**Uitwerking Tentamen Wiskundige Statistiek voor TW – module 5 (201400218)** Vrijdag 28 oktober 2016

**Opgave 1**

**a.** Twee onafhankelijke binomiale aantallen: $X$= "aantal keer dat dichtslibben voorkomt in groep A" $\sim B(450, p_1)$ en $Y \sim B(400, p_2)$ voor groep B, dus gebruik interval met grenzen (formuleblad):

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

Hierin is $n_1 = 450, n_2 = 400,\ \hat{p}_1 = \frac{70}{450}$, $\hat{p}_2 = \frac{45}{400}$ en $c = 1.96$ zodat $\Phi(c) = 0.975$, dus:

95%-BI-($p_1$- $p_2$) $= \left(\left(\frac{70}{450} - \frac{45}{400}\right) - 1.96 \times 0.02327, \left(\frac{70}{450} - \frac{45}{400}\right) + 1.96 \times 0.02327\right) \approx (-0.3\%, 8.9\%)$

**b.** Het verschil in dichtslib-kansen ($p_1 - p_2$) ligt met betrouwbaarheid 95% tussen -0.3% en 8.9%. Herhaalde bepaling van zo'n interval zal in zo'n 95% van de gevallen leiden tot een interval dat de werkelijke waarde van het verschil $p_1 - p_2$ bevat.
(*Foute interpretaties: "het verschil van kansen ligt met kans 95% in dit interval" of "bij 100 herhalingen van de steekproeven zal het verschil $\hat{p}_1 - \hat{p}_2$ dan wel het verschil $p_1 - p_2$ in zo'n 95 gevallen in dit interval liggen")*

**c.** De gegeven voorwaarde betekent: de lengte van het interval $2c\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq 0.01$

Als we de schattingen voor $p_1$ en $p_2$ uit a. kiezen, $n_1 = n_2 = n$ en $c = 1.96$ nemen, vinden we:

$2 \times 1.96\sqrt{\frac{0.1556 \times 0.8444}{n} + \frac{0.1125 \times 0.8875}{n}} \leq 0.01$, ofwel $\sqrt{n} \geq \frac{3.92}{0.01}\sqrt{0.2312} \approx 188.5$ of $n \geq 35518$.

(*een alternatief is om $p(1-p) = \frac{1}{4}$ voor zowel als $\hat{p}_2$ te gebruiken: $n \geq \left(\frac{3.92}{0.01}\right)^2 \times \frac{1}{2} = 76832$.*
*Verder is het aantal benodigde patiënten voor beide steekproeven $2n$*)

**Opgave 2**

**a.** 1. Model: de waargenomen *airflow rates* zijn een realisatie van een aselecte steekproef $X_1, \ldots, X_{19}$ uit een normale verdeling met onbekende verwachte *airflow rate* μ en onbekende $\sigma^2$.
(*Dus: we passen de 1 steekproef t-toets toe*)
2. Toets $H_0$: $\mu = 0.8$ tegen $H_1$: $\mu < 0.8$ met α = 0.05
3. $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\overline{X} - 0.8}{S/\sqrt{19}}$
4. $T$ is $t_{19-1}$-verdeeld als $H_0$ waar is.
5. Waargenomen $t = \frac{0.7663 - 0.8}{0.08591/\sqrt{19}} \approx -1.71$
6. Linkseenzijdige toets: als $T \leq c$, dan $H_0$ verwerpen.
Voor α = 0.05 is $c = -1.734$ (*gebruik de $t_{18}$-tabel, zodat $P(T_{18} \geq -c) = 0.05$*)
7. -1.71 = $t > c$ = - 1.734, dus $H_0$ niet verwerpen.
8. Met een onbetrouwbaarheid 5% is de verwachte *airflow rate* niet aantoonbaar lager dan 0.8.

*Alternatief met behulp van de p-waarde (stap 6 en 7):*
De overschrijdingskans is $P(T_{18} \leq -1.734) = P(T_{18} \geq 1.734)$ ligt volgens de $t_{18}$-tabel tussen 5 en 10%, is dus groter dan 5% = α, dus $H_0$ niet verwerpen.

