Løsningsforslag STAT100 Vår 2014

1) Hvis forventet mengde E. coli er lavere i B er altså

$\mu_A > \mu_B \Rightarrow \mu_A - \mu_B > 0$

Dette blir H_1 som vi ønsker å påvise.

H_0 blir da $\mu_A - \mu_B = 0$ (ingen forskjell)

e) er korrekt.

2) Målinger er gjort begge steder i de samme månedene. En evt måned til måned variasjon som påvirker begge steder kan skape avhengighet mellom målinger i samme måned. Denne blir man kvitt ved å analysere som parvise data.

d) er korrekt.

3) Én-utvalgs t-test på differanser.

Testobservator:

$$T = \frac{\hat{\mu}_D - \mu_0}{s_D/\sqrt{n}} = \frac{10.30 - 0}{10.54/\sqrt{10}} = 3.09$$

b) er korrekt.

4) Én utvalgs t-test har generelt $n-1$ frihetsgrader.

Her: $n = 10$, dvs 9 frihetsgrader

a) er korrekt.

5) $H_0: p = 0.5$ $H_1: p > 0.5$

La X være antall måneder av $n=10$ at B har lavere E. coli konsentrasjon enn A.

Da er $X \sim B(10, p)$. Hvis $p > 0.5$ er B bedre enn A (lavere bakteriekons).

Vi observerer $X = 9$.

Under H_0 er $X \sim B(10, 0.5)$.

p-verdi for hypotesetest er $P(X \geq 9 \mid H_0 \text{ er sann})$

dvs $P(X \geq 9)$ når $p = 0.5$

$$P(X \geq 9) = 1 - P(X \leq 8)$$
$$= 1 - 0.989 \quad \text{(jfr. tabell E.1)}$$
$$= 0.011$$

c) er korrekt

6) To-utvalgs t-test

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \dots = \frac{S_1^2 + S_2^2}{2} = \frac{11.83^2 + 10.93^2}{2}$$
$$= 129.71$$

$S_p = 11.39$

Testobservator:

$$T = \frac{213.7 - 203.4}{11.39\sqrt{\frac{1}{10} + \frac{1}{10}}} = 2.02$$

c) er korrekt.

7) $n_1 + n_2 - 2 = 10 + 10 - 2 = 18$ frihetsgrader

e) er korrekt

8) Vet at $MSG = \frac{SSG}{k-1} \Leftrightarrow SSG = MSG \cdot (k-1)$

Her er $k = 3$ (gjødseltyper)

$MSG = 7.4533$

Dermed : $SSG = 7.4533 \cdot (3-1) = 14.907$

b) er korrekt

9)

$$F = \frac{MSG}{MSE} = \frac{7.4533}{1.2178} = 6.12$$

e) er korrekt.

10)

Residual : $Y_{ij} - \bar{Y}_i$ der i er gjødseltype
j er observasjon under gjødseltype

Får vite at for $i = 1$ (type A) for en eller annen j er $Y_{ij} = 26.0$.

Av utskrift ser vi at $\bar{Y}_1 = 25.0$ for $i = 1$

$\hat{\varepsilon}_{ij} = 26.0 - 25.0 = 1.0$

d) er korrekt.

11) $\hat{\theta} = \frac{\bar{Y}_1 - \bar{Y}_2}{2} - \bar{Y}_3 = \frac{(25.0 - 24.7)}{2} - 22.5 = 2.35$

e) er korrekt.

12) A og B er kunstgjødsler. Hvis de har høgere gjennomsnittsforventning enn C vil $\theta > 0$. dermed

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

d) er korrekt.

