

# Besvarelse til eksamen i Statistisk Dataanalyse 1, november 2012

## Opgave 1

- (a) Antallet af celler med  $H_2O_2$  for de to sorter betegnes hhv.  $X_A$  og  $X_B$ , som antages at være udfald fra to binomialfordelinger:  $\text{bin}(n_A, p_A)$  og  $\text{bin}(n_B, p_B)$ , hvor  $n_A = 1451$  og  $n_B = 1435$  er de to totalantal af celler.

Et 95% konfidensinterval for  $p_A$  (Stakado) er

$$\hat{p}_A \pm 1.96 \text{SE}(\hat{p}_A) = \hat{p}_A \pm 1.96 \sqrt{\hat{p}_A(1 - \hat{p}_A)/n_A} = 0.034 \pm 0.0093 = (0.024, 0.043).$$

Tilsvarende fås for sorten Sevin konfidensintervallet

$$0.019 \pm 0.0070 = (0.012, 0.026)$$

for  $p_B$ .

- (b) Vi udfører en homogenitetstest for hypotesen  $p_A = p_B$  svarende til at der ikke er forskel på hyppigheden af celler med  $H_2O_2$  for de to sorter.

Teststørrelsen hertil er chi-i-anden teststørrelsen som er

$$X^2 = \frac{(49 - 38.2)^2}{38.2} + \frac{(1402 - 1412.8)^2}{1412.8} + \frac{(27 - 37.8)^2}{37.8} + \frac{(1408 - 1397.2)^2}{1397.2} = 6.29$$

som under hypotesen er approksimativt  $\chi^2$ -fordelt med  $(2 - 1)(2 - 1) = 1$  frihedsgrad.

- (c) giver en P-værdi, som ifølge opgaveteksten skal antages at være

$$P = 0.012.$$

(Dette er faktisk den korrekte værdi, men de to alternative metoder, henholdsvis med kontinuitetskorrektion og med brug af Fishert's eksakte test, giver p-værdierne 0.017 og 0.014).

Heraf konkluderes at der med ret stor sikkerhed er forskel på hyppigheden af celler med  $H_2O_2$  hos de to sorter, idet forekomsten hos den resistente sort, Stakado, er klart signifikant hyppigere.

- (d) Odds-ratio for forekomst af  $H_2O_2$  på Stakado i forhold til Sevin er  $(p_A/(1 - p_A))/(p_B/(1 - p_B))$  som estimeres til

$$\hat{\text{OR}} = \frac{\hat{p}_A(1 - \hat{p}_B)}{\hat{p}_B(1 - \hat{p}_A)} = 1.82.$$

Konfidensintervallet beregnes først for  $\ln(\text{OR})$ , idet

$$\text{SE}(\ln(\text{OR})) = \sqrt{\frac{1}{49} + \frac{1}{1402} + \frac{1}{27} + \frac{1}{1408}} = 0.2426$$

hvorf 95% konfidensintervallet fås som

$$\ln(1.82) \pm 1.96 \cdot 0.2426 = \ln(1.82) \pm 0.476 = (0.125, 1.076).$$

Ved at tage eksponentialfunktionen til de to grænser får vi 95% konfidensintervallet for OR til  $(1.13, 2.93)$ .

## Opgave 2

- (a) Vi benytter modellen for lineær regression af  $y_4$  på  $y_0$ , dvs. at

$$y_{4i} = \alpha + \beta \cdot y_{0i} + e_i, \quad i = 1, \dots, 32,$$

hvor  $e_1, \dots, e_n$  antages uafhængige og normalfordelt med middelværdi nul og med samme spredning,  $\sigma$ .

Hypotesen om, at der ikke er nogen sammenhæng mellem  $y_0$  og  $y_4$  er  $\beta = 0$ , som testes med et  $t$ -test med teststørrelsen

$$t = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} = \frac{0.8218}{0.2681} = 3.065$$

som giver P-værdien 0.0046, jf. `summary(m1)` i R. Vi kan derfor konkludere, at der med stor sikkerhed er en sammenhæng mellem målingen før og efter måltidet, og at sammenhængen er positiv, således, at personer med relativt høj (lav) fastemåling fortrinsvist også vil være i den høje (lave) ende 4 timer efter måltidet. (Det kan endvidere bemærkes, at estimatet  $\hat{\beta} = 0.82$  viser, at den forventede værdi af  $y_4$  ikke helt “følger med”  $y_0$ , hvilket ville svare til hældning 1. Dette er et velkendt fænomen, som kaldes “regression towards the mean”).

- (b) For at undersøge, om der er nogen ændring fra før til efter måltidet, benytter vi de målte ændringer, differenserne

$$d_i = y_{4i} - y_{0i}, \quad i = 1, \dots, 32,$$

som antages uafhængige og normalfordelt med middelværdi  $\mu_d$  og spredning  $\sigma_d$ . Hypotesen  $\mu_d = 0$ , om ingen ændring, testes ved et parret  $t$ -test med teststørrelsen

$$t = \frac{\hat{\mu}_d}{\text{SE}(\hat{d})} = \sqrt{32} \frac{\bar{d}}{s_d} = 5.7789,$$

som ifølge R-udskriften fra `t.test(y0, y4, paired=TRUE)` giver P-værdien  $2.3 \cdot 10^{-6}$ . Dette viser med meget stor sikkerhed, at der gennemgående sker et fald i koncentrationen af frie fedtsyrer fra før til efter måltidet, idet middelværdien er mindre efter 4 timer end ved fastemålingen.

