Tentamen Kanstat
WI1275TA
30 juni 2010, 9 – 12 uur

Bij dit examen is het gebruik van een (evt. grafische) rekenmachine toegestaan. Een formuleblad wordt bijgeleverd. 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 (of Engels).

---

1. Op een stuk weg is op 25 verschillende dagen de verkeersdrukte gemeten in de ochtendspits, tussen 8 en 9 uur 's ochtends. De volgende waarnemingen zijn gedaan, waarbij de getallen het aantal gepasseerde auto's per uur geven (in volgorde van grootte geordend) :

2034	2213	3856	4057	4223
4501	4595	4620	4632	4660
4825	5001	5159	5362	5518
5534	5556	6058	6567	7190
7225	7232	7842	8190	8720

a. Bereken mediaan, eerste kwartiel en derde kwartiel van deze dataset.

b. Geef de waarde van de empirische verdelingsfunctie in 5000.

c. Om meer over de dataset te weten te komen wordt een genormaliseerd histogram gemaakt (dwz. een histogram, zodat het totale oppervlak onder de curve gelijk is aan 1). Geef de celhoogten van het histogram, dat begint in 1000 en een celbreedte gelijk aan 2000 heeft, m.a.w. geef de celhoogte van de cellen $[1000, 3000)$, $[3000, 5000)$, $[5000, 7000)$, $[7000, 9000)$. *Er wordt niet gevraagd het histogram te tekenen!*

d. Teken de boxplot bij deze dataset. Geef daarin bij elk relevant punt (uiteinden van de box, uiteinden van de 'snorharen', ...) de bijbehorende getalwaarde aan.

2. In een laboratorium moeten 1000 monsters onderzocht/getest worden op aanwezigheid van sporen van zware metalen. Tests kosten geld, dus de managers willen het onderzoek efficiënt inrichten. Er is na eerdere dergelijke onderzoeken bekend dat in circa 2 procent van de monsters zich daadwerkelijk sporen bevinden. We nemen aan (model!): de kans dat een willekeurig monster sporen bevat ('positief' is), is gelijk aan 0.02.

Een mogelijke aanpak is: voeg telkens gedeelten van tien monsters bij elkaar en test deze op sporen. Indien er in het mengsel sporen worden aangetroffen: test alle tien de monsters individueel.

a. Voer voor één groep van tien monsters de stochast X in, het aantal onderzoeken dat gedaan moet worden voor de betreffende groep. Geef de verdeling van X en bereken de verwachting van X.

b. Op de gegeven set van 1000 monsters: wat is het verwachte aantal tests dat moet worden uitgevoerd? (leg uit!)

Een slim managertje (dat nog in z'n proeftijd zit) stelt de volgende subtielere aanpak voor: verdeel het totaal in groepjes van 25, en als een groepje positief test, deel dit dan op in vijf groepjes van vijf. (En uiteraard: als een groepje van vijf positief test, dan moet elk van de vijf individueel getest worden). Laat in dit geval Y het aantal tests voor een groep van 25 zijn.

c. Welke waarden kan Y aannemen? (Denk eerst even na, voordat je een *toegelicht* antwoord geeft!)

Het wordt te ingewikkeld om voor elke waarde van Y de kans precies uit te rekenen. Echter, de kans dat er in een groep van 25 monsters meer dan drie positieve gevallen zitten is erg klein (en deze hoeven ook nog niet eens in meer dan drie verschillende groepjes van vijf te zitten).

d. Bereken de zojuist genoemde kans.

In het volgende onderdeel laten we de gebeurtenis 'meer dan drie positieven in een groep van 25' buiten beschouwing.
De kans dat er precies één van de vijf groepjes van vijf moet worden hertest is dan gelijk aan de kans dat er 1, 2 of 3 positieven zijn, *die zich bovendien in hetzelfde groepje van vijf bevinden.*

e. Gegeven dat een groep van 25 drie positieven bevat, hoe groot is dan de (*voorwaardelijke*) kans dat die zich in één groepje van vijf bevinden? Geef eerst een vaasmodel (hoeveel ballen, welke kleuren, hoe vaak trekken, met of zonder terugleggen) om deze kans te berekenen, en *bereken* vervolgens deze kans.

f. Al met al is de weg geplaveid om de kans te berekenen dat (in een set van 25) er precies één groepje van vijf minstens één positief monster bevat. Bereken deze kans.

