Løsningsforslag vår 2013 - STAT100

1) $n = 20$ uavhengige forsøk med to mulige utfall i hvert forsøk (ja/nei) → Binomisk
Ingen info om verdi på sannsynligheten p → ukjent

Svar: $X \sim Bin(20, p)$

2) $\hat{p} = \frac{8}{20} = 0.4$

$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4 \cdot 0.6}{20}} = 0.11$

Et tilnærmet 95 % KI for p blir

$[\hat{p} \pm z_{0.025} \cdot SE(\hat{p})] = [0.4 \pm 1.96 \cdot 0.11] \approx [0.18, 0.61]$

3)

$$n \geq 4 \cdot p(1-p)\left(\frac{z_{\frac{\alpha}{2}}}{L}\right)^2 = 4 \cdot 0.5 \cdot 0.5\left(\frac{1.96}{0.05}\right)^2 \approx 1537$$

4) $X \sim B(400, p)$

Utfall: $X = 160$

$\hat{p} = \frac{160}{400} = 0.4$

$H_0: p = p_0 = 0.35 \quad H_1: p > 0.35$

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.4 - 0.35}{\sqrt{\frac{0.35 \cdot 0.65}{400}}} = 2.10$$

5) Skal finne $\hat{E}_{13}$ som under H_0 er lik $n \cdot \hat{p}_{13} = n \cdot \hat{p}(A_1)\hat{p}(B_3)$

$$= n \cdot \frac{R_1}{n} \cdot \frac{K_3}{n} = \frac{R_1 \cdot K_3}{n} = \frac{140 \cdot 240}{437} = 76.89$$

6) Testobservatoren Q er kji-kvadratfordelt med $(3-1)(3-1) = 4$ frihetsgrader. Vi forkaster H_0: Uavhengige rad og kolonnevariabler hvis $Q > \chi^2_{\alpha,4}$ for testnivå α. Vi må finne i χ^2-tabellen den største verdien i rad 4 som er mindre enn $Q = 16.7$.

Det er $\chi^2_{0.005,4} = 14.86$

Det laveste testnivået som gir forkastning er da $\underline{\underline{0.005}}$

7) $$Q = \sum_{i=1}^{3}\sum_{j=1}^{3} q_{ij} \quad \text{der} \quad q_{ij} = \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

I utskriften er alle q_{ij} gitt nederst unntatt

$$q_{32} = \frac{(X_{32} - E_{32})^2}{E_{32}} = \frac{(65 - 77.99)^2}{77.99} = 2.16$$

Dermed blir

$$Q = 0.06 + 0.05 + 3.17 + 2.60 + 2.63 + 2.16 = \underline{\underline{10.67}}$$

8) Dette er den betingede sannsynligheten for at en nordmann alltid bruker bilbelte, dvs

$P(\text{alltid bilbelte} \mid \text{nordmann}) = \frac{75}{140} = \underline{\underline{0.54}}$

PS. (Her får vi ikke beskjed om å anta noen H_0 er sann. Dersom vi skulle anta H_0 er sann, ville det bli slik:

$$P(\text{alltid bilbelte} \mid \text{nordmann}) = P(B_3 \mid A_1) = \frac{P(A_1 \cap B_3)}{P(A_1)}$$

$$= \frac{\cancel{P(A_1)} \cdot P(B_3)}{\cancel{P(A_1)}} = P(B_3) = \frac{240}{437} = \underline{\underline{0.55}}$$)

9) $Y \sim N(\mu, \sigma)$ Én-utvalgs situasjon

Skal teste om avl har ført til at μ har økt og blitt større enn $\mu_0 = 0.8$

Da blir riktig hypoteser:

$H_0: \mu = 0.8 \qquad H_1: \mu > 0.8$

10) Ukjent σ så vi må bruke T-test

$$T = \frac{\bar{Y} - \mu_0}{S_y/\sqrt{n}} = \frac{0.90 - 0.80}{0.211/\sqrt{14}} = \underline{\underline{1.77}}$$

11) En variant av minste kvadraters estimatoren for β er:

$$\hat{\beta} = r \cdot \frac{S_y}{S_x} = 0.95 \cdot \frac{0.211}{0.766} = \underline{\underline{0.026}}$$