- (c) Vi antager, at  $y_{4_1}, \dots, y_{4_{32}}$  er uafhængige og normalfordelt med middelværdi  $\mu_4$  og spredning  $\sigma_4$ . Det interval, der spørges om, er et 95% prædiktionsinterval, som er givet ved

$$\hat{\mu}_4 \pm t_{0.975,31}s_4 \sqrt{1 + \frac{1}{32}} = 5.7513 \pm 2.0395 \cdot 0.5937 \sqrt{33/32} = 5.75 \pm 1.23$$

som er intervallet  $(4.52, 6.98)$ .

- (d) Her benytter vi den lineære regressionsmodel fra spørgsmål (a) og beregner et 95% prædiktionsinterval svarende til værdien  $y_0 = 6.0$ . Lad  $s_0$  betegne stikprøvespredningen for  $y_0$  og

$$SS_0 = \sum (y_{0_i} - \bar{y}_0)^2 = 31 \cdot s_0^2.$$

Intervallet er givet ved formlen

$$\hat{\alpha} + \hat{\beta} \cdot 6.0 \pm t_{0.975,30}s \sqrt{1 + \frac{1}{32} + \frac{(6.0 - \bar{y}_0)^2}{SS_0}}$$

som ved indsættelse af værdierne giver

$$= 5.52 \pm 2.042 \cdot 0.5266 \sqrt{1 + 1/32 + \frac{(6.0 - 6.284)^2}{31 \cdot 0.353^2}} = 5.52 \pm 1.10$$

som er intervallet  $(4.41, 6.62)$ , (hvis afrundinger i mellemregninger undgås).

### Opgave 3

- (a) Lad  $y_1, \dots, y_n$  betegne  $y_4$  målingerne for de  $n = 32$  forsøgspersoner,  $k_i$  deres køn (Male eller Female) og  $a_i$  deres aldersgruppe (A eller B). Der er tale om et forsøg med to faktorer, køn og aldersgruppe, med hver to niveauer. Som model benytter vi modellen for tofaktorforsøg med vekselvirkning:

$$y_i = \mu(k_i, a_i) + e_i,$$

hvor  $e_1, \dots, e_n$  antages uafhængige og normalfordelt med middelværdi nul og med samme spredning,  $\sigma$ . Samme model kan opskrives med opdeling af de fire mulige middelværdier  $\mu(k, a)$  i intercept samt hoved- og vekselvirkninger:

$$y_i = \mu + \alpha(k_i) + \beta(a_i) + \gamma(k_i, a_i) + e_i.$$

Modellen er kørt i R som `model7`, og udskriften fra `summary(model7)` viser estimatet  $s = 0.5162$  for spredningen  $\sigma$ .

- (b) Med udgangspunkt i ovenstående model tester vi først hypotesen,  $\gamma(k, a) = 0$  for alle værdier af  $k$  og  $a$ . Dette er hypotesen om ingen vekselvirkning, som også kan formuleres som, at modellen kan reduceres til den additive model

$$y_i = \mu + \alpha(k_i) + \beta(a_i) + e_i,$$

som er kørt som `model6` i R. Fra `summary(model7)` ser vi, at vekselvirkningen kun vedrører en enkelt parameter og derfor kan testes med et *t*-test som giver  $t = -0.315$  svarende til en  $P - vrdi$  på 0.75. Da denne er langt fra at være signifikant, tyder intet på vekselvirkning mellem aldersgruppe og køn. Modellen kan dermed reduceres til den additive `model6`.

Vi tester nu hypotesen  $\alpha(\text{male}) = \alpha(\text{female})$  svarende til reduktion til `model3`. Fra `summary(model6)` har vi *t*-teststørrelsen  $t = -1.886$  som giver en P-værdi på 0.069. Omend denne P-værdi ikke er konventionelt signifikant, er den ret lille, og vi noterer os derfor, at noget kunne tyde på en forskel på de to køn, omend det ikke er helt overbevisende.

På samme måde tester vi på basis af den additive `model6` hypotesen  $\beta(A) = \beta(B)$  om, at der ikke er forskel på de to aldersgrupper. Fra `summary(model6)` har vi *t*-teststørrelsen  $t = 3.124$  som giver en P-værdi på 0.0040 som med stor sikkerhed viser, at der er en forskel på koncentrationen af frie fedtsyre hos unge og ældre. Dette kunne også have været testet på baggrund af `summary(model3)`, hvis vi havde valgt at reducere modellen ved at tage effekten af køn ud. Det ville have givet  $t = 2.999$  og P-værdien 0.0054 med samme konklusion.

- (c) Konklusionen af analysen i (b) er, at koncentrationen af frie fedtsyrer bestemt afhænger af aldersgruppen og muligvis også af kønnet, omend det ikke er overbevisende. Vi angiver estimat og konfidensinterval for hver af disse to effekter på basis af den additive `model6`, men kunne også have valgt kun at angive effekt af aldersgruppen på basis af `model3`.

Estimat og konfidensinterval for  $\beta(B) - \beta(A)$ , som er forskellen “ældre - yngre”, er

$$0.561 \pm t_{0.975,29} \text{SE}(\hat{\beta}(B) - \hat{\beta}(A)) = 0.561 \pm 2.045 \cdot 0.1797 = 0.56 \pm 0.37,$$

som er intervallet (0.19, 0.93). Fra den reducerede `model3` ville vi have fået

$$0.561 \pm t_{0.975,30} \text{SE}(\hat{\beta}(B) - \hat{\beta}(A)) = 0.561 \pm 2.042 \cdot 0.1872 = 0.56 \pm 0.38,$$

svarende til intervallet (0.18, 0.94).