Hier laten we het maar bij. Een precieze beschouwing wijst uit dat het mannetje wel een beetje gelijk heeft: met zijn aanpak is het verwachte aantal uit te voeren tests ongeveer een kwart lager dan bij de eerste methode.

3. Een stochast X heeft de volgende dichtheid:

$$f(x) = \begin{cases} 0\ , & \text{als} \quad |x| > 2 \\ c(4 - x^2), & \text{als} \quad -2 \leq x \leq 2 \end{cases}$$

waarbij c een door jou te bepalen constante is

a. Welke waarde moet c hebben?

(Als je dit onderdeel niet kunt oplossen: reken door met $c = 3/32.$)

b. Noem de verdelingsfunctie van X maar even F. Geef een algemene formule voor $F(x)$, voor $-2 \leq x \leq 2$, en ook voor $x \geq 2$.

c. Bereken de verwachting van X^2.

d. Bereken de dichtheid f_Y van de stochast $Y = \frac{1}{2}X$. (Aanwijzing: ga eerst na welke gevallen je moet onderscheiden.)

4. Op het eiland Quaboum bevinden zich twee vulkanen: de Sjöllofyll en de Pacuyo. De tijdstippen van uitbarstingen van beide worden gemodelleerd met behulp van onafhankelijke Poissonprocessen met intensiteiten $\lambda_S = 0.04$ en $\lambda_P = 0.025$ (uitbarstingen per jaar).

a. Bereken het verwachte totale aantal uitbarstingen in honderd jaar. Licht je antwoord duidelijk toe!

b. Bereken de kans dat er de komende 20 jaar geen enkele uitbarsting plaatsvindt.

c. Bereken de kans dat de eerste uitbarsting van de Pacuyo tussen de dertig en zestig jaar vanaf heden zal plaatsvinden.

Noem de tijdstippen vanaf heden tot de eerste uitbarsting van de Sjöllofyll en de Pacuyo S_1 respectievelijk T_1.
(Voor het gemak: S_1 heeft een Exp(0.04) verdeling.)
Laat verder X_1 het tijdstip zijn van de eerste uitbarsting van een van beiden.

d. Bereken de verdelingsfunctie van X_1.

e. Bereken de verwachting van X_1.

5. De hoeken van een gelijkbenige driehoek zijn α, α en β. Ze worden onafhankelijk van elkaar gemeten, en daarbij worden meetfouten gemaakt. We geven de metingen aan met T_1, T_2 en T_3. We nemen aan dat T_1, T_2 en T_3 zuivere schatters zijn voor α, α respectievelijk β, en dat ze dezelfde variantie σ^2 hebben. Bekijk de volgende combinaties:

$$U = \tfrac{1}{2}(T_1 + T_2), \quad \text{en} \quad V = U + \tfrac{1}{3}(\pi - 2U - T_3)$$

a. Toon aan dat U een zuivere schatter is voor α.

b. Bereken $\mathrm{Var}(U)$.

c. Ga na of V een zuivere schatter is voor α.

d. Bereken $\mathrm{Var}(V)$.

6. Om de nauwkeurigheid van een weeginstrument te onderzoeken wordt een serie metingen van een standaardgewicht van 10 gram uitgevoerd. De aannamen zijn dat de metingen normaal verdeeld zijn met (onbekende) verwachting μ, en standaardafwijking $\sigma = 0.04$ gram.

a. Stel het ijkgewicht wordt tien keer gewogen en het gemiddelde van de tien metingen is 10.056 gram. Geef een 98% betrouwbaarheidsinterval voor μ.

b. Hoeveel metingen moeten er (minimaal) gedaan worden om een 98 % betrouwbaarheidsinterval te krijgen dat korter is dan 0.02 gram?

c. Na een nieuwe ijking willen we een toets uitvoeren om te zien of deze goed heeft plaatsgevonden.
We nemen als nulhypothese $\mu = 10$ en als alternatieve hypothese $H_1 : \mu \neq 10$, en als significantieniveau $\alpha = 2\,\%$.
Er worden opnieuw tien (onafhankelijke) metingen (met standaardafwijking 0.04) gedaan. Deze leveren een gemiddelde waarde van 9.965 gram. Onderzoek of de nulhypothese wel of niet verworpen wordt.