**b.** De p-waarde 13.2% betekent dat de aanname van normaliteit niet verworpen wordt bij een gebruikelijke keuze van α, dus tussen 1% en 10%. (*De toets in a. is dus gebaseerd op een redelijke veronderstelling*)

**Opgave 3**

**a.** $F_U(u) = P(U \leq u) = 1 - P(\min(X_1, \ldots, X_n) > \text{u}) = 1 - P(X_1 > u) \times \ldots \times P(X_n > u)$
$= 1 - e^{-\lambda u} \times \ldots \times e^{-\lambda u} = 1 - e^{-n\lambda u}$ $(u \geq 0)$ , omdat de $X_i$'s o.o. en $Exp(\lambda)$-verdeeld zijn.
$f_U(u) = \frac{d}{du}F_U(u) = +n\lambda e^{-n\lambda u}$ $(u \geq 0)$, dus $U$ is $Exp(n\lambda)$-verdeeld.

**b.** $\overline{X}$ is een zuivere schatter van $\mu$, omdat $E(\overline{X}) = \mu$ (*bekende eigenschap van $\overline{X}$*)
$T = nU$ is ook een zuivere schatter want $E(T) = nE(U) = n \times \frac{1}{n\lambda} = \frac{1}{\lambda} = \mu$ (omdat $U \sim Exp(n\lambda)$, zie a.)

**c.** Met de eigenschap $MSE(T) = (ET - \theta)^2 + var(T)$ en de aangetoonde zuiverheid in b. vinden we:
$MSE(\overline{X}) = var(\overline{X}) = \frac{\sigma^2}{n} = \frac{1}{n\lambda^2}$ en
$MSE(T) = var(T) = var(nU) = n^2 var(U) = n^2 \times \frac{1}{(n\lambda)^2} = \frac{1}{\lambda^2}$.
Dus voor $n = 1$ zijn beide schatters even goed (identiek) en voor $n \geq 2$ is $\overline{X}$ beter.

**d.** We weten dat een schatter consistent is als $\lim_{n\to\infty} MSE = 0$.
$\lim_{n\to\infty} MSE(\overline{X}) = \lim_{n\to\infty} \frac{1}{n\lambda^2} = 0$ , dus $\overline{X}$ is een consistente schatter van $\mu$.
$\lim_{n\to\infty} MSE(T) = \lim_{n\to\infty} \frac{1}{\lambda^2} > 0$, maar daar volgt niet automatisch uit dat $T$ geen consistente schatter van $\mu$ is (*De eigenschap zegt dat de schatter consistent is als* $\lim_{n\to\infty} MSE = 0$, *niet andersom*)
Als we de definitie van consistentie toepassen vinden we ( $\epsilon > 0$):

$$\begin{aligned} P(|T-\mu| > \epsilon) &= P(T < \mu - \epsilon) + P(T > \mu + \epsilon) = P(nU < \mu - \epsilon) + P(nU > \mu + \epsilon) \\ &= P\left(U < \frac{\mu-\epsilon}{n}\right) + P\left(U > \frac{\mu+\epsilon}{n}\right) \\ &= P\left(U < \frac{\mu-\epsilon}{n}\right) + P\left(U > \frac{\mu+\epsilon}{n}\right) \\ &= 1 - e^{-n\lambda \cdot \frac{\mu-\epsilon}{n}} + e^{-n\lambda \cdot \frac{\mu+\epsilon}{n}} = 1 - e^{-\lambda(\mu-\epsilon)} + e^{-\lambda(\mu+\epsilon)} \end{aligned}$$

convergeert dus niet naar 0, dus $T$ is <u>niet</u> consistent.

**Opgave 4**

**a.** $L = \prod_{i=1}^{n} P(X = x_i) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p = (1-p)^{\sum x_i - n} p^n$
Dus $\ln(L) = (\sum x_i - n)\ln(1-p) + n\ln(p)$
$\frac{d}{dp}\ln(L) = 0$ als $-\frac{\sum x_i - n}{1-p} + \frac{n}{p} = 0$, ofwel $\frac{-(\sum x_i - n)p + n(1-p)}{p(1-p)} = \frac{n - p\sum x_i}{p(1-p)} = 0$ als $p = \frac{n}{\sum x_i} = {}^{1}\!/_{\overline{x}}$.
Dit is een maximum voor $\ln(L)$, omdat $\frac{d}{dp}\ln(L) > 0$ voor $p < {}^{1}\!/_{\overline{x}}$ en $\frac{d}{dp}\ln(L) < 0$ voor $p > {}^{1}\!/_{\overline{x}}$ .
Dus $\hat{p} = {}^{1}\!/_{\overline{X}}$ is de meest aannemelijke ($mle$) van $p$.

