Løsningsforslag eksamen i STAT100 høst 2012

1) Forventningsrett estimator er $\hat{\mu} = \bar{x}$. Fra utvalget finner vi $\bar{x} = 60.4$
2) En-utvalgs t-test med ukjent, men estimert standardavvik.
   Hypoteser: $H_0 : \mu = 52$ mot $H_1 : \mu > 52$
   Testobservator:

$$T = \frac{\bar{x} - 52}{s/\sqrt{n}} = \frac{61.8 - 52}{11.6/\sqrt{16}} = 3.38$$

3) To-utvalgs test med felles populasjonsstandardavvik. Estimat:

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n1 + n2 - 2}} = \sqrt{\frac{15 \cdot 11.6^2 + 9 \cdot 8.2^2}{16 + 10 - 2}} = 10.5$$

4) Hvis forventet fart har økt i løpet av året vil E(Z) > E(Y) og dermed vil $\mu_D = E(D)$ bli negativ (alternativ hypotese). Ved forkastning er derfor alternativ b) være riktig svar.
5) SS_G har k-1 = 3-1=2 frihetsgrader
6) SS_T har alltid N-1 = 21-1 = 20 frihetsgrader
7) MSE = SSE/(N-k) = 229376/18 = 12743.11 ≈ 12743
8) Kan bruke at MSE er et felles estimat for variansen i alle grupper, dvs en S^2_{pooled} som er et veid gjennomsnitt av de tre gruppevariansene:

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{n_1 + n_2 + n_3 - 3}$$

   Som dersom vi har 7 observasjoner i hver gruppe, men samme varianser, blir lik:

$$= \frac{6 \cdot 130.6^2 + 6 \cdot 113.5^2 + 6 \cdot 80.8^2}{7 + 7 + 7 - 3} = 12155.75$$

9) Dette er p-verdien for testen.
10) Vi trenger fra t-tabellen: $t_{0.1/2,18} = t_{0.05,18} = 1.734$. Da blir intervallet

$$[\hat{\theta} \pm t_{0.05,18} SE(\hat{\theta})] = [34.17 \pm 1.734 \cdot 60.96] = [-71.5, 139.9]$$

11) Dette er en standard ANOVA-modell som vi har hatt i kurset som er gitt ved alternativ a)
12) $F = MS_G/MS_E = 344.6/36.4 = 9.47$
13) $R^2 = SS_G/SS_T = SS_G/(SS_G + SS_E) = 1034/(1045 + 582) = 0.64$
14) Den er et mål på hvor mye av variasjonen i spiringsprosent som kan forklares ved modellen.
15) $\hat{\theta} = (\bar{y}_2 + \bar{y}_3 + \bar{y}_4)/3 - \bar{y}_1 = (83 + 90 + 91)/3 - 73 = 15$

16) Merk at vi kan skrive: $\theta = \sum_{i=1}^{4} c_i\mu_i = (-1)\mu_1 + 1/3\mu_2 + 1/3\mu_3 + 1/3\mu_4$ som gir konstantene $c_1, \ldots c_4$.

$$SE(\hat{\theta}) = \sqrt{MS_E \sum_{i=1}^{2} c_i^2/n_i} = \sqrt{36.4((-1)^2/5 + 3 \cdot (1/3)^2/5)} = 3.12$$

17) Vi skal teste om det er generell forskjell, dvs tosidig alternativ hypotese. Hvis det ikke er forskjell vil $\theta = 0$. Riktig svar er alternativ a)

18)

$$\theta = \mu_3 - (\mu_2 + \mu_4)/2 = 0 \cdot \mu_1 + (-1/2) \cdot \mu_2 + 1 \cdot \mu_3 + (-1/2)\mu_4$$

Dvs alternativ c)

19) Som et mål på variasjon mellom målinger av responsen gjort innen samme nivå av gruppevariabelen.

20) Størrelsen er SS_T

21) I figur c) er det minst tegn til lineær sammenheng mellom X og Y, mao korrelasjon nær 0

22) Figur a)

23) Dersom alder øker med ett år er estimert endring i maxpuls lik $\hat{\beta} = -1.1405$. Ved 10 års økning vil dermed forventet endring i maxpuls være $10 \cdot \hat{\beta} = -11.405$.

24) For å teste $H_0 : \beta = 0$ mot $H_1 : \beta \neq 0$ bruker vi

$$T = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{-1.1405}{0.1612} = -7.075$$

25) I enkel lineær regresjon vil MS_E som er et estimat på støyvariansen, ha N-2 frihetsgrader, dvs her får vi 10-2 = 8 frihetsgrader. Siden MS_E inngår i beregningen av $SE(\hat{\beta})$ så vil også testobservatoren *T* funnet i oppgave 24 ha 8 frihetsgrader.

26) For å teste om $\beta = -1$ må vi bruke testobservatoren:

$$T = \frac{\hat{\beta} - (-1)}{SE(\hat{\beta})} = \frac{-0.1405}{0.1612} = -0.87$$

27) Testobservatoren blir den samme, men bruken av den blir annerledes, dvs hvilken tabellverdi den skal sammenliknes med.

28) $[\hat{\beta} \pm t_{0.025,8} SE(\hat{\beta})] = [-1.1405 \pm 2.306 \cdot 0.1612] = [-1.51, -0.77]$

29) Et KI for α er $[\hat{\alpha} \pm t_{0.025,8} SE(\hat{\alpha})]$. Det betyr f.eks at øvre grense er:
$222.27 + t_{\alpha/2,8} \cdot 6.0254 = 233.5$. Da finner vi at
$t_{\alpha/2,8} = (233.5 - 222.27)/6.0254 = 1.86$
Fra tabell D.5 over t-tabellen og i rad 8 finner vi at $\alpha/2 = 0.05$ og $\alpha = 0.10$.
Intervallet er derfor et 90% KI.