Estimat og konfidensinterval for  $\alpha(\text{male}) - \alpha(\text{female})$  er tilsvarende

$$-0.3388 \pm t_{0.975,29} \text{SE}(\hat{\alpha}(\text{male}) - \hat{\alpha}(\text{female})) = -0.34 \pm 2.045 \cdot 0.1797 = -0.34 \pm 0.37,$$

som er intervallet (-0.03, 0.71).

# Besvarelse til eksamen i Statistisk Dataanalyse 1, januar 2013

## Opgave 1

- (a) Modellen der beskriver den retlinede sammenhæng er

$$\text{Inv}_i = a + b \cdot \text{Ind}_i + e_i$$

hvor  $e_1, \dots, e_n$ ,  $n = 46$ , antages uafhængige og normalfordelt med middelværdi 0 og samme spredning  $\sigma$ . Modellen hedder `lm1` i R-udskriften, hvor vi fra `summary(lm1)` aflæser estimererne

$$\hat{a} = -8.62, \quad \hat{b} = 2.44, \quad s = 0.655$$

hvor  $s$  er estimatet for residualspredningen,  $\sigma$ , i modellen.

- (b) Fra samme tabel fra `summary(lm1)` ser vi, at standard error for estimatet for hældningen er  $\text{SE}(\hat{b}) = 0.292$ . I  $t$ -fordelingen med 44 frihedsgrader er 97.5% fraktilen 2.015, så 95% konfidensintervallet for hældningen,  $b$ , bliver

$$2.438 \pm 2.015 \cdot 0.292 = 2.44 \pm 0.59 = (1.85, 3.03).$$

- (c) Den første anova vedrører følgende tre modeller,

$$\begin{aligned}\text{lm0} : \text{Inv}_i &= b \cdot \text{Ind}_i + e_i \\ \text{lm1} : \text{Inv}_i &= a + b \cdot \text{Ind}_i + e_i \\ \text{lm0} : \text{Inv}_i &= a + b \cdot \text{Ind}_i + c \cdot \text{Ind}_i^2 + e_i\end{aligned}$$

hvor  $e_1, \dots, e_n$  i alle modeller antages uafhængige normalfordelte med middelværdi 0 og samme spredning,  $\sigma$ . De tre modeller beskriver logaritmen til dråbehastigheden som som funktion af logaritmen til dråbediameteren i form af henholdsvis en retlinie gennem punktet  $(0, 0)$ , en ret linie uden restriktioner, og et andengradspolynomium.

I det sidste af de to test er modellen et andengradspolynomium, og hypotesen er, at andengradsleddet er nul. Med en teststørrelse på 11.67 og tilhørende p-værdi på 0.0014 kan hypotesen klart forkastes: der er ikke tale om en retlinet sammenhæng.

I det første af de to test antages, at der er tale om en retlinet sammenhæng, og hypotesen er, at linien går gennem  $(0, 0)$ . P-værdien er her  $1.77 \cdot 10^{-7}$ , og hypotesen kan klart afvises med meget stor sikkerhed. Testet har ikke nogen relevans her, da førstnævnte test viste, at antagelsen om en retlinet sammenhæng ikke er rimelig.

I anden anova-funktionskald er de to modeller

$$\begin{aligned}\text{lm3} : \text{Inv}_i &= b \cdot \text{Ind}_i + c \cdot \text{Ind}_i^2 + e_i \\ \text{lm2} : \text{Inv}_i &= a + b \cdot \text{Ind}_i + c \cdot \text{Ind}_i^2 + e_i\end{aligned}$$

hvor modellen lm3 er et andengradspolynomium uden konstantled, dvs. en parabel som går gennem  $(0, 0)$ , mens lm2 som før er en parabel uden restriktioner. Med en p-værdi på 0.0002 ( $F = 10.5$ ) kan også denne (ikke særligt relevante hypotese) klart afvises.

Testene viser således, at den retlinede sammenhæng ikke passer med data, og at hvis en parabel beskriver sammenhængen, går den i hvert fald ikke gennem  $(0, 0)$ .

- (d) Figuren viser residualerne tegnet mod predikterede værdier, og den bruges til
- at validere den systematiske del af modellen (den rette linie),
  - at validere antagelsen om varianshomogenitet.

Første antagelse (om ret linie) er klart forkert, idet der er et "U-formet" mønster, som afgiver systematisk fra en middel-residual på nul, uanset den prediktede værdi. Antagelsen om varianshomogenitet ser heller ikke for godt ud, idet der tilsyneladende er større variation "lodret set" i højre side af figuren — altså stigende spredning med stigende middelværdi.

- (e) Modellen er

$$\ln v_i = a + \mu(\ln d_i) + e_i$$

hvor logaritmen til diameteren ved en fejltagelse nu benyttes som faktor. Da alle diametrerne formodentlig er forskellige, har faktoren et niveau, og dermed en middelværdi, for hver observation. Dermed kan modellen tilpasses perfekt (og komplet unyttigt) til data, og det er ikke muligt at estimere en spredning. Fænomenet omtales unertiden som "overfitting".

