STK 1110 2016, eksamen 28.11, skisse av løsning

Nils Lid Hjort, 28.11.2016

Oppgave 1

(a) Ved den vanlige transformsjonsformelen har $V_i = 2X_i/\theta$ tettheten $g(v) = f(\frac{1}{2}\theta v)\frac{1}{2}\theta$, og dette forenkles til $\frac{1}{2}\exp(-\frac{1}{2}v)$, for $v > 0$. Dette er ganske riktig χ^2_2-tettheten.

(b) Fra

$$\mathrm{E}\,\frac{2X_i}{\theta} = 2 \quad \text{og} \quad \mathrm{Var}\,\frac{2X_i}{\theta} = 4$$

følger $\mathrm{E}\,X_i = \theta$ og $\mathrm{Var}\,X_i = \theta^2$.

(c) Estimatoren $\widehat{\theta} = \bar{X}$ får forventning θ, og er dermed forventningsrett, og varians θ^2/n.

(d) Sentralgrenseteoremet medfører her at $\widehat{\theta}$ er tilnærmet normalfordelt, med sin forventning θ og sin varians θ^2/n. Dette igjen medfører at

$$t_1 = \frac{\widehat{\theta} - \theta}{\widehat{\theta}/\sqrt{n}} \approx_d \mathrm{N}(0,1).$$

Det følger av dette at

$$\widehat{\theta} \pm 1.96\widehat{\theta}/\sqrt{n} = \widehat{\theta}(1 \pm 1.96/\sqrt{n})$$

er et konfidensintervall for θ med dekningsgrad tilnærmet 95%.

– Man kan også ta utgangspunkt i at

$$t_2 = \frac{\widehat{\theta} - \theta}{\theta/\sqrt{n}} = \sqrt{n}(\widehat{\theta}/\theta - 1) \approx_d \mathrm{N}(0,1),$$

hvilket etter transformasjonen fra starten av oppgaven er ensbetydende med

$$\sqrt{n}\{\chi^2_{2n}/(2n) - 1\} \approx_d \mathrm{N}(0,1).$$

Med dette som utgangspunkt, altså $\theta/\sqrt{n}$ som nevner og ikke $\widehat{\theta}/\sqrt{n}$, ledes man til

$$\frac{\widehat{\theta}}{1 + 1.96/\sqrt{n}} \le \theta \le \frac{\widehat{\theta}}{1 - 1.96/\sqrt{n}}$$

som tilnærmet 95% konfidensintervall. Endelig kan man også konstruerer et eksakt intervall, via egenskapen $\widehat{\theta}/\theta \sim \chi_{2n}/(2n)$.

(e) Her vil $\bar{Y} - \bar{X}$ ha forventning $\theta_2 - \theta_1$, som er null under H_0, og varians $\theta_1^2/n + \theta_2^2/n$. En fornuftig testobservator er derfor

$$t_1 = \frac{\bar{Y} - \bar{X}}{(\widehat{\theta}_1^2/n + \widehat{\theta}_2^2/n)^{1/2}},$$

med den den egenskap at $t_1 \approx_d \mathrm{N}(0,1)$ under H_0. Det samme gjelder

$$t_2 = \frac{\bar{Y} - \bar{X}}{(\widehat{\sigma}_1^2/n + \widehat{\sigma}_2^2/n)^{1/2}},$$

med de empiriske variansene. Her finner man

$$t_1 = (4.444 - 2.222)/\sqrt{4.444^2/n + 2.222^2/n} = 2.000,$$
$$t_2 = (4.444 - 2.222)/\sqrt{2.468^2/n + 4.987^2/n} = 1.786.$$

Den første er en anelse mer nøyaktig enn den andre, idet man utnytter eksplisitte formler for variansene, men den andre kan alltids skryte av at den er modellrobust. Verdiene t_1 og t_2 ligger på kanten av forkastning for en 0.05-nivå-test.

