Antwoorden 2e deeltentamen Statistiek
27 juni 2012

Opgave 1 Stel dat er een onderzoek is gedaan naar de associatie van 20 genen en een bepaalde ziekte. De volgende p-waarden voor de associaties zijn gevonden: $p_i = \frac{1}{2}e^{-\frac{1}{2}(i-1)}$ voor $i \in \{1, \ldots, 20\}$. Welke genen beoordeelt u als statistisch significant geassocieerd als u zoveel mogelijk statistisch significante associaties wilt vinden en

a 10 pt) de kans dat u minstens 1 gen ten onrechte als geassocieerd benoemt kleiner moet zijn dan $0,05$.
ANTWOORD: Gebruik de Bonferroni correctie: Het aantal p-waardes kleiner dan 0,05/20 is 9. Dit volgt uit: $0,05/20 = p_i \Rightarrow 0,0025 = \frac{1}{2}e^{-\frac{1}{2}(i-1)} \Rightarrow i = 1 - 2\log(0,005) \approx 11,6$. Dus voor $i \in \{12, \ldots, 20\}$ geldt $p_i < 0,05/20$.

b 10 pt) u accepteert dat het verwachte aantal genen dat ten onrechte als significant geassocieerd is gekwalificeerd maximaal 5% is van het aantal genen dat als significant geassocieerd is gekwalificeerd, d.w.z., maximaal 5% van de positieve resultaten is fout-positief.
ANTWOORD: Gebruik de methode van Benjamini en Hochberg. Zij k de grootste i zodanig dat $p_{(i)} = p_{21-i} \leq \frac{i}{20}0,05$. Uitproberen geeft dat $k = 14$. De 14 kleinste p-waardes worden dus als significant bestempeld.

Opgave 2 In een experiment is de bloeddruk gemeten van 10 proefpersonen waarvan de body mass index (BMI) bekend is. De onderzoeksvraag is of er een relatie is tussen bloeddruk en BMI. De resultaten van het experiment staan in de tabel hieronder:

individu	BMI	bloedruk (in mmHg)
i	x_i	y_i
1	18	110
2	18	114
3	20	115
4	21	121
5	22	117
6	24	112
7	24	120
8	25	119
9	26	117
10	30	121

7 pt) Vind de regressielijn voor de bloeddruk als functie van de BMI.
ANTWOORD: $\bar{x} = 22,8$, $\bar{y} = 116,6$.
$\widehat{\beta_1} = \frac{\sum_{i=1}^{10}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{10}(x_i-\bar{x})^2} = \frac{193}{319} \approx 0,61$, $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = \frac{32795}{319} \approx 102,81$. De regressielijn is dus: $\hat{y} = \frac{32795}{319} + \frac{193}{319}x$.

9 pt) Bepaal de 95% betrouwbaarheidsintervallen voor het snijpunt (β_0) en de helling (β_1).
ANTWOORD: $\frac{\widehat{\beta_i}-\beta_i}{s_{\widehat{\beta_i}}} \sim t_{n-2}$.
$s^2 = \frac{RSS}{n-2} = \frac{\sum_{i=1}^{10}(y_i-(\widehat{\beta_0}-\widehat{\beta_1}x_i)^2}{n-2} = \frac{13349}{1276} \approx 10,46$.
$\widehat{Var(\hat{\beta}_0)} = \frac{s^2\sum_{i=1}^{10}x_i^2}{10\sum_{i=1}^{10}x_i^2-(\sum_{i=1}^{10}x_i)^2} = \frac{s^2 5326}{10\cdot 5326-51984} \approx 43,67$

$\widehat{Var(\hat{\beta}_1)} = \frac{ns^2}{10\sum_{i=1}^{10} x_i^2 - (\sum_{i=1}^{10} x_i)^2} = \frac{10s^2}{10 \cdot 5326 - 51984} \approx 0,082.$

Hieruit volgt dat de 95% betrouwbaarheidsintervallen gegeven worden door: $\hat{\beta}_i \pm t_8(0,025)\sqrt{\widehat{Var(\hat{\beta}_i)}}$. $t_8(0,025) = 2,306$. Dus de 95% betrouwbaarheidsintervallen voor β_0 en β_1 zijn respectievelijk: (87,6-118,0) en (-0,06-1,27).

9 pt) Bepaal een 95% betrouwbaarheidsinterval voor de verwachte bloeddruk van een persoon met een BMI van $x_0 = 23$.

