Uitwerking voorbeeld toets Statistiek INF-BIT

Opgave 1

a. 25% van 25 is 6.25, dus $Q_1 = x_{(7)} = 11$ en $Q_3 = x_{(19)} = 33$, dus KA = 33 – 11 = 22.
$(Q_1 - 1.5 \times KA,\ Q_3 + 1.5 \times KA) = (-22, 66)$, dus 1 (potentiële) uitschieter: 84

b. De mediaan = 18 (= $x_{(13)}$) en $\overline{x} = 26.16$: het steekproefgemiddelde is groter dan de mediaan, hetgeen wijst op een naar rechts scheve verdeling en/of uitschieters aan de rechterkant.

c. - Uit a en b blijkt dat de verdeling scheef naar rechts is en een uitschieter aan de rechterkant.
- De numerieke waarden van de scheefheidcoëfficiënt 1.30 (> 0, dus scheefheid naar rechts) en de kurtosis 4.38 wijken af van de referentiewaarden 2 resp. 9 van de exponentiële verdeling, maar ook van de referentiewaarden van de normale verdeling (0 resp. 3).
Vanwege de scheefheid naar rechts lijkt de exponentiële verdeling de beste keuze.

d. Ja, het normale QQ-plot vertoont een duidelijk patroon (middenstuk boven de lijn *y* = *x* en de rest eronder), terwijl het exponentiële QQ-plot de lijn beter volgt.

e. 1. De coëfficiënt $a_3 = -0.2543$ uit de Shapiro-Wilk tabel met $n = 25$.
2. het kritieke gebied voor $\alpha = 10\%$ is: $W \leq c = 0.931$
3. $W = 0.870 < c$, dus H_0 verwerpen: de verdeling van het aantal dagen is niet normaal met een onbetrouwbaarheid van 5%.

f. Het interval $\overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$ is niet bruikbaar vanwege het niet normaal zijn van de verdeling. Ook kunnen wegens de lage waarde van n geen gebruik maken van een met de standaardnormale verdeling benaderd betrouwbaarheidsinterval $\overline{X} \pm c \frac{S}{\sqrt{n}}$ met c uit de N(0, 1)-verdeling.

Opgave 2

a. De zuivere schatter van p is $\frac{X}{n}$: de verwachte kwadratische fout is $var\left(\frac{X}{n}\right) = \frac{p(1-p)}{n}$ $(n = 400)$

b. We gebruiken 95%-BI voor *p* met grenzen: $\hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (zie formuleblad)

Hierin is: $c = 1.96$ ($N(0,1)$ -verdeling), $\hat{p} = \frac{73}{400} = 0.1825$, $n = 100$, invullen:

$$\left(0.1825 - 1.96\frac{\sqrt{0.1825 \times 0.8175}}{\sqrt{400}}, 0.22 + 1.96\frac{\sqrt{0.1825 \times 0.8175}}{\sqrt{400}}\right)$$

$$= (0.1825 - 0.0379, 0.1825 + 0.0379) = (0.145, 0.220)$$

Interpretatie: "met een betrouwbaarheid van 95% ligt het percentage van alle auto's in de regio die niet voldoen aan de eisen tussen 14.5 en 22%"

c. 1. $\alpha = P(X \geq 55|H_0) \overset{\text{c.c.}}{=} P(X \geq 54.5|H_0) \approx P\left(Z \geq \frac{54.5 - 400 \cdot 0.1}{\sqrt{400 \cdot 0.1 \cdot 0.9}}\right) = 1 - \Phi(2.42) = 0.78\%$

2. Het onderscheidend vermogen is voor $p = 0.15$:

$P(X \geq 55|p = 0.15) \overset{\text{c.c.}}{=} P(X \geq 54.5|p = 0.15) = P\left(Z \geq \frac{54.5 - 400 \cdot 0.15}{\sqrt{400 \cdot 0.15 \cdot 0.85}}\right) = \Phi(0.92) = 82.12\%$

(of via de kans op een fout van de 2de soort: $P(X < 55|p = 0.15) = \cdots = 17.88\%$)

Opgave 3

a. De toetsingsprocedure toegepast: zijn de legeskosten in steden hoger dan in plattelandsgemeenten?

1. Model: het gaat hier om twee onafhankelijke steekproeven, 1 uit de grote steden en 1 uit de plattelandsgemeenten, waarbij de (kwantitatieve) variabele legeskosten wordt gemeten.
   We veronderstellen voor beide populaties normale verdelingen met onbekende μ's (μ_1 resp. μ_2) en gelijke, maar onbekende σ's.
2. We toetsen H_0: $\mu_1 = \mu_2$ tegen H_1: $\mu_1 > \mu_2$ met α = 5%
3. Toetsingsgrootheid $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2\left(\frac{1}{8}+\frac{1}{8}\right)}}$ met $S^2 = \frac{7S_1^2 + 7S_2^2}{8+8-2}$
4. T is onder H_0 t-verdeeld met $df = n_1 + n_2 - 2 = 14$
5. Waargenomen: $S^2 = \frac{80^2 + 58^2}{2} = 4882$ $(s \approx 70)$, dus $t = \frac{492-452}{\sqrt{4882\left(\frac{1}{8}+\frac{1}{8}\right)}} \approx 1.145$
6. De toets is rechtseenzijdig: verwerp H_0 als $T \geq c$
   met $c = 1.761$ uit de t_{14}-tabel zodat $P(T_{14} \geq c) = 5\%$
7. $t = 1.145$ ligt niet in het kritieke gebied, dus H_0 niet verwerpen.
8. Conclusie op 5%-significantieniveau: in grote steden verschillen legeskosten voor verbouwingen niet aantoonbaar van die van plattelandsgemeenten.

