Tentamen toegepaste statistiek voor TA (WI1275TA)

14 april 2014, 14:00-17:00

Afdeling Toegepaste Wiskunde, Faculteit Elektrotechniek, Wiskunde en Informatica

---

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.
Rekenmachine toegestaan. Formuleblad/boek/aantekeningen niet toegestaan.

---

1. Bedrijf AAA gaat ieder van haar 1000 werknemers testen op het gebruik van drugs. Het management schat dat 1% van de werknemers drugs gebruikt. De te gebruiken test is 98% accuraat (als iemand daadwerkelijk drugs gebruikt, dan wordt dat in 98% van de gevallen door de test ook aangegeven). De test heeft 3% vals-positieven (dat wil zeggen: met kans 0.03 geeft de test een positieve uitslag voor mensen die geen drugs gebruikt hebben).

   (a) Een zeker persoon test positief. Wat is de kans dat deze persoon ook daadwerkelijk drugs heeft gebruikt?

   (b) Als een persoon negatief test, wat is dan de kans dat hij/zij drugs heeft gebruikt?

2. Een computer bestand met 1000 huishoudens heeft komens variërend van 5800$ tot 98600$. Per ongeluk krijgt het hoogste inkomen een extra nul en wordt dus geregistreerd als 986000$.

   (a) Wordt het steekproefgemiddelde hierdoor beïnvloed? Indien dit zo is, met hoeveel?

   (b) Wordt de steekproefmediaan hierdoor beïnvloed? Indien dit zo is, met hoeveel?

3. Uit ervaring weet een docent dat het cijfer dat een student voor het eindexamen haalt een stochastische variabele is met gemiddelde 7,5 and standaard deviatie 0,8. Hoeveel studenten moeten aan het eindexamen deelnemen zodat met kans tenminste 0,95 het het gemiddelde eindcijfer van alle studenten tenminste 7,3 is?

   Hint: gebruik een benadering met behulp van de centrale limietstelling.

4. Gegeven is de stochastische variabele X met kansdichtheid

$$f(x) = \begin{cases} x & \text{als } x \in [0,1] \\ \frac{1}{2} & \text{als } x \in (1,2] \\ 0 & \text{voor alle andere waarden van } x \end{cases}$$

   (a) Bereken de verdelingsfunctie F van X. Dus: bereken de $F(x)$ voor alle $x \in \mathbb{R}$.

   (b) Geef aan hoe je een realisatie van X kunt genereren met behulp van een toevalsgetal u. Maak onderscheid tussen de gevallen $u \in (0, 1/2)$ en $u \in [1/2, 1)$.

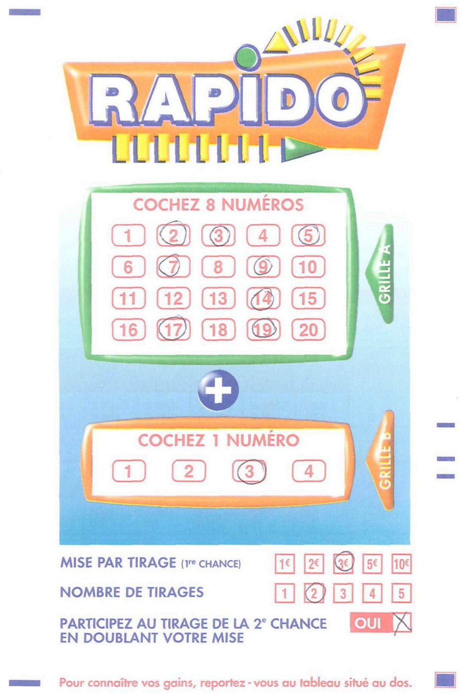

Figure 1: Invulformulier bij Rapido.

5. In Franse cafés wordt het spel Rapido gespeeld. We bekijken hier een ietwat vereenvoudigde opzet. Het spel gaat als volgt. Er zijn 2 ronden. Bij iedere ronde vul je een rooster in (”grille A” en ”grille B”). Zie figuur 1. Bij ronde A kies je 8 getallen uit de verzameling $\{1, 2, \ldots, 20\}$. Bij ronde B kies je één getal uit de getallen $\{1, 2, 3, 4\}$. Als je het formulier hebt ingevuld, dan zie je vervolgens op een monitor 8 getallen verschijnen die getrokken worden bij ronde A. Vervolgens zie je nog één getal verschijnen dat getrokken wordt bij ronde B. We gaan ervan uit dat de computer de 8 getallen bij ronde A willekeurig zonder terugleggen kiest uit de verzameling $\{1, 2, \ldots, 20\}$. Net zo gaan we ervan uit dat de computer bij ronde B een willekeurig getal uit $\{1, 2, 3, 4\}$ kiest. Verder nemen we aan dat de uitkomsten bij ronde A en ronde B onafhankelijk van elkaar worden gegenereerd.