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

$c_1 = 1 \quad c_2 = -1 \quad c_3 = 0$

Forventningsrett estimator er

$\hat{\theta} = 1 \cdot \bar{Y}_1 + (-1) \cdot \bar{Y}_2 + 0 \cdot \bar{Y}_3$

Denne har standardfeil

$$SE(\hat{\theta}) = \sqrt{MSE \cdot \sum_{i=1}^{3} \frac{c_i^2}{n_i}}$$

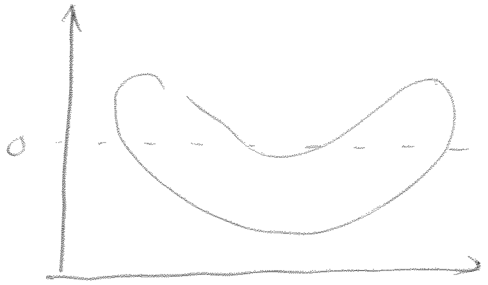
$$= \sqrt{1.2178 \left(\frac{1^2}{4} + \frac{(-1)^2}{4} + \frac{0^2}{4} \right)}$$

$$= 0.78$$

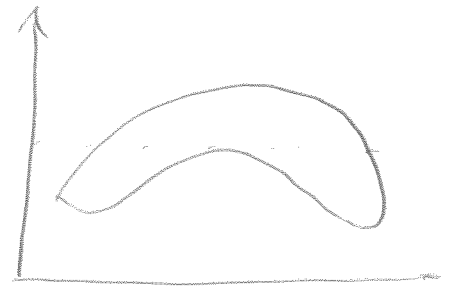
a) er korrekt.

14) c) er korrekt.

15) Linearitet er OK siden det ikke er noe kurve å se i plottet.
Dvs ikke noe slikt:

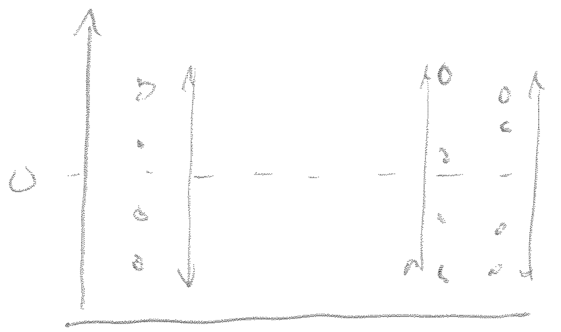


eller



f.eks.

Konstant varians er OK siden variasjonen synes å være den samme i alle tre grupper



Samme spredning rundt 0 for residualene

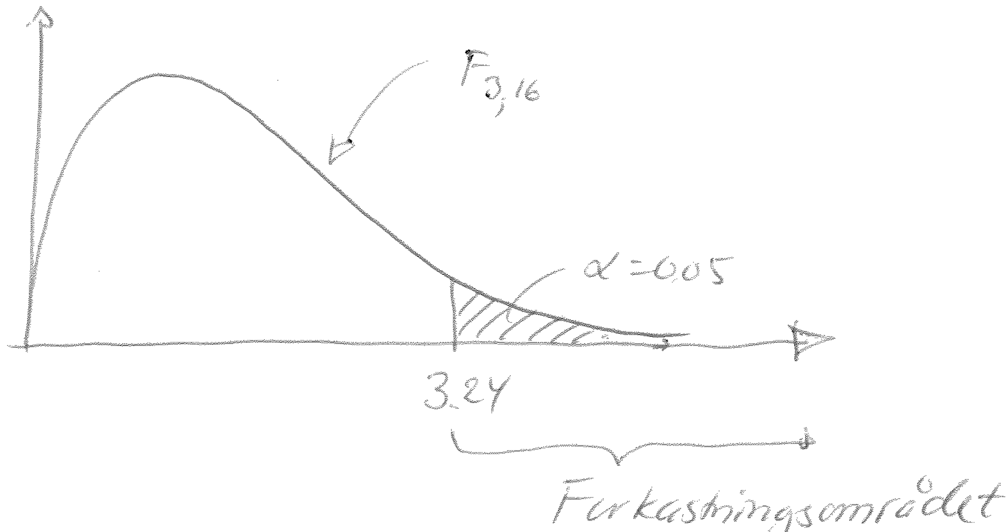
d) er korrekt.

16) σ beskriver uforklart variasjon, dvs den som ikke kan fanges opp av faktoren gjødseltype. Denne antas å være den samme i alle grupper. Mao, for en eller annen gjødseltype vil avling variere av andre årsaker. Denne variasjonen beskrives av σ.

a) er korrekt.