## Opgave 2

- (a) Det antages, at antallet af insekter, der er døde efter tre dage, er binomialfordelt med antal  $n = 56$  for begge stammer, mens sandsynligheden for, at insektet er dødt er henholdsvis  $p_A$  og  $p_B$  for de to stammer. Vi ønsker at teste hypotesen  $p_A = p_B$  svarende til, at dødeligheden er den samme for de to stammer. Dette gøres med et homogenitetstest i en  $2 \times 2$  antalstabell. Svarende til observationerne

$$\begin{pmatrix} 38 & 18 \\ 28 & 28 \end{pmatrix}$$

er de forventede værdier

$$\begin{pmatrix} 56 \cdot 66/112 & 56 \cdot 46/112 \\ 56 \cdot 66/112 & 56 \cdot 46/112 \end{pmatrix} = \begin{pmatrix} 33 & 23 \\ 33 & 23 \end{pmatrix}$$

Teststørrelsen bliver hermed

$$X^2 = \frac{(38 - 33)^2}{33} + \frac{(18 - 23)^2}{23} + \frac{(28 - 33)^2}{33} + \frac{(28 - 23)^2}{23} = 3.69$$

som under hypotesen er en observation fra en  $\chi^2$ -fordeling med 1 frihedsgrad. (P-værdien skulle ikke beregnes, men bliver 0.055.) Med en  $p$ -værdi på 0.08 konkluderes, at vi ikke overbevisende kan slutte, at der er forskel på de to dødeligheder.

Spørgsmålet kunne også besvares ved at beregne et konfidensinterval for forskellen,  $p_A - p_B$ . Et 95% konfidensinterval bliver

$$p_A - p_B : 0.18 \pm 0.18 = (0.0, 0.36)$$

hvilket viser at forskellen er lige på grænsen til at være signifikant på 5% niveauet.

- (b) Modellen for overlevelsandsynligheden,  $p(t)$ , hvor  $t$  er tiden i dage, er givet ved

$$\text{logit}(p(t)) = a + bt$$

Hvis  $p(t) = 0.5$ , svarende til, at halvdelen overlever, er

$$\text{logit}(p(t)) = \log(p/(1-p)) = 0,$$

så  $t = -a/b$ . Ud fra estimererne i udskriften får vi denne “halveringstid” til  $-6.6038/(-2.2278) = 2.96$  dage.

- (c) Logaritmen til odds for at overleve en dag ekstra vokser med hældningen, dvs. falder med 2.23 da hældningen er negativ. Odds ratio for at overleve i fire dage i forhold til to dage kan derfor estimeres til

$$\text{OR} = \exp(2 \cdot (-2.23)) = 0.012$$

ud fra den logistiske regressionsmodel.

- (d) Odds for at overleve i fire dage kan ud fra tallene estimeres til  $\text{odds}_4 = 13/99 = 0.13$ , mens odds for at overleve i to dage er  $\text{odds}_2 = 104/8 = 13$ . Odds ratio for at overleve i fire dage i forhold til to dage kan derfor estimeres til

$$\text{OR} = 13 \cdot 8 / (99 \cdot 104) = 0.010$$

i god overensstemmelse med det modelbaserede estimat.

- (e) Først og fremmest er antagelsen om uafhængighed ikke opfyldt, da det er de samme insekter, der indgår flere dage (så længe de overlever). Den anden antagelse, der kan sættes spørgsmålsteget ved, er antagelsen om, at sandsynligheden for at overleve følger en logistisk funktion ( $p = \exp(a + bt)/(1 + \exp(a + bt))$ ) for passende valgt  $a$  og  $b$ . Denne antagelse er svær at kontrollere med så få forskellige tider som her.

### Opgave 3

- (a) Et 95% “normalområde” for gevinsten er

$$10000 \pm 1.96 * 5000 = 10000 \pm 9800 = (200, 19800)$$

(b) Lad  $G_A$  betegne gevinsten. Sandsynligheden for en negativ gevinst er

$$P(G_A < 0) = P\left(\frac{G_A - 10000}{5000} < \frac{0 - 10000}{5000}\right) = \Phi(-2) = 1 - \Phi(2) = 1 - 0.977 = 0.023$$

hvori  $\Phi$  betegner fordelingsfunktionen for en standard normalfordeling.

(c) Med tilsvarende betegnelser som ovenfor får vi

$$\mathbb{E}(G_B - G_A) = 15000 - 10000 = 5000$$

og

$$\text{var}(G_B - G_A) = \text{var}(G_B) + \text{var}(G_A) = 5000^2 + 9000^2$$

og dermed

$$P(G_B - G_A > 0) = 1 - \Phi\left(\frac{-5000}{\sqrt{5000^2 + 9000^2}}\right) = 0.686.$$

- (d) Spredningen må være op til 10000, da beregningen fra spørgsmål (b) viser, at det er middelværdien divideret med spredningen, der afgør hvad tabsrisikoen bliver.
- (e) Vi antager, at der er tale om 10 uafhængige observationer fra en normalfordeling med middelværdi  $\mu$  og spredning  $\sigma$ . Estimaterne er

$$\hat{\mu} = 4868.5, \quad s = 3408.9,$$

så et 95% konfidensinterval for  $\mu$  bliver

$$\hat{\mu} \pm t_{9,0.975} \cdot \frac{s}{\sqrt{10}} = 4868.5 \pm 2.262 \cdot 1078.0 = 4869 \pm 2439$$

som bliver intervallet (2430, 7307).

