le 9) = F(9) - F(3) = 1 - \sqrt{3}/2 = 0.134$.

2b $f(x) = F'(x) = 1/(4\sqrt{x})$ als $0 \le x \le 4$, en $f(x) = 0$ daarbuiten.
$\mathrm{E}[X] = \displaystyle\int_0^4 x \cdot \frac{1}{4\sqrt{x}}\,dx = \int_0^4 \tfrac{1}{4}\sqrt{x}\,dx = ... = \frac{4}{3}$.

3a Inkopper! $\mathrm{P}(Y = 0) = 0.10 + 0.05 + 0.05 + 0.05 = 0.25$.
$\mathrm{P}(X = 1 \mid Y = 0) = \mathrm{P}(X = 1, Y = 0)\,/\mathrm{P}(Y = 0) = 0.05/0.25 = 0.20$.

3b $\mathrm{P}(X = 1 \mid Y = 0) \neq \mathrm{P}(X = 1) = 0.25$, dus X en Y zijn afhankelijk.

3c Beetje gereken: $\mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$.
$\mathrm{E}[X] = 0 \cdot 0.25 + 1 \cdot 0.25 + 2 \cdot 0.25 + 3 \cdot 0.25 = 1.5$,
$\mathrm{E}[Y] = -1 \cdot 0.30 + 0 \cdot 0.25 + 1 \cdot 0.45 = 0.15$, en
$\mathrm{E}[XY] = (-1) \cdot 1 \cdot 0.10 + (-1) \cdot 2 \cdot 0.05 + (-1) \cdot 3 \cdot 0.05 + \ldots + 1 \cdot 1 \cdot 0.15 = 0.5$,
dus $\mathrm{Cov}(X, Y) = 0.5 - 1.5 \cdot 0.15 = 0.275$.

4 Het onderscheid normaal $\leftrightarrow$ exponentieel gaat via symmetrie: histogram III en empirische vdfs I en II horen bij exponentiële verdeling. (De boxplots passen trouwens ook niet vanwege de aanwezigheid van negatieve datapunten.)
Verder geldt voor een Exp(2) variabele X dat $\mathrm{P}(X \ge 4) = e^{-2\cdot 4} \approx 0.003$, wat (verreweg) het best past bij emp vdf II, terwijl de andere twee beter passen bij een Exp(1/2) verdeling: als $X \sim \mathrm{Exp}(1/2)$, dan $\mathrm{P}(X \ge 4) \approx 0.14$.
De normale datasets zijn te onderscheiden via mediaan en spreiding/kwartielen. histogram I, boxplot III: $\mathcal{N}(3, 3^2)$, histogram II, boxplot I: $\mathcal{N}(1, 1)$, en de overgeblevenen: emp.vdf. III en boxplot II: $\mathcal{N}(1, 3^2)$.
Kort samengevat:

	I	II	III
histogram	$\mathcal{N}(3,3^2)$	$\mathcal{N}(1,1)$	$\text{Exp}(1/2)$
emp. vdf	$\text{Exp}(1/2)$	$\text{Exp}(2)$	$\mathcal{N}(1,3^2)$
boxplot	$\mathcal{N}(1,1)$	$\mathcal{N}(1,3^2)$	$\mathcal{N}(3,3^2)$

5a $\mathrm{P}(Y=1)=p$.

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

5b $\mathrm{P}(Y>2)=(1-p)^2$ wordt geschat door $(1-\hat{p})^2\approx 0.34$.
$T_2=(1-T_1)^2$.

$$S_2=\frac{\text{aantal } X_i>2}{474};\quad \text{schatting } p^*=\frac{474-198-107}{474}\approx 0.36.$$

5c Een relatieve frequentie van een gebeurtenis (hier: $Y=1$ resp. $Y>2$) als schatter voor de kans op die gebeurtenis is altijd een zuivere schatter (zie H.19) dus T_1 en S_2 zijn zuiver.
$T_2=g(T_1)=(1-T_1)^2$; $g'(x)=2>0$, 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. $\mathrm{P}(Y>2)$.

5d Er geldt: $\mathrm{P}(Y=k)=(1-p)^{k-1}p$.

$$L(p)=C\cdot p^{198}\cdot((1-p)p)^{107}\left((1-p)^2p\right)^{55}\cdots\left((1-p)^{11}p\right)^{6}=\ldots=C\,p^{474}(1-p)^{811}$$

Waarbij C het aantal volgorde is voor "198 enen, 107 tweeën, enz".