17) Testobservatoren F har 3 og 16 frihetsgrader. Av tabell for F med testnivå $\alpha = 0{,}05$ finner vi i kolonne 3 og rad 16 tallet 3.24.

Altså



b) er korrekt.

18) Vi uttaler oss aldri om sannsynligheten for at en hypotese er korrekt, så e) må være feil. Alle de andre er korrekte utsagn om testnivået α.

e) er korrekt.

19) Et 99% KI for θ blir

$$\hat{\theta} \pm t_{0.005,n-k} \cdot SE(\hat{\theta}) \quad \text{der } t_{0.005,16} = 2.921$$

Dette gir

$$-32 \pm 2.921 \cdot 10.95 = [-63.98, -0.015]$$

c) er korrekt.

20) Hypoteser : $H_0 : \theta = -10$ $H_1 : \theta < -10$

Under H_0 er testobservatoren :

$$T = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} = \frac{-32 - (-10)}{10.95} = -2.01$$

t-fordelt med $n-k = 20-4 = 16$ frihetsgrader.

Forkast H_0 hvis $T < -t_{0.05,16} (= -1.746)$

Vi ser at $-2.01 < -1.746$ og vi forkaster H_0 på 5% nivå

d) er korrekt.

21)

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

Standardfeilen til estimatoren blir

$$SE(\hat{\theta}) = \sqrt{MSE \left(\frac{1^2}{5} + \frac{(-1)^2}{5} \right)} = 10.95$$

Siden øvre grense i intervallet er lik

$$\hat{\theta} + t_{\frac{\alpha}{2},16} \cdot SE(\hat{\theta}) = -2.78 \quad \text{(oppgitt)}$$

og $\hat{\theta} = \bar{y}_3 - \bar{y}_4 = 196 - 222 = -26$

$SE(\hat{\theta}) = 10.95$

finner vi at

$$-26 + t_{\frac{\alpha}{2},16} \cdot 10.95 = -2.78$$

$$\Updownarrow$$

$$t_{\frac{\alpha}{2},16} = \frac{-2.78 + 26}{10.95} = \underline{2.12}$$

Av rad 16 i t-tabell finner vi at $t = 2.12$ for areal lik 0.025. Dette arealet er altså $\frac{\alpha}{2}$. Dermed er $\alpha = 0.05$ og vi har et 95% intervall. c) er korrekt.

22) $R^2 = \frac{SS_G}{SS_T} = \frac{SS_G}{SS_G + SS_E} = \frac{4820}{4820 + 4800} = 0.501$

c) er korrekt.

23) La utvalgsstandardavviket for gård i være S_i.

Da vil $MS_E = \frac{(n_1-1) \cdot S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2 + (n_4-1)S_4^2}{n-k}$

(utvidet S_p^2 til fire grupper)

Siden det er like mange observasjoner i alle grupper kan den forenkles til

$$MS_E = \frac{S_1^2 + S_2^2 + S_3^2 + S_4^2}{4}$$

Alle tall er kjente unntatt S_3. Setter inn:

$$300 = \frac{15.81^2 + 18.71^2 + S_3^2 + 19.24^2}{4}$$

$$\Updownarrow$$

$$4 \cdot 300 - 15.81^2 - 18.71^2 - 19.24^2 = S_3^2$$

$$\Updownarrow$$

$$S_3 = \sqrt{4 \cdot 300 - 15.81^2 - 18.71^2 - 19.24^2} = \underline{15.16}$$

b) er korrekt.

24) Fra modell 1:

β er forventet endring i utgifter dersom antall senger økes med 1.

10 flere senger gir forventet endring lik 10β

Siden $\hat{\beta} = 1412.1$ vil estimert endring med 10 ekstra senger være $10 \cdot 1412.1 = \underline{14121}$

e) er korrekt

25) 95% KI for β:

$\hat{\beta} \pm t_{0.025, 8} \cdot SE(\hat{\beta})$

$= 1412.1 \pm 2.306 \cdot 224.2$

$= [\,895.1\,,\ 1929.1\,]$

a) er korrekt.

