

Sandsynlighedsteori og Statistik: Ugeseddel 13

Levi van Boekel

Uge 48, 2024

Opgave 2

2.1

Et histogram over observationerne er tegnet nedenfor:

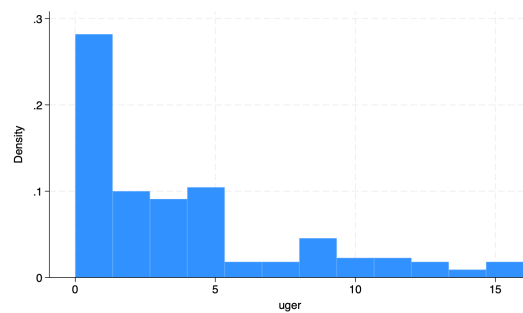


Figure 1: Histogram for uger uden fravær

Der er samtidig bestemt følgende deskriptive statistiske mål:

uger					
Percentiles		Smallest			
1%	0	0			
5%	0	0			
10%	0	0			
25%	1	0	Obs		165
			Sum of wgt.		165
50%	2		Mean		3.690909
		Largest	Std. dev.		3.976483
75%	5	15			
90%	10	15	Variance		15.81242
95%	12	16	Skewness		1.301522
99%	16	16	Kurtosis		3.912086

Figure 2: Deskriptive mål for uger uden fravær

Vi ser, at den gennemsnitlige elev møder ca. 4 uger i skole ugen fravær, samtidig med, at der er stor varians blandt eleverne. Skævheden fortæller os desuden, at fordelingen er højreskæv, svarende til, at der er mere sandsynlighedsmasse til højre for middelværdien. Vi kan desuden se, at topstejlheden er højere end for normalfordelingen (3), svarende til, at der er mere masse i halerne. Vi ser desuden, at ca. 95 pct. af eleverne har 12 eller færre uger uden fravær, mens *kun* 50 pct. af eleverne har 2 uger eller færre uger uden fravær.

2.2

Log likelihood funktionen er:

$$\log L_n(\theta \mid y_1, \dots, y_{165}) = \sum_{i=1}^{165} [\log(\theta) + y_i \cdot \log(1 - \theta)] \quad (1)$$

Vi antager her (1), at observationerne kommer fra en eksponentialfordeling, hvilket som udgangspunkt er en OK antagelse, da nogen sygdom er *memoryless*, dvs. indtræder med en given sandsynlighed uanset hvor længe det er siden man sidst fik sygdom. Der er dog også mange typer sygdomme, hvor det ikke er tilfældet og hvor man (1) bliver resistent eller (2) bliver ved med at være syg. Vi antager (2) at sandsynligheden for at blive syg er den samme for hver barn i klassens, hvilket oplagt heller ikke er sandt, idet det også afhænger af sanitære forhold i hjemmet og f.eks. kost. Afsluttende antager vi uafhængighed, hvilket heller ikke forekommer rimeligt, idet sygdom smitter. I det hele taget virker modellen altså svag.

2.3

Vi finder først Score funktionen:

$$\frac{\delta \log L_n(\theta \mid Y_i)}{\delta \theta} = \sum_{i=1}^{165} \left[\frac{1}{\theta} - \frac{Y_i}{1 - \theta} \right]$$

Og løser nu FOC:

$$\begin{aligned} \sum_{i=1}^{165} \left[\frac{1}{\theta} - \frac{Y_i}{1 - \theta} \right] &= 0 \\ \frac{n}{\theta} &= \frac{\sum_{i=1}^{165} Y_i}{1 - \theta} \\ n - n\theta &= \theta \cdot \sum_{i=1}^{165} Y_i \\ n &= \theta \cdot \left(\sum_{i=1}^{165} Y_i + n \right) \\ \hat{\theta} &= \frac{n}{\sum_{i=1}^{165} Y_i + n} \end{aligned}$$

Ovenstående er altså MLE estimatoren. For at finde estimatet indsætter vi realisationerne:

$$\hat{\theta}(y_1, \dots, y_{165}) = \frac{165}{609 + 165} \approx 0.213 \quad (2)$$