   a. Bereken de kans p dat je bij ronde A precies 5 getallen goed hebt gegokt.

   b. Voor de eenvoud stellen we nu dat je alleen een prijs wint als je er precies 5 goed hebt gegokt (in werkelijkheid win je natuurlijk ook als je er meer dan 5 goed hebt gegokt). De hoogte van je winst hangt af van het resultaat bij ronde B. Als je (naast de 5 goed gegokte getallen bij ronde A), ook bij ronde B wint, dan krijg je 5 keer je inzet terug. Als je na de 5 goed gegokte getallen vervolgens bij ronde B verliest, dan krijg je 2 keer je inzet terug. Bereken je verwachte winst als je 1 euro inzet.
   Indien je het antwoord op vraag **a** *niet weet, neem dan* $p = 0.15$.

   c. Wat is de kans dat je bij ronde A alle getallen goed hebt, en ook bij ronde B het goede getal hebt gegokt?

6. Beschouw de steekproef $Y_1, ..., Y_n$ uit een exponentiële verdeling met parameter θ. De dichtheid van iedere Y_i wordt dus gegeven door

$$f(y) = \begin{cases} \theta e^{-\theta y} & y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Noteer met $Y_{(1)} = \min(Y_1, ..., Y_n)$ de kleinste waarde in de steekproef.

(a) Laat zien dat $Y_{(1)}$ een exponentiële verdeling heeft met parameter $n\theta$.
Hint: bereken $\mathrm{P}\big(Y_{(1)} > y\big)$ *voor* $y > 0$.

(b) Laat zien dat $T = nY_{(1)}$ een zuivere schatter voor $1/\theta$ is.

(c) Bereken de MSE (Mean Squared Error) van T voor het schatten van $1/\theta$.
N.B. er geldt $\mathrm{Var}(Y_2) = 1/\theta^2$.

7. Het CvB van de universiteit probeert zijn medewerkers te ontmoedigen met de auto te komen. Daartoe wordt beweerd dat het gemiddeld 30 minuten duurt om naar het werk te komen (gemeten vanaf de afzit Delft-Zuid van de A13). Een medewerker denkt dat het niet zo lang duurt. Bij de laatste 5 keer dat hij naar de campus reed, heeft hij de tijd gemeten die het duurt om op zijn werk te komen. Het gemiddelde en standaard-deviatie van deze metingen zijn 20 en 6 respectievelijk. Schrijf dit probleem als een toetsingsprobleem en toets met significantie-niveau $\alpha = 0.1$ of de medewerker gelijk heeft.

Je mag aannemen dat de tijd die het duurt om een parkeerplek te vinden normaal verdeeld is.

8. Een steekproef uit een Poissonverdeling met parameter θ levert zes keer een 0 en drie keer een 1 op.

(a) Bereken de maximum-likelihood (meest-aannemelijke) schatting voor θ.

(b) Stel dat we nog een extra trekking uit dezelfde Poissonverdeling verkrijgen waarvan alleen bekend is dat de uitkomst groter dan of gelijk aan 2 is. Geef de likelihood-functie als je ook deze extra informatie tot je beschikking hebt.

N.B. Als X een Poisson verdeling heeft met parameter θ, dan geldt

$$\mathrm{P}(X = k) = e^{-\theta}\frac{\theta^k}{k!}, \qquad k = 0, 1, \ldots.$$

Puntenverdeling bij vragen:

Opgave:	1a	1b	2a	2b	3	4a	4b	5a	5b	5c	6a	6b	6c	7	8a	8b
Punten:	2	2	2	2	3	2	2	2	2	2	2	2	2	3	3	1

Antwoorden:

1 (a) Let D denote a person who is using drugs , pos denote a positive test, and neg denote a negative test. Then: $P(D|pos) = \frac{P(pos|D)P(D)}{P(pos|D)P(D)+P(pos|D')P(D')} = \frac{0.98\cdot 0.01}{0.98\cdot 0.01+0.03\cdot 0.99} = 0.2481$

(b) $P(D|neg) = \frac{P(neg|D)P(D)}{P(neg|D)P(D)+P(neg|D')P(D')} = \frac{0.02\cdot 0.01}{0.02\cdot 0.01+0.97\cdot 0.99} = 0.0002$

2 (a) Yes. Suppose the average with the true value is $\bar{X}$. Then the sum of all incomes is $1000\cdot\bar{X}$. With the mistaken value, the sum of all incomes is $1000\cdot\bar{X}+(986000-98600)$, or the computed average will go up by $\frac{986000-98600}{1000} = 887.4(\$)$

(b) No. The identity of the middle household does not change.