26) Fra modell 2 estimatet

$\hat{y} = -719.3 + 1866.2 \cdot x$

for $x = 40$

$\hat{y} = -719.3 + 1866.2 \cdot 40 = 73928.7$

e) er korrekt.

27) Et 95 % PI for y

$$\hat{y} \pm t_{0.025,8} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \underbrace{\frac{(x-\bar{x})^2}{(s/SE(\hat{\beta}))^2}}_{\substack{\text{for } x=\bar{x} \\ \text{blir denne } = 0}}}$$

Vet at $\hat{y} = 102295$

$s = 30920$

$n = 10$

$t_{0.025,8} = 2.306$

Dermed :

$$102295 \pm 2.306 \cdot 30920 \cdot \sqrt{1 + \frac{1}{10}}$$

$= [27513 \quad , \quad 177077]$

b) er korrekt.

28) c) er korrekt

29) For en regresjonsmodell

$y_i = \alpha + \beta x_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma)$

vil $E(y|x) = \alpha + \beta x_i$

Altså er d) ikke antatt

d) er korrekt <u>svar</u>

30) Hvis sammenheng er positiv vil $\beta > 0$
Dermed er $H_1: \beta > 0$
og $H_0: \beta = 0$

a) er korrekt.

31) For modell 2, er det effekt av sykehussenger?
$H_0: \beta = 0$ $H_1: \beta \neq 0$
Testobservator:

$$T = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{1866.2}{527.7} = 3.536$$

e) er korrekt.

32) Bare alternativ d) sier noe om hvor godt modellen forklarer variasjon i responsen.
d) er korrekt

33) La X_{33} være antall som jobber med salg/service i varehandelen.
Har observert $X_{33} = 56$
Under H_0 om uavhengighet mellom næring og yrke vil denne ha forventning:

$$E(X_{33} \mid H_0 \text{ sann}) = E_{33} = \frac{R_3 \cdot K_3}{n} = \frac{79 \cdot 70}{182} = 30.4$$

a) er korrekt.

34) La q_{31} være bidraget.

$$q_{31} = \frac{(X_{31} - E_{31})^2}{E_{31}} = \frac{(13 - 27.8)^2}{27.8} = 7.9$$

c) er korrekt.

35) Testobservatoren Q er kji-kvadratfordelt ~~med~~ med $(r-1)(k-1) = (4-1)(3-1) = 6$ fr.grader.
Fra tabell over X_6^2 finner vi for $\alpha = 0.01$ at $X^2_{0.01,6} = 16.81$

d) er korrekt.

36) Fra tabell over kji-kvadrat komponentene finner vi summen av disse, $Q = 1.04$.
Under H_0 om uavhengighet er $Q \sim X_4^2$ og fra tabell finner vi at $X^2_{0.05,4} = 9.49$
Siden $Q < X^2_{0.05,4}$ beholdes H_0 om uavhengighet.

e) er korrekt.

37) Q er egentlig tilnærmet X^2-fordelt under H_0. Tilnærmingen blir dårlig dersom forventet antall i noen celler er lavt. (parallell til $np \geq 5$ & $n(1-p) \geq 5$ regelen i binomisk normaltilnærming)

b) er korrekt.

38) SS_E har $n-k$ frihetsgrader i ANOVA-modellene vi betrakter i STAT100.
Her er $n = 13 \cdot 3 = 39$ og $k = 3$
Dermed $39 - 3 = 36$ frihetsgrader
c) er korrekt.

39) $$SS_G = \sum_{i=1}^{3} n_i (\bar{Y}_i - \bar{Y})^2$$

$$= 13(29.47 - 27.83)^2 + 13 \cdot (30.86 - 27.83)^2 + 13(23.15 - 27.83)^2$$
$$= 439$$
a) er korrekt.

40) Figur øverst til venstre antyder at variasjonen er forskjellig i de tre gruppene.
Figur øverst til høyre gir ikke klar indikasjon om at residualenes fordeling ikke er normal.
Nederste figur viser at residualene har en trend over tid innen de tre gruppene.
Dette antyder at ε_i-ene ikke er uavhengige.
b) er mest korrekt uttalelse.