6a Onder $H-0$ geldt $X\sim\text{Bin}(30,\,0.8)$.

6b Normale benadering, incl. cont.correctie

$$\mathrm{P}(X\le 20)=\mathrm{P}(X\le 20.5)=\mathrm{P}\left(\frac{X-np}{\sqrt{np(1-p)}}\le\frac{20.5-np}{\sqrt{np(1-p)}}=-1.5975\right)$$

en dat wordt hier, na invullen $n=30$, $p=0.8$

$$\mathrm{P}\left(\frac{X-np}{\sqrt{np(1-p)}}\le -1.5975\right)\approx\mathrm{P}(Z\le -1.5975)\approx 0.055$$

6c Gezocht: de grootste k zodat voor $X \sim \text{Bin}(30,\, 0.8)$ geldt

$$\text{P}\,(X \leq k) \leq 0.10$$

Dit kan via trial and error met de rekenmachine (functie voor binomiale verdeling), en dan vind je $\text{P}\,(X \leq 20) = 0.061$ en $\text{P}\,(X \leq 21) = 0.13$. Het kritieke gebied wordt dus $\{0,\, 1\, 2\, \ldots, 20\}$. Alternatief: weer via de normale benadering: vind grootste k zodat

$$\text{P}\left(Z \leq \frac{k + 0.5 - np}{\sqrt{np(1-p)}} < 0.10\right)$$

Daarvoor moet $\dfrac{k + 0.5 - 24}{\sqrt{4.8}} \leq z_{0.1} = -1.28,$
omgerekend $k \leq 23.5 - 1.28 \cdot \sqrt{4.8} \approx 20.7$, dus, daar k geheel moet zijn: $k = 20$.

6d Voor $X \sim \text{Bin}(30,0.8)$, dus onder H_0, geldt $P_0 = \text{P}\,(X = 20) = \binom{30}{20} 0.8^{20}\, 0.2^{10}$, en voor $X \sim \text{Bin}(30,0.5)$, onder H_1 dus, is die kans $P_1 = \binom{30}{20} 0.5^{20}\, 0.5^{10} = \binom{30}{20} 0.5^{30}$
Op elkaar delen geeft

$$\frac{P_1}{P_0} = \frac{0.5^{20}}{0.8^{20}\, 0.2^{10}} \approx 0.79 < 1.$$

Dus voor $k = 20$ geldt $P_1 < P_0$, en H_0 wordt niet verworpen.

6e Gezocht: alle k waarvoor $\text{P}(X = k \,|\, p = 0.5) > \text{P}(X = k \,|\, p = 0.8)$.
Onder H_0 ligt de verdeling gecentreerd rond 24, en voor H_1 rond 15. Het is duidelijk dat er een grens k_0 is zodat $P_1 > P_0$ precies voor $k \leq k_0$. Dit kan ook weer met trial and error, of als volgt:

$$\begin{aligned}
P_1 > P_0 \quad &\Leftrightarrow \quad 0.5^{30} > 0.8^k \cdot 0.2^{30-k} \quad \text{(de bin.coeffn doen er niet toe)} \\
&\Leftrightarrow \quad 0.5^{30} > \left(\frac{0.8}{0.2}\right)^k \cdot 0.2^{30} \\
&\Leftrightarrow \quad \left(\frac{0.8}{0.2}\right)^k = 4^k < \frac{0.5^{30}}{0.2^{30}} = 2.5^{30} \\
&\Leftrightarrow \quad \ln\left(4^k\right) = k \ln 4 < \ln\left(2.5^{30}\right) = 30 \ln 2.5 \\
&\Leftrightarrow \quad k < 30 \ln 2.5 / \ln 4 = 19.82
\end{aligned}$$

En daar nog steeds k geheel moet zijn: $k \leq 19$.

NORMEN:

1. 3×1; **2.** $3 + 4$; **3a.** 1+2; **3bc.** 2+3; **4.** 6; **5a.** 1+1+1; **5b.** 1+1+1; **5c.** 1+2; **5d.** 3; **6.** $2 + 4 + 2 + 2 + 2$.