12) Et 95% KI for α (konstantleddet i modellen) er

$$\left[\hat{\alpha} \pm t_{\frac{\alpha}{2}, n-2} \cdot SE(\hat{\alpha})\right] = \left[0.130045 \pm 2.179 \cdot 0.002487\right]$$
$$= \underline{\underline{[-0.035 \,,\, 0.295]}}$$

(Da $t_{0.025, 12} = 2.179$)

13) Fra R-utskriften ser vi at p-verdien for å teste $H_0: \alpha = 0$ mot $H_1: \alpha \neq 0$ er 0.111.

Generelt forkaster man H_0 hvis p-verdi $< \alpha$.

Mao, dersom $\alpha > 0.111$ forkastes H_0

testnivået

Generelt har vi også at dersom man forkaster H_0 for testnivå α (tosidig test) vil et $(1-\alpha)\cdot 100\%$ konfidensintervall ikke inneholde verdien under nullhypotesen.

Dermed vil et $(1-0.111)\cdot 100\%$ KI for konstantleddet α i modellen akkurat inneholde verdien 0, mens f.eks et $(1-0.12)\cdot 100\%$ altså et 88% KI ikke vil inneholde 0.
(Dette svarer til en test på 12% nivå, noe som vil gi forkastning av H_0)

Riktig svar: 88% intervall

14) Et 95% KI for $E(y|x)$ er

$$\left[\hat{y} \pm t_{0.025,12} \cdot s \cdot \sqrt{\frac{1}{n} + \left(\frac{x-\bar{x}}{s/SE(\hat{\beta})}\right)^2}\right]$$

Bredden av intervallet er

$$2 \cdot t_{0.025,12} \cdot s \cdot \sqrt{\frac{1}{n} + \left(\frac{x-\bar{x}}{s/SE(\hat{\beta})}\right)^2}$$

$$= 2 \cdot 2.179 \cdot 0.06864 \cdot \sqrt{\frac{1}{14} + \left(\frac{30.56-27.56}{0.06864/0.002487}\right)^2}$$

$$= 0.081$$

15) Denne kan løses på flere måter, bl.a.:

a) I enkel lineær regresjon vil R^2 være lik r^2 der r = korrelasjonen mellom x og y.

Dermed: $R^2 = 0{,}95^2 = \underline{\underline{0{,}90}}$

b) Pr definisjon er generelt (også i multippel regresjon)

$$R^2 = \frac{SS_R}{SS_T}$$

Vi må finne SS_T som ikke er gitt i utskriften.
Har at $SS_T = SS_R + SS_E$
Videre vet vi at $MS_E = \frac{SSE}{n-2}$ i enkel regresjon
dvs $SSE = MS_E \cdot (n-2)$
Til slutt: $MS_E = s^2$ som er gitt i utskrift
Da får vi samlet sett

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_R}{SS_R + SS_E} = \frac{SS_R}{SS_R + MS_E(n-2)} = \frac{SS_R}{SS_R + s^2(n-2)}$$

$$\frac{0{,}52221}{0{,}52221 + 0{,}06864^2 \cdot (14-2)} = \underline{\underline{0{,}90}}$$

16) R^2 er andel av variasjon i responsvariabelen som kan forklares ved den lineære sammenhengen til x (i regresjon), eller ved at de tilhører ulike nivå av en gruppevariabel (i ANOVA-modeller).

Her har vi regresjon og $\underline{\underline{\text{alt. c)}}}$ passer

17) σ er et mål på hvor mye responsvariabelen y vil variere dersom man holder x fast og gitt på én verdi. I regresjon antar vi at denne variasjonen er den samme uansett hvilken verdi av x vi observerer y ved.
Alt. a) er riktig i denne sammenhengen

18) SSG i ANOVA med én gruppevariabel med k nivåer har alltid k-1 frihetsgrader. Her har vi k=5, dermed 4 frihetsgrader

19) Hypoteser:
$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ mot H_1: minst to μ_i'er forskjellige

Forkaster H_0 dersom $F = \frac{MSG}{MSE}$ blir større enn $F_{\alpha, k, n-k}$.

