Tentamen Statistische Methoden
MST
17 april 2015, 14.00–17.00 uur

Normering: MC vraag elk 0.4 punt. Open vragen allebei 2 punten.

1. Van de gebeurtenissen A en B is bekend dat $P(A) = 1/3$, $P(B) = 1/4$ en $P(A \mid B) = 2/3$. Bereken $P(A \cup B)$.

a. 1/3 **b.** 1/4 **c.** 3/4 **d.** 5/12 **e.** 7/12 **f.** 1/2

2. De leugendetector wordt vooral in de VS veel ingezet bij het opsporen van misdadigers. We beschrijven dit in kansrekeningtermen: A is de gebeurtenis dat de verdachte schuldig is; B is de gebeurtenis dat de leugendetector de verdachte als schuldig aanwijst. Het is bekend dat $P(B \mid A) = 0.90$ en $P(B \mid A^c) = 0.10$ en $P(A) = 0.10$. Stel dat de leugendetector de verdachte aanwijst als schuldig. Hoe groot is dan de kans dat de verdachte daadwerkelijk schuldig is?

a. 0.50 **b.** 0.90 **c.** 0.20 **d.** 0.60 **e.** 0.40 **f.** 0.10

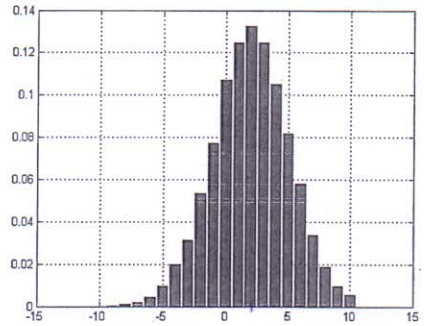
3. De stochasten $X_1, \ldots, X_{100}$ zijn onafhankelijk en identiek verdeeld, met verwachting 1 en variantie 9. Bereken de kans dat de som van deze honderd stochasten groter dan 115 is, via de centrale limietstelling.

a. 0.13 **b.** 0.87 **c.** 0.31 **d.** 0.69 **e.** 0.52 **f.** 0.48

4. De stochasten X, Y nemen alleen waarden aan tussen 0 en 1. De simultane dichtheidsfunctie is $f(x, y) = 6x^2y$ voor $0 \leq x, y \leq 1$. Bereken $P(X > Y)$.

a. 0.2 **b.** 0.3 **c.** 0.4 **d.** 0.5 **e.** 0.6 **f.** 0.7

5. Het volgende histogram representeert een normaal verdeelde dataset. Welke waarden van μ en σ passen het best bij dit histogram?



a. $\mu = 1; \sigma = 1$ **b.** $\mu = 1; \sigma = 2$ **c.** $\mu = 1; \sigma = 3$
d. $\mu = 2; \sigma = 1$ **e.** $\mu = 2; \sigma = 2$ **f.** $\mu = 2; \sigma = 3$

6. Ik zet de getallen 1 tot en met 5 random op een rij en ik ben benieuwd of de getallen 2 en 3 naast elkaar komen te staan. Bijvoorbeeld, in de randomrij 5, 3, 2, 4, 1 staan 2 en 3 naast elkaar. In de randomrij 3, 4, 1, 5, 2 staan ze niet naast elkaar. Hoe groot is de kans dat de getallen 2 en 3 naast elkaar komen te staan?

a. 0.1 b. 0.2 c. 0.3 d. 0.4 e. 0.5 f. 0.6

7. Welke verdeling past het best bij de volgende dataset:

6	1	3	3	4
1	0	2	3	2
2	0	2	1	1
5	1	2	3	5

a. Bin(20,0.2) b. Bin(11,0.2) c. Bin(10,0.5)
d. Bin(6,0.5) e. Geo(0.2) f. Geo(0.4)