**b.** De MP-toets is de toets die $H_0: p = 0.1$ verwerpt ten gunste van $H_1: p = 0.2$ voor grote waarden van

$$r(x_1, \dots, x_n) = \frac{\prod_{i=1}^{n} P(X = x_i | p = 0.2)}{\prod_{i=1}^{n} P(X = x_i | p = 0.1)} = \frac{0.8^{\sum x_i - n} 0.2^n}{0.9^{\sum x_i - n} 0.1^n} = 2^n \left(\frac{0.8}{0.9}\right)^{n(\overline{x}-1)}$$

$r$ is een dalende functie in $\overline{x}$, dus de MP-toets verwerpt voor **kleine** waarden van $\overline{x}$:
**verwerp $H_0$ als $\overline{X} \leq c$.**
Als we $H_0: p = 0.1$ toetsen tegen $H_1: p = p_1$, voor willekeurige $p_1 > 0.1$ blijft de conclusie hetzelfde want dan is $r(x_1, \dots, x_n) = \frac{\prod_{i=1}^{n} P(X = x_i | p = p_1)}{\prod_{i=1}^{n} P(X = x_i | p = 0.1)} = \frac{(1-p_1)^{\sum x_i - n} p_1^n}{0.9^{\sum x_i - n} 0.1^n} = \left(\frac{p_1}{0.1}\right)^n \left(\frac{1-p_1}{0.9}\right)^{n(\overline{x}-1)}$ is nog steeds een dalende functie in $\overline{x}$, want $\frac{1-p_1}{0.9} < 1$, omdat $p_1 > 0.1$.
Conclusie: de toets die **$H_0$ verwerpt als $\overline{X} \leq c$ is Uniform Most Powerful (UMP).**

**Opgave 5**

Chebyshev: $P(|T-\theta| > \varepsilon) \leq \frac{E(T-\theta)^2}{\varepsilon^2}$, waarin $E(T-\theta)^2 = MSE(T)$ als T een schatter van $\theta$ is.
Bewijs: $Z$ heeft de volgende verdeling: $Z = \varepsilon^2$, als $(T-\theta)^2 > \varepsilon^2$, en dus met kans $P(|T-\theta| > \varepsilon)$ en
$Z = 0$, als $(T-\theta)^2 \leq \varepsilon^2$, en dus met kans $P(|T-\theta| \leq \varepsilon)$.
Merk op dat $T \leq (T-\theta)^2$ en dus ook $E(T) \leq E(T-\theta)^2$, dus:
$E(Z) = \varepsilon^2 \times P(|T-\theta| > \varepsilon) + 0 \times P(|T-\theta| \leq \varepsilon) = \boldsymbol{\varepsilon^2 \times P(|T-\theta| > \varepsilon) \leq E(T-\theta)^2}$, waaruit Chebyshevs regel volgt.

(*Je kunt ook eerst bewijzen dat* $P(|X| > \varepsilon) \leq \frac{E(X^2)}{\varepsilon^2}$ *en vervolgens* $X = T - \theta$ *substitueren, want:*

$$\begin{aligned} E(X^2) = \sum_x x^2 P(X = x) &= \sum_{|x| \leq \varepsilon} x^2 P(X = x) + \sum_{|x| > \varepsilon} x^2 P(X = x) \\ &\geq \sum_{|x| > \varepsilon} x^2 P(X = x) \geq \sum_{|x| > \varepsilon} \varepsilon^2 P(X = x) = \varepsilon^2 \sum_{|x| > \varepsilon} P(X = x) = \varepsilon^2 P(|X| > \varepsilon) \end{aligned}$$

*voor discrete* $X$*: continue geval is volkomen analoog*)