Altså $F = \frac{MSG}{MSE} = \frac{0{,}08}{0.018} = \underline{4.44}$

20) $\hat{\theta} = \frac{(\bar{y}_1 + \bar{y}_2 + \bar{y}_3)}{3} - \frac{(\bar{y}_4 - \bar{y}_5)}{2} = \frac{(4.225 + 4.250 + 4.150)}{3} - \frac{(4.0 + 3.925)}{2}$

$= \underline{0.25}$

21) θ kan skrives slik:

$$\theta = \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \frac{1}{2}\mu_4 - \frac{1}{2}\mu_5$$

$$c_1 = \frac{1}{3} \quad c_2 = \frac{1}{3} \quad c_3 = \frac{1}{3} \quad c_4 = -\frac{1}{2} \quad c_5 = -\frac{1}{2}$$

Alt. b) er korrekt

22) Hvis snittet av forventningene for 1978 og 1980 er lavere enn snittet av de 3 årene før vil θ bli ~~...~~ større enn 0

Vi skal derfor teste:

$H_0: \theta = 0$ mot $H_1: \theta > 0$

23) Vi kan skrive slik

$$\theta = \mu_3 - \mu_4 = 0 \cdot \mu_1 + 0 \cdot \mu_2 + 1 \cdot \mu_3 + (-1) \cdot \mu_4 + 0 \cdot \mu_5$$

som gir $c_1 = 0$, $c_2 = 0$, $c_3 = 1$, $c_4 = -1$, $c_5 = 0$

$SE(\hat{\theta})$ er gitt ved

$$SE(\hat{\theta}) = \sqrt{MS_E \cdot \sum \frac{c_i^2}{n_i}}$$

$$= \sqrt{MS_E\left(\frac{c_1^2}{4} + \frac{c_2^2}{4} + \frac{c_3^2}{4} + \frac{c_4^2}{4} + \frac{c_5^2}{4}\right)}$$

$$= \sqrt{MS_E\left(\frac{c_3^2}{4} + \frac{c_4^2}{4}\right)} \quad \text{siden } c_1 = c_2 = c_5 = 0$$

$$= \sqrt{0{,}018\left(\frac{1^2}{4} + \frac{(-1)^2}{4}\right)} = \underline{\underline{0{,}095}}$$

24) Dette er $\underline{\underline{MSE}}$

25) MSE er et felles variansmål basert på utvalgsvariansene i alle k gruppene siden vi antar at alle grupper har samme σ.
Det kan nevnes at for $k = 2$ er $S_p^2 = MSE$ og for $k > 2$ utvider vi S_p^2 formelen til k ledd som en variant av MSE:

$$MSE = (S_p^2 =) = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \dots + (n_k-1)S_k^2}{n_1+n_2+\dots+n_k - k}$$

$$= \frac{(n_1-1)\cdot S_1^2 + (n_2-1)\cdot S_2^2 + \dots + (n_k-1)S_k^2}{n-k}$$

Telleren er her det samme som SSE.
Siden alle $n_i = 4$ kan vi skrive:

$$3\cdot S_1^2 + 3\cdot S_2^2 + \dots 3\cdot S_5^2 = SSE$$

S_1 er ukjent og er den som skal beregnes. De andre er kjent. Dermed

$$3\cdot S_1^2 + \underbrace{3\cdot 0{,}129^2 + 3\cdot 0{,}129^2 + 3\cdot 0{,}141^2 + 3\cdot 0{,}171^2}_{=0{,}2472} = \underset{\uparrow \atop SSE}{0{,}27}$$

$$3\cdot S_1^2 = 0{,}27 - 0{,}2472$$

$$S_1 = \sqrt{\frac{0.27 - 0.2472}{3}} = \underline{\underline{0.087}}$$

26) β i modellen angir forventet endring i y når x øker med én enhet.

$\hat{\beta}$ er den estimerte forventede endringen i y når x øker med én enhet.

At $\hat{\beta} = -1.17$ betyr at vi estimerer y til å avta med 1.17 når avstanden øker med 1 meter.

Alt. b) er korrekt.