(f) Fra transformasjonene over finner vi at

$$\sum_{i=1}^{n} V_i = \frac{2n\bar{X}}{\theta} \sim \chi^2_{2n}.$$

Av dette følger at

$$F = \frac{\bar{Y}}{\bar{X}} = \frac{\theta_2}{\theta_1}\frac{A/(2n)}{B/(2n)} = \rho F_0,$$

der A og B er uavhengige og χ^2_{2n}-fordelte, som betyr at F_0 er F-fordelt med frihetsgrader $(2n, 2n)$. Fra $\Pr(c \le F_0 \le d) = 0.95$, der $c = 0.533$ og $d = 1.875$, pr. tabellen bakerst, følger intervallet $[F/d, F/c]$, som med $F = 2.000$ gir $[1.067, 3.752]$. Data tyder altså på at $\rho > 1$, altså $\sigma_2 > \sigma_1$, på nivå 0.05. Denne analysen er mer presis, altså skarpere, enn den via approksimasjonene over.

Oppgave 2

(a) Siden $\widehat{\theta}_i \sim \mathrm{N}(\theta_i, \sigma^2)$, med kjent $\sigma = 2.112$, er $\widehat{\theta}_i \pm 1.96\,\sigma$ et 95% konfidensintervall for θ_i. Altså $13.37 \pm 1.96\,\sigma = [9.230, 17.510]$ for Leicester og $7.43 \pm 1.96\,\sigma = [3.290, 11.570]$ for Leeds.

(b) Vi har

$$\widehat{\theta}_i = \theta_i + \varepsilon_i = \theta_0 + \delta_i + \varepsilon_i \sim \mathrm{N}(\theta_0, \sigma^2 + \tau^2)$$

siden en sum av uavhengige normalfordelte størrelser er normalfordelt, og da er det jo nok å beregne forventning og varians.

(c) Størrelsen S^2 er en forventningsrett estimator for variansen til $\widehat{\theta}_i$-ene, altså $\sigma^2 + \tau^2$. Da er det bare å trekke fra og ta en kvadratrot, så $\widehat{\tau} = (S^2 - \sigma^2)^{1/2} = 2.366$ er et fornuftig estimat for τ.

(d) Under $H_0\colon \tau = 0$ er S^2 altså den vanlige estimator for σ^2, og $S^2 \sim \sigma^2\chi^2_\nu/\nu$, med $\mu = 10$. Vi forkaster H_0 dersom S^2 er stor nok, nemlig $(S^2/\sigma^2)\nu \ge 18.3070$, der denne skranken er 0.95-kvantilen for χ^2_ν. Dette svarer til $S^2/\sigma^2 \ge 1.8307$ som forkastningsgrense, og her er $S^2/\sigma^2 = 2.2554$. Altså forkaster vi H_0 og tror at $\tau > 0$.

(e) Vi har $\Pr(c \le \chi^2_\nu \le d) = 0.95$ med $c = 3.2470$ og $d = 20.4832$. Siden $S^2 \sim (\sigma^2 + \tau^2)\chi^2_\nu/\nu$ får vi intervallet $[S^2\nu/d, S^2\nu/c] = [4.9117, 20.9846]$ for parameteren $\sigma^2 + \tau^2$. Ved igjen å trekke fra σ^2 og kvadratrote det til får vi 95% intervallet $[0.6717, 5.1501]$ for τ.

Oppgave 3

(a) Vi har

$$\ell(p) = x \log p + (n - x) \log(1 - p) + \log \binom{n}{x},$$

med derivert

$$\frac{x}{p} - \frac{n - x}{1 - p}.$$

Settes denne lik 0 får vi $\widehat{p} = x/n$, og en sjekk viser at dette virkelig er er makspunkt.

(b) Her får vi

$$\begin{aligned}\ell(p, q, r) = x \log p + (n - x) \log(1 - p) + y \log q + (n - y) \log(1 - q) \\ + z \log r + (n - z) \log(1 - r) + c,\end{aligned}$$

der

$$c = \log \binom{n}{x} + \log \binom{n}{y} + \log \binom{n}{z}$$

er en konstant vi i grunnen ikke får bruk for. Om $p = q = r$ holder kan de tre leddene slås sammen, med maksimand $\bar{p} = (x + y + z)/(3n)$.