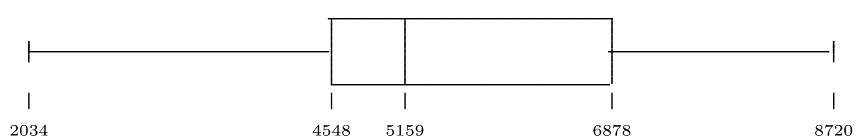
Antwoorden open vragen

1a Bij een oneven aantal n is de mediaan het middelste getal; hier is dat 5159.
Voor het eerste kwartiel: neem de algemene formule $q_n(p) = x_{(k)} + \alpha(x_{(k+1)} - x_{(k)})$, $k = \lfloor p(n+1) \rfloor$, $\alpha = p(n+1) - k$.
Voor $p = \frac{1}{4}$: $k = \lfloor \frac{1}{4} \cdot 26 \rfloor = 6$ en $\alpha = \frac{1}{4} \cdot 26 - 6 = \frac{1}{2}$, dus het eerste kwartiel wordt $x_{(6)} + \frac{1}{2}(x_{(7)} - x_{(6)}) = 4548$.
Net zo: $q_n(\frac{3}{4}) = x_{(19)} + \frac{1}{2}(x_{(20)} - x_{(19)}) = 6878.5$.

1b $F_{25}(5000) = \dfrac{\text{aantal punten } \leq 5000}{25} = \dfrac{11}{25} = 0.44.$

1c Voor cel i: $h_i = \dfrac{\text{aantal punten in cel } i}{n \cdot \text{lengte van cel } i}.$
Hier levert dat de hoogtes $\dfrac{2}{25 \cdot 2000} = \dfrac{2}{50000}$, $\dfrac{9}{50000}$, $\dfrac{8}{50000}$, resp. $\dfrac{6}{50000}$.

1d De randen van de box liggen bij het eerste resp. het derde kwartiel. De snorharen eindigen *hoogstens* 1.5 IQR ($= 1.5 \cdot (6878.5 - 4548) = 3496$, afgerond) buiten de box, en dat is hier (afgerond) binnen $(4548 - 3496,\ 6878.5 + 3496) = (1052,\ 10374)$.
Uitbijters: punten buiten $(1052, 10374)$, zijn er niet.
Dus: snorharen eindigen bij het maximum en het minimum.
Plaatje:



2a Er zijn twee mogelijkheden: indien er geen sporen zijn: $X = 1$, indien er wel sporen zijn: $X = 11$. De kans dat er geen sporen zijn in een set van 10 monsters: $\mathrm{P}(X = 1) = 0.98^{10} \approx 0.817$.
De verdeling van X wordt:

k	1	11
$\mathrm{P}(X = k)$	0.817	0.183

dus de verwachting: $\mathrm{E}[X] = 0.817 \cdot 1 + 0.183 \cdot 11 = 2.83$

2b Er zijn honderd sets van 10, dus het totale aantal tests is te schrijven als $T = X_1 + X_2 + \ldots + X_{100}$.
De verwachting wordt dan $\mathrm{E}[T] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \ldots + \mathrm{E}[X_{100}] \approx 100 \cdot 2.83 = 283$.

2c Als er geen positieve monsters zijn: $Y = 1$. Als er 1 positief monster is: de 'tweede ronde' levert minstens 1 set van 5 die een keer over moet. Dus het aantal uit te voeren tests wordt $1 + 5 + 5 \cdot$ (aantal sets van 5 met een positief monster), en dat wordt een van de waarden 11, 16, 21, 26, 31.
Dus: mogelijke waarden voor Y: 1, 11, 16, 21, 26, 31.