ANTWOORD: Voor 2 stochasten X en Y en een constante a geldt: $Var(x + aY) := E(X + aY - E(X + aY))^2 = E(X^2) + a^2E(Y) + 2aCov(X, Y)$, dus geldt: $Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = Var(\hat{\beta}_0) + x_0^2 Var(\hat{\beta}_1) + 2x_0 Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{10} x_i^2}{10\sum_{i=1}^{10} x_i^2 - (\sum_{i=1}^{10} x_i)^2} + x_0^2 \frac{n\sigma^2}{10\sum_{i=1}^{10} x_i^2 - (\sum_{i=1}^{10} x_i)^2} + 2x_0 \frac{-\sigma^2 \sum_{i=1}^{10} x_i}{10\sum_{i=1}^{10} x_i^2 - (\sum_{i=1}^{10} x_i)^2} = \sigma^2 \left(\frac{1}{10} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{10} (x_i - \bar{x})^2} \right)$. Een 95% betrouwbaarheidsinterval voor de bloeddruk is dan $\hat{\beta}_0 + x_0\hat{\beta}_1 \pm s^2 \left(\frac{1}{10} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{10} (x_i - \bar{x})^2} \right) t_8(0,025) = \hat{\beta}_0 + x_0\hat{\beta}_1 \pm s^2 \frac{32}{319} t_8(0,025) = \hat{\beta}_0 + x_0\hat{\beta}_1 \pm 1,05 t_8(0,025) = \hat{\beta}_0 + x_0\hat{\beta}_1 \pm 1,05 \cdot 2,306 = \hat{\beta}_0 + x_0\hat{\beta}_1 \pm 2,42$. Dit is gelijk aan: $(114,3 - 119,1)$.

Opgave 3 Er is een onderzoek gedaan met 40 personen met depressieve klachten. 20 personen hebben een antidepressivum gekregen, de 20 andere personen een placebo. Na een maand is de deelnemers gevraagd om op een schaal van één tot vijf aan te geven hoe de klachten zijn t.o.v. een maand eerder. De betekenis van de schaal is als volgt: de klachten zijn I: sterk verminderd, II: verminderd, III: gelijk gebleven, IV: verergerd, V: sterk verergerd. De resultaten van het experiment staan in de tabel:

Behandeling	Uitkomst				
	I	II	III	IV	V
Antidepressivum	6	5	4	3	2
Placebo	4	4	4	5	3

De onderzoeksvraag is of het antidepressivum de depressieve klachten vermindert. De onderzoekers gaan uit van een significantieniveau (α) van 0,05.

a 6 pt) Geef de nulhypothese en de alternatieve hypothese.

ANTWOORD: H_0: Het antidepressivum en het placebo hebben dezelfde invloed op de klachten na 1 maand, i.e., de distributie over de 5 categorieën is hetzelfde voor het antidepressivum en het placebo. Het betreft een eenzijdige toets: H_A: het antidepressivum leidt gemiddeld tot betere uitkomsten dan het placebo.

b 8 pt) Bereken de Mann Whitney statistiek (U). Als twee patiënten dezelfde uitkomst hebben en een patiënt het antidepressivum heeft gekregen en de ander het placebo, schrijf dan 50% van het gewicht van dat paar toe aan de placebo-arm (alsof de patiënt met het placebo een beter resultaat had dan de patiënt met het antidepressivum) en 50% van het gewicht van dat paar aan de behandeling (alsof de patiënt met de behandeling een beter resultaat heeft geboekt).

ANTWOORD: Het aantal paren patiënten waarbij het antidepressivum betere resultaten boekte dan de placebo is: $\frac{1}{2}6 \cdot 4 + 6(4 + 4 + 5 + 3) + \frac{1}{2}5 \cdot 4 + 5(4 + 5 + 3) + \frac{1}{2}4 \cdot 4 + 4(5 + 3) + \frac{1}{2}3 \cdot 5 + 3(3) + \frac{1}{2}2 \cdot 3 = \frac{475}{2} = 237,5$

c 8 pt) Gebruik de de Mann Whitney statistiek om te bepalen of het antidepressivum als werkzaam beschouwd mag worden op basis van dit experiment. (Indien u onderdeel b niet heeft berekend, neem dan aan dat $U = 238$.)

ANTWOORD: De kans op een waarde van de U-statistiek van 237,5 of hoger, gegeven de nulhypothese kan berekend worden met de normale benadering.
$\sigma = \sqrt{mn(m+n+1)/12} = \sqrt{\frac{4100}{3}} \approx 37,0$ ($m = n = 20$). $\frac{U-mn/2}{\sigma}$ is bij benadering standaardnormaal verdeeld onder de nul-hypothese. $(237,5 - 200)/37,0 \approx 1,01$. $1 - \phi(1,01) = 0,1562 > \alpha$. We kunnen op basis van dit experiment dus niet concluderen dat het antidepressivum werkzaam is. ALTERNATIEF: Gebruik tabel 8 van Rice. Voor $n_1 = n_2 = 20$, is de kritische waarde voor een 1-zijdige test met $\alpha = 0,05$ voor de Wilcoxon rank sum test gelijk aan 348. Dit is een afwijking van ((20(20+20+1)/2)-348=62 van het gemiddelde. In het voorbeeld is de afwijking 37,5, dat is minder, dus is er geen significant verschil tussen de placebo en het antidepressivum in deze test. (De alternatieve methode is gebaseerd op het feit dat de Mann-Whitney en de Wilcoxon rank sum statistiek op een constante na identiek aan elkaar zijn).