2.4

For at være sikre på, at ovenstående er maksimum tages den anden ordens afledte (hessematricen):

$$\frac{\delta \log L_n(\theta | Y_i)}{\delta \theta \delta \theta} = \left[-\frac{1}{\theta^2} - \frac{Y_i}{(1-\theta)^2} \right]$$

2.5

Informationen bestemmes nedenfor:

$$\begin{aligned} I(\theta) &= -\mathbb{E}(H(\theta)) \\ &= -\mathbb{E} \left[-\frac{1}{\theta^2} - \frac{Y_i}{(1-\theta)^2} \right] \\ &= \frac{1}{\theta^2} + \frac{\frac{(1-\theta)}{\theta}}{(1-\theta)^2} \\ &= \frac{1}{\theta^2} + \frac{1}{\theta \cdot (1-\theta)} \\ &= \frac{(1-\theta) + \theta}{\theta^2(1-\theta)} \\ &= \frac{1}{\theta^2(1-\theta)} \end{aligned}$$

2.6

Hvorfor den asymptotiske varians er:

$$\text{Var}(\hat{\theta}(Y_1, \dots, Y_{165})) = \frac{\theta^2(1-\theta)}{n} \quad (3)$$

Som i vores estimat bliver: $\frac{0.23^2 \cdot (1-0.23)}{165} \approx 0.0002468$, med standardafvigelse: $\sqrt{0.0002468} \approx 0.0157$

2.7

95 pct. konfidensintervallet er:

$$\theta_0 \in [0.23 - 1.96 \cdot 0.0157, 0.23 + 1.96 \cdot 0.0157] = [0, 199228, 0, 260772] \quad (4)$$

2.8 & 2.9

Vi bestemmer ssh. for nogle enkelte udfald:

$$\begin{cases} P(Y = 0) = (1 - 0.213)^0 \cdot 0.213 = 0.213 & \text{Faktisk ssh: 0.224} \\ P(Y = 1) = (1 - 0.213)^1 \cdot 0.213 = 0.167 & \text{Faktisk ssh: 0.1515} \\ P(Y = 2) = (1 - 0.213)^1 \cdot 0.213 = 0.1314 & \text{Faktisk ssh: 0.1333} \end{cases} \quad (5)$$

Det ser altså ud til at passe nogenlunde...

Opgave 3

3.1

Vi starter med at antage, at kønnet af børnene er eksogent givet, dvs:

$$D_i \text{ er eksogent givet.} \quad (6)$$

Vi antager desuden, at sygefraværet givet køn, stadigvæk følger en eksponentialfordeling, men hvor parameteren afhænger af d_i og y_i gennem funktionen $g_1(\cdot)$, dvs:

$$Y_i \mid D_i \sim \text{Geo}(\theta) \quad (7)$$

Hvor det i opgaven er givet, at:

$$\theta_i = \beta_1 + \beta_2 D_i \quad (8)$$

Log likelihood funktionen bliver da:

$$\log L_n(\theta \mid y_1, \dots, y_{165}) = \sum_{i=1}^{165} [\log(\beta_1 + \beta_2 d_i) + y_i \cdot \log(1 - \beta_1 - \beta_2 d_i)] \quad (9)$$

Hvor parameterrummet er: $\theta \in \Theta\{\beta_1, \beta_2 \in \mathbb{R} \mid \beta_1 + \beta_2 d_i > 0\}$. Fortolkningen af parameteren er nu, at β_1 beskriver sygefraværet givet, at man ikke er dreng, mens β_2 beskriver, hvordan sygefraværet ændres, hvis man er dreng.