2d Het aantal positieve monsters in een groep van 25 heeft een Bin(25, p) verdeling, met $p = 0.02$.
De kans op meer dan 3 is gelijk aan (1 − de kans op hoogstens 3) =
$= 1 - (\binom{25}{0} p^0(1-p)^{25} + \binom{25}{1} p^1(1-p)^{24} + \binom{25}{2} p^2(1-p)^{23} + \binom{25}{3} p^3(1-p)^{22}) =$
$= 1 - (0.6035 + 0.3079 + 0.0754 + 0.0118) = 0.0014.$

2e Het probleem is te zien als drie keer trekken, zonder terugleggen, uit een vaas met vijf rode, vijf gele, vijf blauwe, vijf groene en vijf oranje ballen. (Vijf ballen van dezelfde kleur correponderen met een groepje van vijf dat tegelijk getest wordt.) Gevraagd wordt dan de kans dat de drie ballen dezelfde kleur hebben. Deze kans wordt: $\frac{25}{25} \cdot \frac{4}{24} \cdot \frac{3}{23} = \frac{1}{46}$. Toelichting: $\frac{3}{23}$ is de kans dat de derde bal dezelfde kleur heeft als de eerste twee, *gegeven dat* de eerste twee dezelfde kleur hebben.

Opmerking: het model: 5 keer trekken uit een vaas met 3 zwarte en 22 witte ballen, en dan de kans berekenen dat je de drie zwarte trekt is fout; de 3 zwarte moeten 'in een setje van 5' zitten, maar dat hoeft niet per se het setje te zijn dat getrokken wordt. De zojuist beschreven kans is bij nader inzien een factor 5 te klein.

2f Er zijn drie mogelijkheden:

— Er is 1 positief monster; de kans hierop: $\binom{25}{1} p^1(1-p)^{24} = 0.3079$;

— Er zijn 2 positieve monsters, en die zitten in hetzelfde groepje van vijf; de kans hierop: $\binom{25}{2} p^1(1-p)^{23} \cdot \frac{25}{25} \cdot \frac{4}{24} = 0.0754 \cdot 0.1667 = 0.0126$;

— Er zijn 3 positieve monsters, in hetzelfde groepje van vijf; de kans hierop: $\binom{25}{3} p^1(1-p)^{22} \cdot \frac{1}{46} \approx 0.00026$.

Optellen van deze drie kansen (de derde blijkt ook wel te verwaarlozen) geeft het antwoord: ($\approx$) 0.32.

3a De totale kansmassa, $\int_{-\infty}^{\infty} f(x)\,dx$ moet gelijk zijn aan 1:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-2}^{2} c(4-x^2)\,dx = c\left[4x - \frac{x^3}{3}\right]_{-2}^{2} = \tfrac{32}{3}c$$

Dus c moet gelijk zijn aan $\frac{3}{32}$.

3b Het is evident dat $F(x) = 1$, vanaf $x = 2$. Voor $-2 \leq x \leq 2$

$$F(x) = \int_{-2}^{x} c(4-t^2)\,dt = \ldots = \tfrac{3}{32}\,(4x + \tfrac{16}{3} - \tfrac{1}{3}x^3) = \tfrac{3}{8}x - \tfrac{1}{32}x^3 + \tfrac{1}{2}$$

hetgeen voor $x = 2$ inderdaad levert $F(2) = \frac{3}{32}\,(8 + \frac{16}{3} - \frac{8}{3}) = \frac{3}{32} \cdot \frac{32}{3} = 1.$

3c

$$\mathrm{E}\left[X^2\right] = \int x^2 f(x)\,dx = c\int_{-2}^{2} 4x^2 - x^4\,dx = \ldots = \frac{3}{32}\cdot\frac{128}{15} = \frac{4}{5}$$