# Besvarelse til eksamen i Statistisk Dataanalyse 1, november 2013

## Opgave 1

- (a) Beregnes række- og søjlemarginaler i  $2 \times 2$ -tabellen fås

Fornavn	Efternavn		Total
	Jensen	Hansen	
Jens	3752	1997	5749
Hans	2160	2807	4967
Total	5912	4804	10716

Vi kan derefter bruge formlen nederst på side 310 i lærebogen, som siger at det forventede antal (under antagelsen om uafhængighed mellem for- og eternavn) kan findes som

$$\text{rækemarginal} * \text{søjlemarginal} / \text{total}.$$

F.eks. er det forventede antal danskere med navnet *Jens Jensen* givet ved

$$\frac{5749 * 5912}{10716} = 3171.714$$

Tilsvarende beregnes de andre tre forventede antal, så vi får de forventede antal vist i tabellen nedenfor:

Fornavn	Efternavn	
	Jensen	Hansen
Jens	3171.714	2577.286
Hans	2740.286	2226.714

- (b) Vi tester hypotesen om uafhængighed i en 2 gange 2 antalstabell. Idet vi er i en  $2 \times 2$  tabel kan teststørrelsen findes ved

$$X^2 = \frac{10716 * (3752 * 2807 - 2160 * 1997)^2}{5749 * 4967 * 5912 * 4804} = 510.9263$$

Idet "tommelfingerreglen" at alle forventede værdier er større end 5, klart er opfyldt (se spørgsmål (a)), så gælder at  $X^2$  er approksimativt  $\chi^2$ -fordeling med  $(2 - 1) * (2 - 1) = 1$  frihedsgrader under hypotesen om uafhængighed mellem for- og eternavn. Idet  $X^2$  er langt større end 0.9999 fraktilen i  $\chi^2$ -fordelingen med 1 frihedsgrad fås, at p-værdien er meget mindre end 0.0001. Vi konkluderer, at der ikke er uafhængighed mellem for- og eternavn. Sammenlignes det observerede antal med det forventede antal så ses, at navnene *Jens* og *Hans* vælges mere hyppigt familier med det tilsvarende eternavn.

- (c) Estimatet for odds ratio for at hedde *Jens* frem for *Hans*, hvis man hedder *Jensen* i forhold til at man hedder *Hansen*, estimeres ved

$$\widehat{OR} = \frac{3752 * 2807}{2160 * 1997} = 2.441594$$

som kan afrundes til 2.44. Standard error for estimatet af den naturlige logaritme af odds ratio beregnes udfra formel (12.4):

$$SE(\log(\widehat{OR})) = \sqrt{\frac{1}{3752} + \frac{1}{2160} + \frac{1}{1997} + \frac{1}{2807}} = 0.03983078$$

Et approksimativt 95% konfidensinterval for  $\log(OR)$  er dermed givet ved

$$\log(2.4415) \pm 1.96 \cdot 0.03983 = (0.8145; 0.9706)$$

Ved at tage eksponentialfunktionen til de to grænser får vi 95% konfidensintervallet for OR til  $(2.2581; 2.6397)$ , som kan afrundes til  $(2.26, 2.64)$ .

## Opgave 2

- (a) For at undersøge om der er forskel på de to målemetoder målt på de *samme* grise, benyttes et *parret t-test*. Dette svarer til at lave et *enkelt stikprøve t-test* på differenserne

$$D_i = X_i - Y_i, \quad i = 1, \dots, 20,$$

hvor  $X_i$  og  $Y_i$  betegner målingerne med metode A henholdsvis B. Differenserne antages uafhængige og normalfordelt med middelværdi  $\mu_D$  og spredning  $\sigma_D$ . Nulhypotesen at de to målemetoder er ens, er givet ved  $\mu_D = 0$  og testes via teststørrelsen

$$t = \frac{\hat{\mu}_D}{SE(\hat{\mu}_D)} = \sqrt{20} \frac{\bar{D}}{s_D} = \sqrt{20} \frac{9.64}{7.292} = 5.912$$

Under nulhypotesen er denne teststørrelse *t*-fordelt med  $20 - 1 = 19$  frihedsgrader. I denne fordeling er 0.9999 fraktilen 4.590, og da den observerede teststørrelse er endnu større, er p-værdien mindre end  $2 * 0.0001 = 0.0002$ . Dette viser med meget stor sikkerhed, at der er en systematisk forskel på de to målemetoder i retning af, at A giver højere målinger.

- (b) Et 95% konfidensinterval for middelforskellen,  $\mu_D$ , på de to målemetoder er

$$\hat{\mu}_D \pm t_{0.975, df=19} * SE(\hat{\mu}_D) = 9.64 \pm 2.093 * 7.292 / \sqrt{20} = 9.64 \pm 3.41$$

som er intervallet  $(6.23, 13.05)$ .

- (c) Her er tale om et 95% prædiktionsinterval for en ny differens  $X - Y$ . Intervallet er

$$\hat{\mu}_D \pm t_{0.975, df=19} * 7.292 * \sqrt{1 + \frac{1}{20}} = 9.64 \pm 15.64$$

som er intervallet  $(-6.00, 25.28)$ .

- (d) Der spørges efter sandsynligheden for, at en ny differens  $X - Y$  er negativ. Differenserne er antaget normalfordelt med middelværdi og spredning, som er estimeret til  $\hat{\mu}_D = 9.64$  og  $s_D = 7.292$ . Derved fås

$$P(D < 0) = \Phi\left(\frac{-9.64}{7.292}\right) = \Phi(-1.32) = 0.093$$

Her er  $\Phi$  fordelingsfunktionen for standardnormalfordelingen, og sandsynligheden er fundet ved opslag i appendix C.2 i lærebogen.