8. Van de stochasten X en Y is bekend dat $\text{Var}(X) = 1$ en $\text{Var}(Y) = 2$ en dat $\text{Cov}(X, Y) = 2$. Bereken de variantie van $X - 2Y$.

a. -1 b. 1 c. 3 d. 5 e. 7 f. 9

9. Volgens de ongelijkheid van Chebyshev geldt dat

$$P\left(\left|\frac{X-\mu}{\sigma}\right| \geq k\right) \leq \frac{1}{k^2}$$

Stel dat $E[X] = 1$ en $\text{Var}(X) = 3$. Hoe groot is de kans dat X in het interval $(-2, 4)$ ligt volgens de ongelijkheid van Chebyshev?

a. ≤ 0.25 b. ≤ 0.33 c. ≤ 0.5 d. ≥ 0.5 e. ≥ 0.66 f. ≥ 0.75

10. De stochast X is $U(-1, 1)$ verdeeld. Bereken de variantie van X^3.

a. 1 b. $\frac{2}{3}$ c. $\frac{1}{2}$ d. $\frac{2}{5}$ e. $\frac{1}{3}$ f. $\frac{2}{7}$

11. Een dataset van 16 metingen heeft gemiddelde $\bar{x}_{16}$ gelijk aan 3.26 en steekproef-standaardafwijking s_{16} is 0.12. Zoals bij elke meting mag je ervan uitgaan dat de 16 metingen normaal verdeeld zijn. Een 90 procent betrouwbaarheidsinterval voor $\bar{x}_{16}$ is van de vorm

$$(3.26 - x, 3.26 + x)$$

Bepaal de juiste waarde van x, afgerond op 1 honderste.

a. 0.01 b. 0.02 c. 0.03 d. 0.04 e. 0.05 f. 0.06

12. Bij paardenraces blijkt dat er relatief vaak gewonnen is door een paard dat is gestart in de eerste van de 8 banen. In 100 races werd er 21 keer gewonnen door een paard in baan 1. Dit is reden om te veronderstellen dat paarden in deze baan een voordeel hebben. Hoe test je deze veronderstelling met hypothesen? Noteer de winstkans voor een paard in baan 1 met p.

a. $H_0: p \leq 1/8$; $H_1: p = 0.21$ b. $H_0: p = 0.21$; $H_1: p > 0.21$
c. $H_0: p = 0.21$; $H_1: p < 1/8$ d. $H_0: p = 1/8$; $H_1: p > 1/8$
e. $H_0: p = 1/8$; $H_1: p = 0.21$ f. $H_0: p \neq 1/8$; $H_1: p > 1/8$

13. De stochast X neemt waarden aan tussen -1 en 1 en heeft kansdichtheidsfunctie $f(x) = kx^4$ voor $-1 < x < 1$. Bepaal de constante k.

a. 1 b. 1.5 c. 2 d. 2.5 e. 3 f. 3.5

14. We voeren een t-test uit met nulhypothese $H_0 : \mu = 10$ met een dataset van 9 elementen met gemiddelde $\bar{x}_9 = 10.28$ en steekproef standaardafwijking $s_9 = 0.21$. De alternatieve hypothese is $\mu > 10$. Het significantieniveau is 0.01. De nulhypothese wordt verworpen indien het gemiddelde $\bar{x}_9$ ligt boven de kritieke waarde c. Hoe groot is c, afgerond op 5 honderste?

a. 10.10 **b.** 10.15 **c.** 10.20 **d.** 10.25 **e.** 10.30 **f.** 10.35

15. Gegeven de dataset

$$1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 5, 6, 100$$

Bepaal de MAD.