3.2 & 3.3

Vi får nu:

$$\hat{\beta}_1 = 0.165$$

$$\hat{\beta}_2 = 0.119$$

For drenge får vi:

$$\begin{cases} P(Y = 0) = (1 - 0.046)^0 \cdot 0.284 = 0.284 & \text{Faktisk ssh: 0.27} \\ P(Y = 1) = (1 - 0.213)^1 \cdot 0.213 = 0.27 & \text{Faktisk ssh: 0.2045} \\ P(Y = 2) = (1 - 0.213)^2 \cdot 0.25 = 0.1477 & \text{Faktisk ssh: 13.33} \end{cases} \quad (10)$$

Som altså er mindre præcist end den ikke betingede model... For piger får vi:

$$\begin{cases} P(Y = 0) = (1 - 0.046)^0 \cdot 0.284 = 0.165 & \text{Faktisk ssh: 0.1688} \\ P(Y = 1) = (1 - 0.213)^1 \cdot 0.213 = 0.1377 & \text{Faktisk ssh: 0.909} \\ P(Y = 2) = (1 - 0.213)^2 \cdot 0.25 = 0.115 & \text{Faktisk ssh: 0.1169} \end{cases} \quad (11)$$

Den ubetingede model virker derfor bedre.

3.4

Vi formulerer null hypotesen som:

$$\mathcal{H}_0 : \hat{\beta}_2 = 0.1 \quad \Theta_0\{\beta_1, \beta_2 \in \mathbb{R} \mid \beta_1 + \beta_2 d_i > 0, \beta_2 = 0.1\} \quad (12)$$

Mens den alternative hypotese bliver:

$$\mathcal{H}_A : \hat{\beta}_2 \neq 0.1 \quad \Theta_0\{\beta_1, \beta_2 \in \mathbb{R} \mid \beta_1 + \beta_2 d_i > 0, \beta_2 \neq 0.1\} \quad (13)$$

Wald statistikken er:

$$\frac{\hat{\beta}_2 - \beta_0^2}{se(\hat{\beta}_2)} = \frac{0.119 - 0.1}{.0309243} \approx 0.3848$$

Vi ved, at ovenstående er asymptotisk standard normalfordelt, hvis null hypotesen er sand. Derfor tjekker vi nu, om det er sandsynligt, at ovenstående realisation kommer fra en standard normal-

fordeling. Ved $\alpha = 0.05$ er de kritiske værdier

$$\Phi^{-1}(0.025) \approx -1.96$$

$$\Phi^{-1}(0.975) \approx 1.96$$

Dvs. null hypotesen afvises, hvis $Z_n > 1.96$ eller $Z_n < -1.96$. Det er klart, at vi derfor ikke kan afvise null hypotesen, da $-1.96 < 0.3834 < 1.96$. Vi kan ligeledes bestemme p-værdien som sandsynligheden for at få et mere ekstremt udfald end z_n som:

$$p_n(\theta_0 = 0.3848) = 2 \cdot (1 - \Phi(0.3848)) \approx 0.7 \quad (14)$$

Der er altså ca. 70 pct. sandsynlighed for at få et mere ekstremt udfald, hvilket er strengt højere end 0.05, hvorfor vi også på den baggrund ikke kan afvise null hypotesen.

Vi laver nu en LR test, hvor vi bemærker, at det ikke restrikerede estimat log likelihood estimat er:

$$L_n(\hat{\theta}) = -393.30716 \quad (15)$$

Det restrikerede bliver:

$$L_n(\tilde{\theta}) = -393.5012 \quad (16)$$

Vi beregner nu LR statistikken som:

$$LR = 2 [-393.30 + 393.5012] \approx 0.4 \quad (17)$$

Vi ved, at ovenstående er χ_1^2 fordelt (fordi vi restrikerer en parameter), den kritiske værdi er her 3.96, og da $0.4 < 3.96$ giver heller ej denne test anledning til at forkaste null hypotesen. P-værdien er her:

$$p_n(\theta_0 = 0.4) = (1 - \zeta(0.4)) \approx 0.52 \quad (18)$$

Hvor ζ er fordelingsfunktionen for en χ_1^2 fordeling.