3d Aangezien X waarden aanneemt van -2 tot en met 2, zal $\frac{1}{2}X$ waarden aannemen tussen $\frac{1}{2}\cdot(-2)$ en $\frac{1}{2}\cdot 2$, oftewel het bereik van Y is het interval $[-1,\,1]$. Daaruit volgt al meteen $f_Y(y) = 0$ als $y \neq [-1,\,1]$.
Voor y in $[-1,\,1]$ gaat een afleiding via de verdelingsfuncties het handigst. Stel $-1 \leq y \leq 1$; dan:

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}\big(\tfrac{1}{2}X \leq y\big) = \mathrm{P}(X \leq 2y) = F_X(2y)$$

Uit $-1 \leq y \leq 1$ volgt $-2 \leq 2y \leq 2$, en op het interval $[-2, 2]$ hadden we al gevonden dat $F_X(x) = \frac{3}{32}\,(4x + \frac{16}{3} - \frac{1}{3}x^3)$,
dus zal $F_Y(y) = \ldots = F_X(2y) = \frac{3}{32}\,(8y + \frac{16}{3} - \frac{1}{3}\cdot 8y^3) = \frac{3}{4}y + \frac{1}{2} - \frac{1}{4}y^3$,
en een keer differentiëren geeft de dichtheid: $f_Y(y) = \frac{3}{4} - \frac{3}{4}y^2$.

4a Geef met N_S en N_P de aantallen uitbarstingen van Sjöllofyll en Pacuyo aan in het interval $[0, 100]$. Dan gelden: $N_S \sim \mathrm{Pois}(0.04\cdot 100)$, dus $\mathrm{E}[N_S] = 0.04\cdot 100 = 4$, en evenzo $\mathrm{E}[N_P] = 0.025\cdot 100 = 2.5$, en het verwachte totale aantal wordt $\mathrm{E}[N_S + N_P] = 6.5$.

4b Hier wordt gevraagd $\mathrm{P}(N_S = 0,\, N_P = 0)$, met nu $N_S \sim \mathrm{Pois}(0.04\cdot 20) = \mathrm{Pois}(0.8)$ en $N_S \sim \mathrm{Pois}(0.5)$. Vanwege de veronderstelde onafhankelijkheid wordt dit $\mathrm{P}(N_S = 0)\,\mathrm{P}(N_P = 0) = e^{-0.8}\cdot e^{-0.5} = 0.2725$

4c Dit is de kans op de gebeurtenis $A \cap B$, waarbij A: geen uitbarsting in $[0, 30]$ en B: minstens 1 uitbarsting in $[30, 60]$. Vanwege de onafhankelijkheid van A en B (Poissonproces!) geldt

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B) = e^{-0.75}\cdot(1 - e^{-0.75}) = 0.249$$

4d
$$\begin{aligned}\mathrm{P}(X_1 > t) &= \text{de kans dat geen van de twee vulkanen uitbarst in } [0, t]\\ &= e^{-0.04t}\cdot e^{-.025t} = e^{-0.065t}\end{aligned}$$
Dus ook: $\mathrm{P}(X_1 \leq t) = 1 - e^{-0.065t}$, voor $t > 0$,
hetgeen betekent dat X_1 een Exp(0.65) verdeling heeft.

4e Uit **4.d** (en het formuleblad) volgt: $\mathrm{E}[X_1] = \dfrac{1}{0.065} \approx 15.4$ (jaar).

5a

$$\mathrm{E}[U] = \mathrm{E}\left[\tfrac{1}{2}(T_1 + T_2)\right] = \tfrac{1}{2}\mathrm{E}[T_1] + \tfrac{1}{2}\mathrm{E}[T_2] = \tfrac{1}{2}\alpha + \tfrac{1}{2}\alpha = \alpha.$$

5b U is het gemiddelde van twee onafhankelijke stochasten met dezelfde verdeling, dus $\mathrm{Var}(U) = \frac{1}{2}\mathrm{Var}(T_1) = \frac{\sigma^2}{2}$.