**Opgave 6**

**a.** We gebruiken de vector-notatie van het multiple lineaire regressie model gebruikend: $Y = X\beta + \varepsilon$:
$E(\hat{\beta}) = \beta$ en $var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$, dus $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$

**b.** De elementen van $\varepsilon$ zijn o.o. en alle $N(0, \sigma^2)$, dus $Y \sim N(X\beta, \sigma^2 I)$ (multivariate normale verdeling)
Maar dan is ook $\hat{\beta} = AY = (X^TX)^{-1}X^T\,Y$ multivariaat normaal verdeeld met
verwachting $E(\hat{\beta}) = E(AY) = AE(Y) = (X^TX)^{-1}X^T\,(X\beta) = (X^TX)^{-1}(X^TX)\beta = \beta$
en variantie
$var(AY) = A^T var(Y)A = (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1} = \sigma^2 I(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2 I(X^TX)^{-1}$
Dus $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$.

**Opgave 7**

**a.** Er is sprake van 2 onafhankelijke aselecte steekroeven dus we passen Wilcoxons rangsomtoets toe:
Noem $X_1, \ldots, X_{14}$ de PCB-metingen op het platteland en $Y_1, \ldots, Y_{15}$ die van de steden.
1. $X_1, \ldots, X_{14}$ en $Y_1, \ldots, Y_{15}$ zijn o.o.: de $X_i$'s zijn verdeeld volgens de onbekende kansdichtheid $f_X$ en de $Y_j$'s volgens de onbekende kansdichtheid $f_Y$.
2. Toets $H_0$: $f_X(x) = f_Y(x)$ tegen $H_1$: $f_X(x) = f_Y(x - a)$ voor een zekere $a \neq 0$.
3. $W = \sum_{i=1}^{14} R(X_i)$
4. Onder $H_0$ heeft $W$ bij benadering een normale verdeling met $\mu = \frac{1}{2}n_1(N+1) = \frac{1}{2} \cdot 14 \cdot 30 = 210$
en $\sigma^2 = \frac{1}{12}n_1n_2(N+1) = \frac{1}{12} \cdot 14 \cdot 15 \cdot 30 = 525$
5. Uitkomst van $W = 1 + 2 + 3 + 4 + 5 + 6 + 8 + 9 + 10 + 11 + 15 + 17 + 23 = 121$
6. Verwerp $H_0$ als de (tweezijdige) overschrijdingskans $\leq \alpha_0 = 5\%$ is.
Overschrijdingskans = $2 \cdot P(W \leq 121|H_0) \approx 2P\left(Z \leq \frac{121.5-210}{\sqrt{525}}\right) \approx 2\big(1 - \Phi(3.86)\big) < 0.0002$
7. De overschrijdingskans is kleiner dan $\alpha_0 = 5\%$, dus $H_0$ verwerpen.
8. We achten dus wel bewezen dat er een verschil is in PCB gehaltes tussen stad en platteland, met een onbetrouwbaarheid van 5%.

Alternatief met kritiek gebied (twee staartkansen van 2.5% en c.c. gebruiken):
Dan vind je als criterium: verwerp $H_0$ als $W \leq c_1 = 164$ of als $W \geq c_2 = 256$ (zelfde conclusie).

**b.** Opvallend bij het beoordelen van de box plots:
1. de verdelingen zijn sterk scheef: een parametrische toets is niet op zijn plaats (maar geen probleem voor het toepassen van verdelingsvrije methoden als Wilcoxon's rangsomtoets).
2. De verdelingen lijken niet "verschoven" ten opzichte van elkaar: de spreiding verschilt ook sterk. Dit roept de vraag op of je wel gelijke verdelingen tegenover het verschuivingsalternatief (zie hypothesen) mag toepassen.

Echter, de log-transformatie laat wel vergelijkbare verdelingen, op verschuiving na: omdat de log-transformatie de rangnummers niet verandert, lijkt toepassen van Wilcoxon`s rangsomtoets hier toch gerechtvaardigd.
(*Opmerking: wellicht is het ook mogelijk normale verdelingen te veronderstellen voor de getransformeerde waarnemingen, waarop we vervolgens een t-toets kunnen toepassen.)*