Opgave 4

4.1 & 4.2

Vi kan kun bruge en Bernoulli fordeling i dette tilfælde, da der er binære udfald. Likelihood funktionen er da:

$$L_n(\theta \mid Y_1, \dots, Y_n) = \prod_{i=1}^n [\theta^{Y_i} \cdot (1 - \theta)^{1-Y_i}] \quad (19)$$

Log likelihood funktionen bliver da:

$$\log L_n(\theta \mid Y_1, \dots, Y_n) = \sum_{i=1}^n [Y_i \log(\theta) + (1 - Y_i) \log(1 - \theta)] \quad (20)$$

Med FOC:

$$\begin{aligned} \frac{\partial \log L_n}{\partial \theta} &= \sum_{i=1}^n \left[\frac{Y_i}{\theta} - \frac{(1 - Y_i)}{(1 - \theta)} \right] \\ \sum_{i=1}^n Y_i(1 - \theta) &= (n - \sum_{i=1}^n Y_i)\theta \\ \sum_{i=1}^n Y_i &= n\theta \\ \hat{\theta} &= \frac{1}{n} \sum_{i=1}^n Y_i \end{aligned}$$

Hvor $\hat{\theta} = p$.

4.3

Vi ved, at 95 pct. konfidensintervallet er givet ved:

$$\theta_0 \in [\hat{\theta}_n - 1.96 \cdot se(\hat{\theta}_n), \hat{\theta}_n + 1.96 \cdot se(\hat{\theta}_n)] \quad (21)$$

Vi skal nu erstatte 1.96 med de tilsvarende værdier for et 99 pct. konfidensintervallet, dvs $\Phi^{-1}(0.005)$ og $\Phi^{-1}(0.995)$, da variansen for en Bernoulli fordeling er: $\frac{p_0 \cdot (1 - p_0)}{n}$ må standardafvigelsen være:

$$se = \sqrt{0.24} \cdot \frac{1}{\sqrt{n}} \quad (22)$$

Hvorfor vi har:

$$\left[\hat{\theta}_n - 2.58 \cdot \sqrt{0.24} \cdot \frac{1}{\sqrt{n}}, \hat{\theta}_n + 2.58 \cdot \sqrt{0.24} \cdot \frac{1}{\sqrt{n}} \right]$$

Hvis afvigelserne nu maksimalt må være 0.04, har vi:

$$\begin{aligned} 2.58 \cdot \sqrt{0.24} \cdot \frac{1}{\sqrt{n}} &\leq 0.04 \\ 2.58 \cdot \sqrt{0.24} \cdot 1 &\leq 0.04 \cdot \sqrt{n} \\ \left[\frac{2.58 \cdot \sqrt{0.24} \cdot 1}{0.04} \right]^2 &\leq n \\ 998.45 &\leq n \end{aligned}$$

4.4

Hvis afvigelserne maksimalt må være 0.02, har vi:

$$\begin{aligned} 2.58 \cdot \sqrt{0.24} \cdot \frac{1}{\sqrt{n}} &\leq 0.02 \\ 2.58 \cdot \sqrt{0.24} \cdot 1 &\leq 0.02 \cdot \sqrt{n} \\ \left[\frac{2.58 \cdot \sqrt{0.24} \cdot 1}{0.02} \right]^2 &\leq n \\ 3993.84 &\leq n \end{aligned}$$

4.5

Vi kan blot finde det generelle resultat ved at indsætte:

$$\begin{aligned} 2.58 \cdot \sqrt{0.24} \cdot 1 &\leq 0.02 \cdot \sqrt{n} \\ \left[\frac{\sqrt{p_0 \cdot (1 - p_0)}}{d} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \right]^2 &\leq n \end{aligned}$$

4.6

Ved $n = 1000$ har vi netop ca. $d = 0.04$ som vist i 4.3.

Opgave 5

5.1

Fordi vi antager uafhængighed kan sandsynlighedsfunktionen under et skrives som:

$$f_{Z_i}(z \mid p_{11}, p_{12}, p_{21}, p_{22}, p_{31}, p_{32}) = p_{11}^{\mathbf{1}(z=(1,1))} \cdot p_{12}^{\mathbf{1}(z=(1,2))} \cdot p_{21}^{\mathbf{1}(z=(2,1))} \cdot p_{22}^{\mathbf{1}(z=(2,2))} \cdot p_{31}^{\mathbf{1}(z=(3,1))} \cdot p_{32}^{\mathbf{1}(z=(3,2))}$$

Idet indikatorfunktionen netop tæller, hvor mange gange observationen fremkommer.