5c

$\mathrm{E}[V] = \mathrm{E}\big[U + \frac{1}{3}(\pi - 2U - T_3)\big] = \mathrm{E}[U] + \frac{1}{3}\mathrm{E}[\pi - 2U - T_3] = \alpha + \frac{1}{3}(\pi - 2\alpha - \beta) = \alpha$
want de $2\alpha + \beta$, zijnde de som van de hoeken van een driehoek, is gelijk aan π.

5d

$V = U + \frac{1}{3}(\pi - 2U - T_3) = \frac{1}{2}(T_1 + T_2) + \frac{1}{3}(\pi - 2\cdot(\frac{1}{2}(T_1 + T_2)) - T_3) = \frac{1}{3}\pi + \frac{1}{6}T_1 + \frac{1}{6}T_2 - \frac{1}{3}T_3$
De drie stochasten in de laatste uitdrukking zijn onafhankelijk, dus geldt

$$\mathrm{Var}(V) = (\tfrac{1}{6})^2\mathrm{Var}(T_1) + (\tfrac{1}{6})^2\mathrm{Var}(T_2) + (\tfrac{1}{3})^2\mathrm{Var}(T_3) = \frac{\sigma^2}{36} + \frac{\sigma^2}{36} + \frac{\sigma^2}{9} = \frac{\sigma^2}{6}$$

Iets eenvoudiger:

$$\mathrm{Var}(V) = \mathrm{Var}\left(\frac{\pi}{3} + \frac{1}{3}U - \frac{1}{3}T_3\right)$$

waarbij U en T_3 ook weer onafhankelijk zijn. Dus volgt

$$\mathrm{Var}(V) = (\tfrac{1}{3})^2\mathrm{Var}(U) + (\tfrac{1}{3})^2\mathrm{Var}(T_3) = \frac{1}{9}\frac{\sigma^2}{2} + \frac{1}{9}\sigma^2 = \frac{\sigma^2}{6}$$

6a Het gemiddelde van 10 metingen heeft een $\mathcal{N}(\mu, \sigma^2/10)$ verdeling. Een 98 % betrouwbaarheidsinterval heeft de grenzen

$$\bar{x}_{10} \pm z_{0.01} \cdot \frac{0.04}{\sqrt{10}}, \text{ waarbij } z_{0.01} = 2.326 \text{ (afgerond: 2.33)}$$

dus dat wordt $[10.027, 10.085]$.

6b Het interval wordt bij n metingen kleiner dan 0.02 gram als

$$2z_{0.01}\frac{0.04}{\sqrt{n}} \leq 0.02$$

dus als

$$\sqrt{n} \geq \frac{2z_{0.01}0.04}{0.02} \approx 9.32$$

Oftewel, als $n \geq 9.32^2 = 86.8...$, dus, aangezien n geheel moet zijn: $n \geq 87$.

6c De gevonden waarde 9.965 ligt – als de nulhypothese waar is – in de linkerstaart van de verdeling (van $\bar{X}_{10}$).
We moeten kijken of de P-waarde, dus $\mathrm{P}\big(\bar{X}_{10} \leq 9.965\big)$ groter dan of kleiner dan 0.01 ($= \frac{1}{2}\alpha$) is. Onder de nulhypothese geldt:

$$\mathrm{P}\big(\bar{X}_{10} \leq 9.965\big) = \mathrm{P}\left(\frac{\bar{X}_{10} - 10}{0.04/\sqrt{10}} \leq \frac{9.965 - 10}{0.04/\sqrt{10}}\right) = \mathrm{P}(Z \leq -2.767) \approx 0.0028 < 0.01$$

dus de nulhypothese wordt verworpen.

Alternatief: het 98 % betrouwbaarheidsinterval bij 9.965 wordt

$$[\bar{x}_{10} - z_{0.01}\cdot\frac{0.04}{\sqrt{10}}, \bar{x}_{10} + z_{0.01}\cdot\frac{0.04}{\sqrt{10}}] = [9.965 - 0.0296, 9.965 + 0.0296] = [9.9354, 9.9946]$$

en dat bevat de waarde 10 niet , dus H_0 wordt verworpen.