- (e) Hændelsen, at differensen er negativ, er indtruffet en gang ud af 20. Da differenserne antages uafhængige, er antallet af negative differenser binomialfordelt med antalsparameter  $n = 20$  og sandsynlighedsparameter  $p$ . Estimatet for  $p$  er  $\hat{p} = 1/20 = 0.05$ , og et approksimativt konfidensinterval for  $p$  bliver

$$0.05 \pm 1.96 \sqrt{\frac{0.05 * 0.95}{20}} = 0.05 \pm 0.0955$$

som er intervallet  $(-0.0455, 0.1455)$ . Idet en sandsynlighed ikke kan være negativ afrundes dette interval til  $(0, 0.146)$ . Alternativt benyttes den mere nøjagtige metode beskrevet i afsnit 10.3.1 i lærebogen (“add two failures and two successes”), hvorfra vi i stedet får

$$\frac{3}{24} \pm 1.96 \sqrt{\frac{(3/24) * (21/24)}{24}} = 0.125 \pm 0.1323$$

Dette er intervallet  $(-0.0073, 0.2573)$ , som afrundes til  $(0, 0.26)$ .

### Opgave 3

- (a) Lad  $L_i$  og  $K_i$  for  $i = 1, \dots, 8$  betegne henholdsvis længde og koncentration for det  $i$ te kar. Som model benytter vi den lineære regressionsmodel

$$L_i = a + b \cdot K_i + e_i,$$

hvor  $e_1, \dots, e_8$  antages uafhængige og normalfordelte  $N(0, \sigma^2)$ . Parametrene er  $a$ ,  $b$  og  $\sigma$ . Fra R output for **summary(m1)** aflæses estimaterne

$$\hat{a} = 9.8949, \quad \hat{b} = -0.2193, \quad s = 0.1602$$

Parameteren  $b$  (hældningen) udtrykker gennemsnitsændringen af kronrørslængden (mm) per enhed øget koncentration, og kvantificerer dermed om sprøjtingen påvirker kronrørslængden. Et 95% konfidensintervallet for denne hældning er

$$\hat{b} \pm t_{0.975, df=6} * SE(\hat{b}) = -0.219 \pm 2.447 \cdot 0.0251 = -0.219 \pm 0.0614$$

som er intervallet  $(-0.2804, -0.1575)$ .

- (b) Af de fem udsagn er kun nummer 3 korrekt.

- (c) Den estimerede formel for den forventede kronrørslængde (i mm) er  $9.895 - 0.2193 * K$ . Denne formel giver 9 mm, hvis koncentrationen er

$$K = \frac{9 - 9.895}{-0.2193} = 4.08$$

- (d) Hvis parametrene i modellen er  $a$ ,  $b$  og  $\sigma$  som ovenfor, er middelværdien af kronrørslængden ved koncentration  $K$  udtrykt ved formlen

$$a + b * K$$

mens spredningen er  $\sigma$ . Derfor er sandsynligheden for, at kronrørslængden er mindre end 9 mm

$$P(L \leq 9) = \Phi\left(\frac{9 - (a + b * K)}{\sigma}\right)$$

hvor  $\Phi$  er fordelingsfunktionen for standardnormalfordelingen. Hvis denne skal være 0.95, skal vi have

$$\frac{9 - (a + b * K)}{\sigma} = 1.645$$

og dermed

$$K = \frac{(9 - a - \sigma * 1.645)}{b}$$

Ved indsættelse af de tre estimerer fås koncentrationen

$$K = \frac{9 - 9.895 - 0.1602 * 1.645}{-0.2193} = 5.28$$

som altså er vores estimat for den ønskede koncentration.

- (e) En mulighed er, at benytte en multipel lineær regressionsmodel, som kunne udføres med følgende programlinier:

```
model1 <- lm(lgd ~ konc1 + konc2)
residualplot(model1)
qqnorm(rstandard(model1))
drop1(model1,test="F")
summary(model1)
```

Besvarelse til reeksamen i Statistisk Dataanalyse 1  
 Januar 2014

## Opgave 1

- (a) Lad  $y_1, \dots, y_n$  betegne tørstofmålingerne for de  $n = 36$  kar,  $a_i$  elefantgræssets alder (1, 2 eller 3 år), og  $j_i$  jordtypen i karret (humusjord, lerjord eller sandjord). Der er tale om et forsøg med to faktorer, jordtype og alder, med hver tre niveauer. Som model benytter vi modellen for tosiddet variansanalyse med vekselvirkning:

$$y_i = \mu(a_i, j_i) + e_i,$$

hvor  $e_1, \dots, e_n$  antages uafhængige og normalfordelt med middelværdi nul og med samme spredning,  $\sigma$ . Modellen er kørt i R som `model.aj2`.

- (b) De korrekte antagelser er nummer 3 og 6.  
 (c) Med udgangspunkt i ovenstående model tester vi først hypotesen om ingen vekselvirkning, altså om modellen kan reduceres til den additive `model.aj1`,

$$y_i = \mu + \alpha(a_i) + \beta(j_i) + e_i$$

Fra `anova(model.aj1, model.aj2)` i R får vi F-teststørrelsen på 0.50 og en  $P$ -værdi på 0.74. Dermed er der intet, der tyder på vekselvirkning mellem alder og jordtype. Modellen kan derfor reduceres til den additive `model.aj1`.

Vi tester nu hypotesen  $\alpha(1) = \alpha(2) = \alpha(3)$  svarende til reduktion til `model.j`. Fra `anova(model.j, model.aj1)` har vi teststørrelsen  $F = 9.05$  som giver en  $P$ -værdi på 0.00008. Der er altså klart en effekt af alder, som ikke kan udelades af modellen.