30) $\hat{y} = 222.27 - 1.1405 \cdot 30 = 188.1$

31) $e = y - \hat{y} = 196 - (222.27 - 1.1405 \cdot 24) = 196 - 194.9 = 1.10$

32) Dersom vi skal lage prediksjonsintervall for gjennomsnittsverdien av x forenkles formelen til:

$$\hat{y} \pm t_{\alpha/2,8} \cdot s\sqrt{1 + \frac{1}{n}} =$$

Der vi trenger å finne $\hat{y} = 222.27 - 1.1405 \cdot 36.8 = 180.3$. Det gir intervallet

$$[180.3 \pm 2.306 \cdot 3.326\sqrt{1 + 1/10}] = [172.3, 188.3]$$

33) Konfidensintervaller beregner vi kun for forventninger (og andre ukjente populasjonsparametre) og ikke for vilkårlige observasjoner av en tilfeldig variabel. Dermed er alternativ a)-c) utelukket. Videre er det snakk om forventningen til Y og ikke X her, så dermed må det bli alternativ d). Det er her altså snakk om forventningen til Y for en gitt verdi av X, dvs E(Y|X=x), og tolkningen i vårt tilfelle er altså: *Det er 99% sannsynlig at intervallet dekker den forventede maxpulsen til 25-åringer.*

34) I en toveistabell med *r* rader og *k* kolonner har *Q* (Kji-kvadrattest-observatoren) $(r-1)(k-1)$ frihetsgrader. Vi får dermed 6 frihetsgrader her.

35) Vi skal finne forventet antall i rad 2, kolonne 2, dvs E_{22}. Den finner vi ved:

$$E_{22} = \frac{R_2 K_2}{N} = \frac{347 \cdot 436}{1000} = 151.3$$

36) Det manglende tallet i tabellen over Kji-kvadrat-komponenter er:

$$q_{11} = \frac{(X_{11} - E_{11})^2}{E_{11}} = \frac{(60 - 62.449)^2}{62.449} = 0.10$$

37) Estimatet for β i en enkel lineær modell tolkes generelt som estimert endring i responsen dersom forklaringsvariabelen øker med én enhet. Alternativt kan vi si at den negative verdien er estimert endringen i y dersom X avtar med én enhet. I denne sammenhengen er dette lik: *For hver meter man kommer nærmere overflaten så er den estimerte, forventede økningen i E.coli-konsentrasjonen lik 1.98*

38) Tosidig testalternativ, 7 frihetsgrader, testnivå 1%. Mao vi må bruke $t_{0.01/2,7} = 3.499$

39) Denne estimatoren vi i det lange løp i gjennomsnitt gi riktig verdi β

40) Utfyllende svar:

Dersom vi definerer

$$W = \frac{1}{n} + \left(\frac{x - \bar{x}}{s/SE(\hat{\beta})} \right)^2$$

Så kan vi skrive formelen for et $(1-\alpha)\cdot 100\%$ konfidensintervall for forventnin $E(y|x)$ slik:

$$\hat{y} \pm t_{\alpha/2,n-2} s\sqrt{W}$$

Tilsvarende vil et $(1-\alpha)\cdot 100\%$ prediksjonsintervall for en ny observasjon y være

$$\hat{y} \pm t_{\alpha/2,n-2} s\sqrt{1+W} \qquad (1)$$

Vi har fått oppgitt et konfidensintervall med grenser $[a,b]$ = [195, 207.8]. Dette gir at den predikerte verdien av responsen må være (midtpunktet i intervallet):

$$\hat{y} = \frac{a+b}{2} = \frac{195 + 207.8}{2} = 201.4$$

Videre har vi fra utskriften at $s = 2.077$ og fra tabell med 7 frihetsgrader og $\alpha/2 = 0.025$ finner vi at $t_{0.025,7} = 2.365$. Dermed kjenner vi alle ledd bortsett fra W som vi må finne fra det kjente konfidensintervallet $[a,b]$. Siden vi lager intervallet ved å trekke fra og plusse på samme faktor $t_{\alpha/2,n-2} s\sqrt{W}$ fra $\hat{y}$, må det bety at bredden på intervallet ($b-a$) er lik 2 ganger denne faktoren, dvs

$$b - a = 2 \cdot t_{\alpha/2,n-2} s\sqrt{W}$$

Løser vi denne likningen mht på W får vi:

$$W = \left(\frac{b-a}{2 \cdot t_{\alpha/2,n-2} \cdot s} \right)^2$$

Dermed kan vi skrive prediksjonsintervallet (1) som funksjon av konfidensintervallgrensene *a* og *b* slik:

$$\frac{a+b}{2} \pm t_{\alpha/2,n-2} s\sqrt{1 + \left(\frac{b-a}{2 \cdot t_{\alpha/2,n-2} \cdot s} \right)^2}$$

Ved å sette inn det vi vet finner vi:

$$\frac{195 + 207.8}{2} \pm 2.365 \cdot 2.077\sqrt{1 + \left(\frac{207.8 - 195}{2 \cdot 2.365 \cdot 2.077} \right)^2}$$

$$201.4 \pm 8.068 = [193.3, 209.5]$$