3 Let $\bar{X_n}$ denote the average test score based on n students. From CLT we know that if the number of students is sufficiently large $\bar{X_n} \sim N(7.5, 0.64/n)$. We need to find n s.t. $P(\bar{X_n} \geq 73) \geq 0.95$.
Then

$$0.95 \leq P(\bar{X_n} \geq 7.3) = P\left(\frac{\bar{X_n}-7.5}{\frac{.8}{\sqrt{n}}} \geq \frac{7.3-7.5}{\frac{.8}{\sqrt{n}}}\right) \approx 1-\Phi\left(\frac{-\sqrt{n}}{4}\right).$$

The above inequality is true only when $\Phi\left(\frac{-\sqrt{n}}{4}\right) \leq Z_{0.95} = -1.645$. Then $n \geq 16(-1.645)^2 = 43.29$. Rounding up we see that if at least 44 students take the exam the conditions will be satisfied.

4 a.

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2}x^2 & 0 \leq x < 1 \\ \frac{1}{2}x & 1 \leq x \leq 2 \\ 1 & x > 2 \end{cases}$$

b. Als $u \in (0, 1/2)$, dan $\frac{1}{2}x^2 = u$ en dus $x = \sqrt{2u}$. Als $u \in [1/2, 1)$, dan $\frac{1}{2}x = u$ en dus $x = 2u$.

5 a. In feite kunnen we kijken naar het volgende probleem. In een vaas zitten 8 groene en 12 rode ballen. Wat is de kans dat we precies 5 groene ballen pakken? Deze wordt gegeven door

$$p = \frac{\binom{8}{5}\binom{12}{3}}{\binom{20}{8}} \approx 0.0978.$$

b. De kansverdeling van de winst W is als volgt

w	4	1	-1
P($W = w$)	p/4	3p/4	1-p

De verwachte winst is dus

$$\mathrm{E}[W] = p + \tfrac{3}{4}p - (1-p) = 2\tfrac{3}{4}p - 1 \approx -0.731.$$

Indien $p = 0.15$, dan is het antwoord dus -0.586.
c. De gevraagde kans is $\frac{1}{4\binom{20}{8}} \approx 7.94 \cdot 10^{-6}\frac{1}{4} \approx 1.98 \cdot 10^{-6}$.

6 a) Let us find the distribution function of Y:

$$F(y) = \begin{cases} 1 - e^{-\theta y} & y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Nu volgt voor $y > 0$

$$\mathrm{P}\big(Y_{(1)} \geq y\big) = \mathrm{P}(Y_1 \geq y)^n = \left(e^{-\theta y}\right)^n = e^{-\theta n y}.$$

Dus

$$\mathrm{P}\big(Y_{(1)} \leq y\big) = 1 - e^{-\theta n y}.$$

wat de verdelingsfunctie van de Exp. verdeling met parameter $n\theta$ is.

(b) Then $\mathrm{E}[T] = nE\left(Y_{(1)}\right) = n\frac{1}{n\theta} = 1/\theta$, so T is an unbiased estimator for $1/\theta$

(c) $MSE(T; 1/\theta) = var(\theta) = n^2 var(Y_{(1)}) = n^2\frac{1}{\theta^2 n^2} = 1/\theta^2$.

7 Set the null and alternative hypotheses:

$$H_0 : \mu \geq 30$$
$$H_a : \mu < 30$$

Calculate the test statistic $T = \frac{\bar{x}-\mu_0}{\frac{s_n}{\sqrt{n}}} = \sqrt{5}\frac{20-30}{6} = -3.727$. The critical region is given by $(-\infty, t_{5,0.9}] = (-\infty, -1.53]$. Hence, we can conclude that our test statistic is in the critical region, therefore we reject H_0 in favor of H_a.

8 (a) Nu geldt

$$L(\theta) = \left(e^{-\theta}\right)^6 \left(\theta e^{-\theta}\right)^3 = \theta^3 e^{-9\theta}.$$

Dus de loglikelihood is gelijk aan

$$\ell(\theta) = 3\log\theta - 9\theta.$$

Afgeleide gelijk stellen aan nul geeft

$$\ell'(\theta) = 3/\theta - 9.$$

Na checken van het tekenoverzicht volgt dat $\hat{\theta} = 1/3$.

(b) Er geldt $\mathrm{P}(X \geq 2) = 1 - \mathrm{P}(X = 0) - \mathrm{P}(X = 1) = 1 - e^{-\theta} - \theta e^{-\theta}$ en dus volgt met de extra meting dat

$$L(\theta) = \theta^3 e^{-9\theta}(1 - e^{-\theta} - \theta e^{-\theta}).$$