På samme måde tester vi på basis af den additive `model.aj1` hypotesen

$$\beta(\text{humusjord}) = \beta(\text{lerjord}) = \beta(\text{sandjord})$$

om, at der ikke er forskel på de tre jordtyper. Fra `anova(model.a, model.aj1)` har vi teststørrelsen  $F = 26.4$  som giver en  $P$ -værdi på 0.0000002. Der er altså også klart en effekt af jordtypen. Slutmodellen bliver dermed den additive `model.aj1`.

- (d) Der er med stor sikkerhed vist effekt af både alder og jordtype, mens intet tyder på vekselvirkning mellem dem.

Estimater og konfidensintervaller angives for effekten af alder og af jordtype. Da der ikke er vekselvirkning, kan estimaterne angives for de to faktorer hver for sig. Vi vælger at præsentere estimaterne med alder 1 år og med humusjord som reference. Ved brug af `summary(model.aj1)` får vi

Parameter	Estimat	Konfidensinterval (95%)
$\alpha(2) - \alpha(1)$	2.42	est $\pm 3.36$
$\alpha(3) - \alpha(1)$	6.90	est $\pm 3.36$
$\beta(\text{lerjord}) - \beta(\text{humusjord})$	-2.18	est $\pm 3.36$
$\beta(\text{sandjord}) - \beta(\text{humusjord})$	9.11	est $\pm 3.36$

idet standard error for hver af forskellene er den samme, nemlig 1.646 som multipliceret med 0.975-fraktilen i  $t$ -fordelingen med 31 frihedsgrader giver konfidensintervallet

$$\text{estimat} \pm 1.646 \cdot 2.040 = \text{estimat} \pm 3.36$$

Det ses altså, at tørstofmængden er større ved højere alder af elefantgræsset, og højere ved sandjord end ved de to andre jordtyper.

- (e) Ligesom i forrige spørgsmål angiver vi estimateet for forskellen på tørstofvægt ved de to aldre (2 år og 3 år) fælles for de tre jordtyper ud fra den additive model.aj1. Fra summary for denne model får vi estimat og konfidensinterval til

$$\hat{\alpha}(3) - \hat{\alpha}(2) = 6.90 - 2.42 = 4.48,$$

med 95% konfidensgrænserne  $4.48 \pm 3.36$ .

## Opgave 2

- (a) Modellen er den lineære regressionsmodel:

$$Y_i = a + bx_i + e_i$$

hvor  $Y_i$  og  $x_i$  betegner henholdsvis den relative bakterieaktivitet og fenolkoncentrationen for den  $i$ te prøve, og hvor  $e_i$ ’erne ( $i = 1, \dots, 12$ ) er indbyrdes uafhængige og  $N(0, \sigma^2)$ . Dermed er antagelserne

1. Den relative bakterieaktivitet er en retlinet funktion af fenolkoncentrationen.
2. Der er uafhængighed mellem prøvernes afvigelser fra denne linie.
3. Afgangelserne er normalfordelt
4. med en spredning som ikke afhænger af fenolkoncentrationen.

- (b) Estimaterne for modellens tre parametre ses i R-udskriften:

$$\hat{a} = 1.098, \hat{b} = -0.0517, \hat{\sigma} = 0.0622.$$

Estimatet for den koncentration,  $x_0$ , der giver en relativ bakterieaktivitet på 0, er givet ved

$$\hat{a} + \hat{b} \cdot x_0 = 0$$

som giver  $x_0 = -1.098 / (-0.0517) = 21.2$ . Tilsvarende får vi

$$\hat{a} + \hat{b} \cdot x_{0.5} = 0.5$$

som giver

$$x_{0.5} = (1.098 - 0.5) / 0.0517 = 11.6$$

- (c) Vi ønsker at undersøge, om  $a$  kan være 1 og beregner et 95%-konfidensinterval for  $a$ . Konfidensintervallet for  $a$  er givet ved

$$\hat{a} \pm t_{10,0.975} \text{SE}(\hat{a}) = 1.098 \pm 0.044 \cdot 2.228 = 1.098 \pm 0.098$$

som lige akkurat grænser op til den formodede værdi  $a = 1$ . Der kunne således være noget der tyder på, at modellen ikke er ideel for disse data.

Spørgsmålet kunne også besvares ved at teste hypotesen  $a = 1$  ved et  $t$ -test med

$$t = \frac{\hat{a} - 1}{\text{SE}(\hat{a})} = \frac{0.098}{0.044} = 2.228,$$

som ved et tosided test i en  $t$ -fordeling med 10 frihedsgrader giver en p-værdi på 0.05, som er lige på kanten af det konventionelt signifikante.

### Opgave 3

- (a) Det antages, at antallet,  $X_i$ , af syge køer i den  $i$ te besætning er binomialfordelt med sandsynlighedspараметer  $p_i$  og antalsparameter  $n_i$ , som er det samelede antal køer i besætningen. Vi ønsker at teste hypotesen  $p_1 = \dots = p_6$  svarende til at sygeligheden er den samme i alle besætninger pånær den tilfældige (binomialfordelte) variation. Det kan gøres ved et homogenitetstest ( $\chi^2$ -test) i en 2 gange 6 tabel, hvor de to rækker repræsenterer henholdsvis de syge og de raske køer. Teststørrelsen er

$$\chi^2 = \sum_i \left( \frac{(X_i - E_i)^2}{E_i} + \frac{((n_i - X_i) - (n_i - E_i))^2}{n_i - E_i} \right)$$

hvor  $E_i = n_i \cdot \sum X_i / N$ , og  $N$  er det totale antal køer. Under hypotesen er teststørrelsen approksimativt  $\chi^2$ -fordelt med  $(2 - 1)(6 - 1) = 5$  frihedsgrader.