Opgave 4 Zij $C = (X, Y, Z)^T \sim N(\mu_C, \Sigma_{C,C})$ met

$\mu_C = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ en $\Sigma_{C,C} = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 4 \end{pmatrix}$. Zij $D = \begin{pmatrix} \mathrm{U} \\ \mathrm{V} \end{pmatrix}$ met $U = 2X + 3Y - Z$ en $V = X + Z$.

a 7pt) Bereken de covariantiematrix Σ_{DD}.

ANTWOORD: $D = AC$ met $A = \begin{pmatrix} 2 & 3 & -1 \\ 1 & 0 & 1 \end{pmatrix}$.

Dus $\Sigma_{DD} = A\Sigma_{CC}A^T = \begin{pmatrix} 2 & 3 & -1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 4 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 3 & 0 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 61 & -1 \\ -1 & 4 \end{pmatrix}$

b 3pt) Zijn U en V onafhankelijk?
ANTWOORD: Nee, $Cov(U, V) = -1 \neq 0$.

c 5pt) Bereken de correlatiecoëfficiënt tussen X en Y.
ANTWOORD: $\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{1}{\sqrt{2 \cdot 3}} \approx 0.41$.

d 5pt) Stel dat je van een steekproef van C van grootte 1 weet dat X de waarde 2 heeft aangenomen. Wat is de verwachte waarde voor Y?

ANTWOORD: De Variantie in X is 2, dus de standaarddeviatie is $\sqrt{2}$. Dat betekent dat $x = 2$, $1/\sqrt{2}$ sd afwijkt van het gemiddelde. Dat betekent dat Y volgens verwachting $\rho/\sqrt{2}$ sd afwijkt van het gemiddelde, dus $E(Y|X = 2) = E(Y) + 0.41\sqrt{3}/\sqrt{2} = 2 + 0.41\sqrt{3}/\sqrt{2} = 5/2$.

Opgave 5 Stel dat we n datapunten hebben van de vorm: $y_i, x_{i,1}, x_{i,2}, \ldots, x_{i,m}$ voor $i \in \{1, n\}$. Stel dat bij een multivariate lineaire regressie analyse met de kleinste kwadratenmethode we de vector $\hat{\beta} := (\beta_0, \beta_1 \ldots, \beta_m)^T$ vinden voor de coëfficiënten van de regressieanalyse. Stel dat we de verklarende variabelen $x_{i,j}$ als volgt transformeren: $u_{i,j} = c_j + a_j x_{i,j}$ ($a_j \neq 0 \quad \forall j \in \{1, m\}$) of in vector notatie: $u_i := (u_{i,1}, \ldots u_{i,m})^T = c + Ax_i$ met $c = (c_1, \ldots, c_m)^T$, $x_i = (x_{i,1}, \ldots, x_{i,m})^T$ en A een $m \times m$-diagonaalmatrix, waarbij het element A_{ii} gelijk is aan a_i.

a 10 pt) Geef formules voor de regressiecoefficienten in de getransformeerde variabelen.

ANTWOORD: $\hat{\beta^u} = \arg\min_{\beta} \left(\sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{m} \beta_j u_{i,j} \right)^2 =$

$$\arg\min_{\beta} \left(\sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{m} \beta_j (c_j + a_j x_{i,j}) \right)^2 =$$

$$\arg\min_{\beta} \left(\sum_{i=1}^{n} y_i - (\beta_0 + \sum_{j=1}^{m} \beta_j c_j) - \sum_{j=1}^{m} \beta_j (a_j x_{i,j}) \right)^2 .$$

$\hat{\beta}$ is de oplossing van: $\arg\min_{\beta} \left(\sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{i,j}) \right)^2$, dus vinden we dat $\hat{\beta}_u$ gegeven wordt door: $\hat{\beta}_0^u = \hat{\beta}_0 - \sum_{j=1}^{m} \frac{\hat{\beta}_j}{a_j} c_j$ en $\hat{\beta}_j^u = \hat{\beta}_j / a_j$ voor $j \in \{1, \ldots, m\}$. Oftewel,

$$\hat{\beta^u} = \begin{pmatrix} 1 & -\frac{c_1}{a_1} & -\frac{c_2}{a_2} & \cdots & -\frac{c_m}{a_m} \\ 0 & \frac{1}{a_1} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{a_2} & & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & 0 & \frac{1}{a_m} \end{pmatrix} \hat{\beta}$$

b 3 pt) Geef een voorbeeld waarbij bovenstaande transformatie handig is.
ANTWOORD: Stel dat de temperatuur als verklarende variabele omgezet wordt van de Kelvin schaal naar de Celsius schaal, of dat een afstand in meters i.p.v. kilometers wordt gegeven.