(c) Her får vi

$$\begin{aligned}\ell_{\max}(\text{big}) = x \log \widehat{p} + (n - x) \log(1 - \widehat{p}) + y \log \widehat{q} + (n - y) \log(1 - \widehat{q}) \\ + z \log \widehat{r} + (n - z) \log(1 - \widehat{r}) + c,\end{aligned}$$

der c er en konstant som blir den samme under både big og under H_0, og der $\widehat{p} = 17/50$, $\widehat{q} = 22/50$, $\widehat{r} = 14/50$. Vi får $-95.9959 + c$. Under H_0 får vi

$$\begin{aligned}\ell_{\max}(H_0) = x \log \bar{p} + (n - x) \log(1 - \bar{p}) + y \log \bar{p} + (n - y) \log(1 - \bar{p}) \\ + z \log \bar{p} + (n - z) \log(1 - \bar{p}) + c = -97.4229 + c.\end{aligned}$$

Differansen blir

$$\Delta = 2\{\ell_{\max}(\text{big}) - \ell_{\max}(H_0)\} = 2.8539.$$

Her er $\Delta \approx_d \chi^2_2$ under H_0. Verdien 2.8539 er for liten til å gjøre inntrykk på en χ^2_2 (p-verdien blir f.ø. 0.240).

Oppgave 4

(a) Verdiene $\widehat{a} = 18.0346$ og $\widehat{b} = 0.0588$ er gitt i utskriften, og

$$\widehat{\sigma} = \sqrt{41388.09/(n - 2)} = 47.9514.$$

(b) Dersom x går ned med 100, går den forventende y ned med $100\,b$, som estimeres med $100\,\widehat{b} = 5.884$ per hundre tusen, eller ca. 59 reddede mennesker per million.

(c) For $x_0 = 3900$ blir estimatet for y lik $\widehat{y}_0 = \widehat{a} + \widehat{b}x_0 = 247.51$.

(d) Vi har

$$Z_0 = Y_0 - \widehat{y}_0 = a + bx_0 + \varepsilon_0 - \widehat{a} - \widehat{b}x_0 = \varepsilon_0 + a - \widehat{a} + (b - \widehat{b})x_0.$$

Den har forventning 0, siden $\widehat{a}$ og $\widehat{b}$ er forventningsrette, og variansen blir

$$\operatorname{Var}\varepsilon_0 + \operatorname{Var}\bar{Y} + (x_0 - \bar{x})^2 \operatorname{Var}\widehat{b} = \sigma^2\{1 + 1/n + (x_0 - \bar{x})^2/M\}.$$

Dessuten er Z_0 normalfordelt, qua lineærkombinasjon av uavhengige normalfordelte variable.

(e) Fra dette får vi

$$t = \frac{Z_0}{\widehat{\sigma}\{1 + 1/n + (x_0 - \bar{x})^2/M\}^{1/2}} \sim t_{n-2}.$$

Et prediksjonsintervall som med sannsynlighet 95% vil inneholde Y_0 er derfor

$$\widehat{y}_0 \pm 1.734\{1 + 1/n + (x_0 - \bar{x})^2/M\}^{1/2} = 247.5101 \pm 56.6170 = [149.34, 345.68].$$

Sannheten er f.ø. at y for USA i 1962 var 256.9. Prediksjonsintervallet bygger på flere forutsetninger, bl.a. at modellen kan ekstrapoleres et stykke mot høyre (ingen andre land røyket så meget som USA, i 1962). Dette kan man ikke tro på uten videre, men her har statistikken altså slått til.

(f) For Ungarn 2014 er det bare å løse ligningen $a + bx_u = y_u = 172.6$ for å komme frem til estimatet $\widehat{x}_u = (y_u - \widehat{a})/\widehat{b} = 2619.48$, antallet sigaretter pr. år pr. person.