5.2

Vi definerer nu $s_{11} = \sum_{i=1}^{111} \mathbf{1}(Z_i = (1, 1))$ osv, og får lader $p_{32} = 1 - p_{11} - p_{12} - \dots - p_{31}$, hvorfor log likelihood funktionen bliver:

$$\log L_n(\theta \mid Y_1, \dots, Y_{111}) = s_{11} \cdot \log(p_{11}) + s_{12} \cdot \log(p_{12}) + s_{21} \cdot \log(p_{21}) + s_{22} \cdot \log(p_{22})$$

$$+ s_{31} \cdot \log(p_{31}) + (111 - s_{11} - s_{12} - s_{21} - s_{22} - s_{13} - s_{31}) \cdot \log(1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31} - p_{13})$$

5.3

FOC er:

$$\frac{\delta L_n(\theta)}{\delta p_i} = 0 \quad \text{for } (1, \dots, 5) \quad (23)$$

Det vil sige:

$$\frac{s_{11}}{p_{11}} - \frac{111 - s_{11} - s_{12} - s_{21} - s_{22} - s_{13} - s_{31}}{1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31} - p_{13}} = 0$$

ned til:

$$\frac{s_{31}}{p_{31}} - \frac{111 - s_{11} - s_{12} - s_{21} - s_{22} - s_{13} - s_{31}}{1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31} - p_{13}} = 0$$

MLE estimaterne (og dermed de relative frekvenser er):

$$\begin{aligned} \hat{p}_{11} &= \frac{17}{111} \\ \hat{p}_{12} &= \frac{43}{111} \\ \hat{p}_{13} &= \frac{3}{111} \\ \hat{p}_{21} &= \frac{10}{111} \\ \hat{p}_{22} &= \frac{27}{111} \\ \hat{p}_{32} &= \frac{11}{111} \end{aligned}$$

Den maksimale værdi af Log likelihood funktionen fås ved indsættelse, dvs:

$$\max \log L_n = 17 \cdot \log\left(\frac{17}{111}\right) + 43 \cdot \log\left(\frac{43}{111}\right) + 3 \cdot \log\left(\frac{3}{111}\right) \dots 11 \cdot \log\left(\frac{11}{111}\right) \approx -171.2 \quad (24)$$

5.4

I de første linjer af koden defineres indikatorfunktionen, hvorefter STATA numerisk finder MLE. Bemærk, at den ikke finder $\bar{p}_{32}$, men det blot er residualt af de andre.

Opgave 6

6.1

For stokastisk uafhængighed gælder det, at den simultane ssh. er givet ved produktet af de marginale.

6.2

Hvis der gælder uafhængighed i modellen, skal vi blot estimere: p_1, p_2, p_3 , hvor p angiver kaffe og x_1, x_2 angiver rygning. Vi kan nu lade:

$$p_3 = 1 - p_1 - p_2 \quad (25)$$

og

$$x_1 = 1 - x_2 \quad (26)$$

Og dermed nøjes med at estimere, p_2, p_3, x_2 , hvorfor vi har 3 frie variable. Vi estimerer i STATA og får:

$$\begin{aligned} \hat{p}_1 &= .2432433 \\ \hat{p}_2 &= .6306306 \\ \hat{p}_3 &= 1 - .2432433 - .6306306 = 0.1262 \\ \hat{x}_1 &= .5675676 \\ \hat{x}_2 &= 1 - .5675676 = 0.4325 \end{aligned}$$

Den maksimale værdi af log-likelihood funktionen er:

$$\log L_n(\tilde{\theta}_n, z_1, \dots, z_{111}) = -175.3515 \quad (27)$$

0.1 6.3

Vi tjekker null hypotesen ved et LR test. Den ikke restrikerede model havde:

$$\log L_n(\hat{\theta}) = -171.2 \quad (28)$$

Derfor bliver LR statistikken:

$$LR = 2 \cdot (-171.2 + 175.35) = 8.3 \quad (29)$$

Vi ved, at LR er asymptotisk χ^2_2 fordelt, idet vi restrikerer to parametre (vi får to frihedsgrader mere). Den kritiske værdi er her:

$$\chi^2(2, 0.95)^{-1} \approx 5.99 \quad (30)$$

Og idet $8.3 > 5.99$ kan vi afvise null hypotesen ved $\alpha = 0.05$. Vi kan desuden regne p-værdien og se, at sandsynligheden for at få et mere ekstremt udfald end ovenstående er:

$$P_n = 1 - \chi^2_2(0.95) = 0.015 \quad (31)$$

Da $0.05 < 0.015$ betyder det, at sandsynligheden for at få et mere ekstremt udfald er under vores krav, hvorfor vi også her afviser null hypotesen.