a. 1 **b.** 2 **c.** 3 **d.** 5 **e.** 6 **f.** 100

OPEN VRAGEN

---

1. De stochast X is $Exp(1)$ verdeeld. De stochast Y is het kwadraat van X, dus $Y = X^2$.

A. Bereken de kansverdelingsfunctie F_Y van de stochast Y.

B. Bereken de dichtheidsfunctie f_Y van Y.

C. Bereken de verwachting van Y.

2. In de plantkunde verdeeld men bladeren in 4 types: zetmeel-groen, suiker-wit, zetmeel-wit, suiker-groen. Volgens de genetica komen deze types voor met kansen $\frac{1}{4}(\theta+2), \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta)$ voor een onbekende parameter θ tussen nul en een. Stel dat men een dataset heeft met daarin 100 bladeren. Dan is het aantal bladeren met een zetmeel-groen type $Bin(100, p_1)$ verdeeld, met $p_1 = \frac{1}{4}(\theta + 2)$. Noteer dit aantal met N_1. Op dezelfde manier, laat N_2 het aantal suiker-witte bladeren zijn. Deze stochast is $Bin(100, p_2)$ met $p_2 = \frac{1}{4}\theta$. Beschouw de twee schatters

$$T_1 = aN_1 - b, \; T_2 = cN_2$$

A. Voor welke waarde van c is T_2 een zuivere schatter voor θ?

B. Voor welke waarden van a, b is T_1 een zuivere schatter voor θ?

C. Vergelijk de variantie van deze twee schatters. Welke schatter heeft de voorkeur, T_1 of T_2?

Formuleblad bij tentamens wi1102CT

NA AFLOOP VAN HET TENTAMEN INLEVEREN!!

N.B.: DIT BLAD IS NIET EEN SAMENVATTING OF OVERZICHT, EN DIENT SLECHTS ALS HULPMIDDEL.

Kansverdelingen

1. **Bernoulli verdeling:** $Ber(p)$.

   $\mathrm{P}(X=1)=p$ en $\mathrm{P}(X=0)=1-p$. $\mathrm{E}[X]=p$; $\mathrm{Var}(X)=p(1-p)$.
2. **Binomiale verdeling:** $Bin(n,p)$.

   $\mathrm{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ voor $k=0,1,\ldots,n$. $\mathrm{E}[X]=np$; $\mathrm{Var}(X)=np(1-p)$.
3. **Geometrische verdeling :** $Geo(p)$.

   $\mathrm{P}(X=k)=p(1-p)^{k-1}$ voor $k=1,2,\ldots$. $\mathrm{E}[X]=1/p$; $\mathrm{Var}(X)=(1-p)/p^2$.
4. **Poisson-verdeling:** $Pois(\mu)$.

   $\mathrm{P}(X=k)=\frac{\mu^k}{k!}e^{-\mu}$ voor $k=0,1,\ldots$. $\mathrm{E}[X]=\mu$; $\mathrm{Var}(X)=\mu$.
5. **Exponentiële verdeling:** $Exp(\lambda)$.

   $f(x)=\lambda e^{-\lambda x}$ en $F(x)=1-e^{-\lambda x}$ voor $x\geq 0$. $\mathrm{E}[X]=1/\lambda$; $\mathrm{Var}(X)=1/\lambda^2$.
6. **Uniforme verdeling** op $[a,b]$: $U(a,b)$.

   $f(x)=\frac{1}{b-a}$ en $F(x)=\frac{x-a}{b-a}$ voor $a\leq x\leq b$. $\mathrm{E}[X]=\frac{1}{2}(a+b)$; $\mathrm{Var}(X)=\frac{1}{12}(b-a)^2$.
7. **Normale verdeling:** $N(\mu,\sigma^2)$.

   $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ en $F(x)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}dt$. $\mathrm{E}[X]=\mu$; $\mathrm{Var}(X)=\sigma^2$.
8. **Pareto verdeling** $Par(\alpha)$:

   $f(x)=\frac{\alpha}{x^{\alpha+1}}$ en $F(x)=1-x^{-\alpha}$ voor $x\geq 1$. $\mathrm{E}[X]=\infty$ voor $0<\alpha\leq 1$.