27) $H_0: \beta = 0$ mot $\beta < 0$

Testobservator:

$$T = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{-1.17}{0.1264} = \underline{\underline{-9.259}}$$

28) Vi forkaster den ensidige testens H_0 dersom T blir veldig liten (stor negativ) og mer spesifikt dersom $T < -t_{\alpha, n-2}$

Her har vi $n-2 = 20-2 = 18$ frihetsgrader

Da får vi fra tabell over t-fordeling:

$$-t_{0.05, 18} = \underline{\underline{-1.734}}$$

29) Med tosidig alternativ vil p-verdien fordobles og det blir mindre sannsynlighet for p-verdi $< \alpha$ og forkastning.

Alt. e) er korrekt.

30) $\hat{y} = 129.1644 + (-1.1704) \cdot 10 = \underline{\underline{117.5}}$

31) Av utskriften merker vi oss at $\bar{x} = 59.95$. Vi skal altså finne et prediksjonsintervall for $x = \bar{x}$ (der det er smalest)

Formelen sier at et $(1-\alpha)\cdot 100\%$ PI for $\hat{y}$ er

$$\hat{y} \pm t_{\frac{\alpha}{2}, n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \left(\frac{x-\bar{x}}{s/SE(\hat{\beta})}\right)^2}$$

Leddet $\left(\frac{x-\bar{x}}{s/SE(\hat{\beta})}\right)^2$ blir 0 når $x = \bar{x}$

$t_{0.025, 18} = 2.101$

$s = 18.15$

$SE(\hat{\beta}) = 0.1264$

$n = 20$

$\hat{y} = 59.0$ (gitt i tekst)

Plugg inn og få

$$\left[59.0 \pm 2.101 \cdot 18.15 \cdot \sqrt{1 + \frac{1}{20} + 0} \right]$$

$$= \underline{\underline{[19.9 \;,\; 98.1]}}$$

32) Tolkning: dersom vi går 59.95 meter fra vegen vil den observerte svevestøvkonsentrasjonen vi får med 95% sannsynlighet være mellom 19.9 og 98.1

$\underline{\underline{\text{Alt. c)}}}$

33) For $x = 59.5$ har vi at $\hat{y} = 59.0$ i følge modellen.
Observert verdi er 45.0
Da blir residualet

$$e = y - \hat{y} = 45.0 - 59.0 = \underline{\underline{-14.0}}$$

34) Utvalgsstandardavvik:

$$S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{4}(y_i - \bar{y})^2} \quad (\text{Definisjon 2.3 i boks})$$

$$= \sqrt{\frac{1}{3-1}\left((1300-1112.5)^2 + (960-1112.5)^2 + (1150-1112.5)^2 + (1040-1112.5)^2\right)}$$

$$= \underline{\underline{147.3}}$$

35) Antar $X \sim N(\mu, \sigma)$

$$\hat{\mu} = \bar{x} = 1145$$
$$\hat{\sigma} = S_x = 156.7$$

$H_0: \mu = 1000$ mot $H_1: \mu \neq 1000$

Testobservator: $T = \dfrac{\bar{x} - \mu_0}{S_x/\sqrt{n}} = \dfrac{1145 - 1000}{156.7/\sqrt{4}} = \underline{\underline{1.85}}$

36) W er t-fordelt med 3 frihetsgrader
(En-utvalgssituasjon med $n = 4$ observasjoner og ukjent standardavvik σ_D)

37) I en ANOVA modell med én gruppevariabel med k nivåer og totalt N observasjoner, vil SS_E ha $N-k$, dvs her: $\underline{\underline{24-4=20}}$ frihetsgrader

38) SS_T har alltid $N-1$ frihetsgrader,
dvs $24-1 = \underline{\underline{23}}$

39) Et forventningsrett estimat for σ^2 er MSE
$\hat{\sigma}^2 = MSE = \underline{\underline{150}}$

40) Sannsynligheten for at en observert F blir større enn F_α under H_0 er lik α. Hvis dette skjer forkastes H_0 feilaktig (type I feil)
Dette kan alternativt uttrykkes ved
$P(\text{Forkaste } H_0 \mid H_0 \text{ er sann})$