6.4

Det er klart, at det ikke er en betinget model, idet udfaldene netop er uafhængige. For at vi kan opfatte det som en betinget model vil vi skulle antage, at rygning er eksogent givet og herefter antage PARVIS uafhængighed.

Opgave 7

7.1

Parameteren, α , kan tolkes som *baseline* CO2 udledningen, hvis virksomheden IKKE er i service-sektoren, mens β kan tolkes som en parameter, der angiver forskellen mellem baseline udledninger og udledninger for servicesektoren. Hvis $\beta < 0$ er udledningerne her mindre og vice versa. Vi ser, at $\beta \approx -0.27$, hvilket svarer til, at servicesektoren i gennemsnit udleder -0.27 ton CO_2 mindre, konstanten er ca. 0.0019.

7.2

Vi tester null hypotesen, der tilsiger, at der ikke er nogen signifikant forskel:

$$\mathcal{H}_0 : \beta = 0 \quad (32)$$

Mens den alternative hypotese er, at der er signifikant forskel:

$$\mathcal{H}_A : \beta \neq 0 \quad (33)$$

Vi tester med et Wald test og bestemmer derfor *Wald statistikken* til:

$$Z_n = \frac{\hat{\beta}_n - \beta_0}{\text{Se}(\hat{\beta}_n)} = \frac{-0.2791 - 0}{0.0545} = -5.12 \quad (34)$$

Vi ved, at ovenstående følger en standard normalfordeling, hvis null hypotesen er sand. De kritiske værdier er her -1.96 og 1.96 (på et 5 pct. signifikansniveau). Det er klart, at $-5.12 < -1.96$, hvorfor vi kan afvise hypotesen om, at der IKKE er nogen signifikant forskel og dermed IKKE KAN AFVISE hypotesen om, at der er en signifikant forskel (som der spørges om i opgaven. Vi kan også finde p-værdien som følgende (idet normalfordelingen er symmetrisk):

$$P_n(5.12) = 2 \cdot (1 - \Phi(5.12)) = 3.055e - 07 \approx 0 \quad (35)$$

Dvs. sandsynligheden for at få noget mere ekstremt er ca. 0, og dermed lagt under vores krav for ikke at kunne afvise null hypotesen.

7.2

Vi tester null hypotesen, der tilsiger, at forskellen er præcis -0.3:

$$\mathcal{H}_0 : \beta = -0.3 \quad (36)$$

Mens den alternative hypotese er, forskellen ikke er præcis -0.3:

$$\mathcal{H}_A : \beta \neq -0.3 \quad (37)$$

Vi tester med et Wald test og bestemmer derfor *Wald statistikken* til:

$$Z_n = \frac{\hat{\beta}_n - \beta_0}{\text{Se}(\hat{\beta}_n)} = \frac{-0.2791 - 0.3}{0.0545} \approx 0.38 \quad (38)$$

Vi ved, at ovenstående følger en standard normalfordeling, hvis null hypotesen er sand. De kritiske værdier er her -1.96 og 1.96 (på et 5 pct. signifikansniveau). Det er klart, at $-1.96 < 0.38 < 1.96$, hvorfor vi IKKE kan afvise null hypotesen. Vi kan også finde p-værdien som følgende (idet normalfordelingen er symmetrisk):

$$P_n(0.38) = 2 \cdot (1 - \Phi(0.38)) \approx 0.7 \quad (39)$$

Dvs. sandsynligheden for at få noget mere ekstremt er ca. 70 pct, og dermed lagt over vores krav til null hypotesen givet $\alpha = 0.05$...