   $\mathrm{E}[X]=\frac{\alpha}{\alpha-1}$ voor $\alpha>1$. $\mathrm{Var}(X)=\frac{\alpha}{(\alpha-1)^2(\alpha-2)}$ voor $\alpha>2$.

Covariantie en correlatie

1. Definitie covariantie: $\mathrm{Cov}(X,Y)=\mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]$.
   Eigenschappen: $\mathrm{Cov}(X,Y)=\mathrm{E}[XY]-\mathrm{E}[X]\cdot\mathrm{E}[Y]$, $\mathrm{Cov}(rX+s,tY+u)=rt\,\mathrm{Cov}(X,Y)$.
2. Definitie correlatiecoëfficiënt: $\rho(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$.
3. $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$.

Empirische verdelingsfunctie

Voor een dataset van n elementen: $F_n(x)=\frac{\text{aantal elementen in de dataset}\leq x}{n}$.

Centrale limietstelling

Als $X_1,X_2,\ldots$ een rij onafhankelijke stochasten is met alle dezelfde verdeling en met verwachting μ and variantie σ^2, dan geldt:

Centrale limietstelling:

$$\lim_{n\to\infty}\mathrm{P}\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\leq a\right)=\mathrm{P}(Z\leq a),$$

waarbij Z een $N(0,1)$ verdeling heeft.

Schatters

De onzuiverheid (bias) van een schatter T voor θ: $\mathrm{E}[T]-\theta$.
Als T_1 en T_2 zuivere schatters voor θ zijn, dan heet T_2 efficiënter dan T_1 als $\mathrm{Var}(T_2)<\mathrm{Var}(T_1)$.
De 'mean squared error' van een schatter T voor θ: $\mathrm{MSE}(T)=\mathrm{E}[(T-\theta)^2]$.
Eigenschap: $\mathrm{MSE}(T)=\mathrm{Var}(T)+(\mathrm{E}[T]-\theta)^2$.

Betrouwbaarheidsintervallen

Gegeven: een dataset $x_1,x_2,\ldots,x_n$, α een getal tussen 0 en 1, en steekproefgrootheden L_n en U_n zó dat $\mathrm{P}(L_n<\theta<U_n)=1-\alpha$ voor elke waarde van θ. Dan is (l_n,u_n), met l_n en u_n bepaald uit de data, een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor θ.

Gestandaardiseerd en gestudentiseerd steekproefgemiddelde

Als X_i een $N(\mu,\sigma^2)$ verdeling hebben voor $i=1,\ldots,n$, en onafhankelijk zijn, dan hebben

$$\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \quad \text{en} \quad \frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}$$

een $N(0,1)$, respectievelijk $t(n-1)$–verdeling. Hierbij is $t(n-1)$ de t-verdeling met $n-1$ vrijheidsgraden.

Regressie model

$Y_i=\alpha+\beta x_i+U_i$ met U_i onafhankelijk, met $\mathrm{E}[U_i]=0$, en $\mathrm{Var}(U_i)=\sigma^2$.

Schatter voor σ^2: $\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2$.

Schatter voor β: $\hat{\beta}=\frac{n\sum x_iy_i-(\sum x_i)(\sum y_i)}{n\sum x_i^2-(\sum x_i)^2}$, $S_b^2=\frac{n}{n\sum x_i^2-(\sum x_i)^2}\hat{\sigma}^2$.

Schatter voor α: $\hat{\alpha}=\bar{y}_n-\hat{\beta}\bar{x}_n$, $S_a^2=\frac{\sum x_i^2}{n\sum x_i^2-(\sum x_i)^2}\hat{\sigma}^2$.

S_b^2 is een schatter voor de variantie van $\hat{\beta}$, S_a^2 is een schatter voor de variantie van $\hat{\alpha}$.

Twee steekproeven

Bij de twee steekproeven t-toets:

Variantie gepoold: $S_p^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)$, ongepoold: $S_d^2=\frac{S_X^2}{n}+\frac{S_Y^2}{m}$.