b. W berekenen eerst Wilcoxon`s W door de $8 + 8$ waarnemingen te ordenen en de rangsom $W = \sum_{i=1}^{8} R(X_i)$ = 77, zie tabel.

geord.	steekpr.	$R(X_{(i)})$
381	2	
397	1	2
402	2	
410	1	4
410	2	
428	1	6
450	2	
457	2	
458	2	
500	1	10
501	2	
511	1	12
519	1	13
528	1	14
560	2	
640	1	16
	W =	77

1. De twee o.o. aselecte steekproeven zijn getrokken uit onbekende verdelingen
2. We toetsen We toetsen H_0: $f_X(x) = f_Y(x)$ tegen H_1: $f_X(x) = f_Y(x + a)$, $a > 0$ met $\alpha = 1\%$
3. $W = \sum_{i=1}^{8} R(X_i)$
4. W is onder H_0 bij benadering normaal verdeeld met
   $\mu_W = 8 \cdot \frac{1+16}{2} = 68$ en $\sigma_W = \sqrt{\frac{1}{12} n_1 n_2 (N+1)} = \sqrt{\frac{1}{12} \cdot 8 \cdot 8 \cdot 17} \approx 9.52$
5. Waargenomen $W = 77$
6. Rechtszijdige toets: als de p-waarde ≤ 5%, verwerpen we H_0.
   p-waarde $= P(W \geq 77 | H_0) \overset{\text{c.c.}}{\approx} 1 - \Phi\left(\frac{76.5 - 68}{9.52}\right) \approx 1 - \Phi(0.89) = 18.94\%$
7. De overschrijdingskans is groter dan α = 5%, dus H_0 niet verwerpen.
8. Bij een onbetrouwbaarheid van 5% zijn de leges in steden niet structureel hoger dan in de plattelandsgemeenten.

Opgave 4

a.

1. De aantallen waarnemingen in de 6 klassen N_{11}, N_{12}, N_{21}, N_{22}, N_{31} en N_{32} (N_{21} is bijv. het aantal positief gestemden onder de PvdA-kiezers) zijn multinomiaal verdeeld met de bijbehorende kansen p_{11}, p_{12}, p_{21}, p_{22} p_{31} en p_{32} (*Of kortweg: de aantallen N_{ij} zijn multinomiaal verdeeld met kansen p_{ij}.)*
2. We toetsen H_0: $p_{i\cdot} \times p_{\cdot j} = p_{ij}$ (de variabelen politieke voorkeur en mening zijn onafhankelijk) tegen H_1: $p_{i\cdot} \times p_{\cdot j} \neq p_{ij}$ voor een paar (i, j) (politieke voorkeur en mening zijn afhankelijk) met $\alpha = 0.01$
3. Toetsingsgrootheid is $\chi^2 = \sum\sum \frac{\left(N_{ij} - \hat{E}_0 N_{ij}\right)^2}{\hat{E}_0 N_{ij}}$ met schattingen $\hat{E}_0 N_{ij} = \frac{\text{kolomsom} \times \text{rijsom}}{n}$
4. Onder H_0 heeft χ^2 heeft een Chi kwadraat verdeling, aantal vrijheidsgraden $df = (r - 1)(c - 1) = 2$
5. We berekenen eerst de verwachte aantallen bij onafhankelijkheid: $\hat{E}_0 N_{ij} = E_{ij}$
   $E_{11} = \frac{300 \times 220}{800} = 82.5$ dus E_{12} = 220 - 82.5 = 137.5, $E_{21} = \frac{190 \times 300}{800} = 71.25$,$E_{22}$= 190 - 71.25 = 118.75,
   $E_{31} = 300 - 82.5 - 71.25 = 146.25$ en $E_{32} = 390 - 146.25 = 243.75$
   Uitkomst:
   $\chi^2 = \frac{(90-82.5)^2}{82.5} + \frac{(130-137.5)^2}{137.5} + \frac{(90-71.25)^2}{71.25} + \frac{(190-118.75)^2}{118.75} + \frac{(120-146.25)^2}{146.25} + \frac{(270-243.75)^2}{243.75} = 15.70$

6. We verwerpen H_0 als $\chi^2 \geq c$. In de χ^2-tabel met *df* = 2 vinden we $c \approx 9.21$
7. De uitkomst 15.70 ligt in het kritiek gebied (> 9.21), dus H_0 verwerpen.
8. Bij significantieniveau 1% is een verband tussen de mening over het kabinet en de politieke voorkeur aangetoond.

b. We passen $\hat{p}_1 - \hat{p}_2 \pm c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ met $\Phi(c) = 1 - \frac{1}{2}\alpha$ toe.

Hierin is $n_1 = 220$, $\hat{p}_1 = \frac{90}{220} \approx 40.9\%$, $n_2 = 190$, $\hat{p}_2 = \frac{90}{190} = 47.4\%$ en $c = 1.96$ met $\Phi(c) = 0.975$

95%-BI$(p_1 - p_2)$ = $(-0.0646 - 0.0962, -0.0646 + 0.0962) \approx (-0.161, 0.032)$