Testet kan fås fra R med linierne

```
cows = matrix(c(32,12,32,11,22,21,38,13,36,1,36,8), nrow=2, ncol=6)
chisq.test(cows)
```

- (b) Under den givne antagelse kan vi slå besætningerne sammen og finde  $X = 66$  syge køer ud af  $N = 262$ . Det giver estimatet

$$\hat{p} = \frac{66}{262} = 0.252$$

for, at en ko er syg. Et (approksimativt) konfidensinterval for  $p$  bliver

$$0.252 \pm 1.96 \sqrt{\frac{0.252 \cdot (1 - 0.252)}{262}} = 0.252 \pm 0.053$$

som er intervallet  $(0.20, 0.30)$ .

- (c) Sandsynligheden for at alle 20 køer er raske er

$$(1 - 0.15)^{20} = 0.039$$

idet vi antager, at der er uafhængighed mellem køerne.

- (d) Udfaldsrummet er mængden af de 66 syge køer, og sandsynlighedsfordelingen er en ligefordeling, således at hver af køerne har sandsynligheden  $1/66$  for at blive udvalgt. Hændelsen at den udtrukne ko kommer fra besætning 1 består af 12 udfald og i ligefordelingen på de 66 udfald har denne hændelse derfor sandsynlighed  $12/66 = 0.182$ .

# Besvarelse til eksamen i Statistisk Dataanalyse 1, november 2014

## Opgave 1

- (a) Lad  $y_1, \dots, y_5$  angive målingerne af forskellene for de 5 spillere med fejlstilling i venstre side, og lad  $x_1, \dots, x_{10}$  angive målingerne af forskellene for de 10 spillere uden fejlstilling i bækkenet. Fra opgaveteksten vides, at vi kan bruge den statistiske model for to uafhængige stikprøver med ens varianser:

$$Y_1, \dots, Y_5 \text{ er uafhængige } N(\mu_Y, \sigma^2), \\ X_1, \dots, X_{10} \text{ er uafhængige } N(\mu_X, \sigma^2).$$

Vi ønsker at teste nul hypotesen  $H_0 : \mu_Y = \mu_X$ , som siger at forskellen mellem fleksibiliteten i venstre og højre baglår er den samme i de to populationer. Den sammenvejede stikprøvespredning er

$$s = \sqrt{\frac{(5-1) \cdot s_y^2 + (10-1) \cdot s_x^2}{5+10-2}} \\ = \sqrt{\frac{4 \cdot 1.30384^2 + 9 \cdot 1.197219^2}{13}} \\ = 1.231009,$$

hvorefter vi finder teststørrelsen

$$t_{\text{obs}} = \frac{\bar{y} - \bar{x}}{s \sqrt{\frac{1}{5} + \frac{1}{10}}} = \frac{-10.3 - 0.6}{1.231009 \sqrt{\frac{3}{10}}} = -16.1661,$$

som er t-fordelt med  $5 + 10 - 2 = 13$  frihedsgrader under nul hypotesen. Det ses umiddelbart at testet er høj signifikant, hvormed nul hypotesen klart afvises. Udfra estimatorne konkluderes at  $\mu_Y < \mu_X$ , hvilket betyder at spillerne med fejlstilling i venstre side af bækkenet har lavere fleksibilitet i venstre baglår ift. højre baglår, end hvad der findes for spillere uden fejlstilling.

- (b) Vi bruger modellen beskrevet i (a) til at lave et 95%-konfidensinterval for  $\mu_X$ . Hvis analysen fortsættes udfra (a), så er det naturligt at bruge den sammenvejede stikprøvespredning. Gøres dette får konfidensintervallet

$$\bar{x} \pm t_{0.975, df=13} \cdot \frac{s}{\sqrt{10}} = 0.6 \pm 2.160 \cdot \frac{1.231009}{\sqrt{10}} = [-0.2408 ; 1.4408].$$

Alternativt kan man vælge at bruge stikprøvespredningen fra den enkelte stikprøve  $x_1, \dots, x_{10}$ , hvormed man får konfidensintervallet

$$\bar{x} \pm t_{0.975, df=9} \cdot \frac{s_x}{\sqrt{10}} = 0.6 \pm 2.262 \cdot \frac{1.197219}{\sqrt{10}} = [-0.2564 ; 1.4564].$$

Uanset hvilken model der benyttes ses, at 0 ligger i det tilhørende 95% konfidensinterval for  $\mu_X$ . Specielt betyder dette, at T-testet for nul hypotesen  $H_0 : \mu_X = 0$  ikke afvises på et 5% signifikansniveau. Vi kan altså antage at venstre og højre baglår er lige fleksible i populationen af spillere uden fejlstilling i bækkenet.

- (c) Idet der spørges til fordelingen af forskellen for en ny spiller laves et 95% prædiktionsinterval. Hvis man bruger den sammenvejede stikprøvespredning fås intervallet

$$\bar{x} \pm t_{0.975, df=13} \cdot s = 0.6 \pm 2.160 \cdot 1.231009 \cdot \sqrt{1 + \frac{1}{10}} = [-2.1888 ; 3.3888],$$

og hvis man betragter  $x_1, \dots, x_{10}$  som en enkelt stikprøve fås intervallet

$$\bar{x} \pm t_{0.975, df=9} \cdot s = 0.6 \pm 2.262 \cdot 1.197219 \cdot \sqrt{1 + \frac{1}{10}} = [-2.2403 ; 3.4403].$$

Uanset hvilken model der benyttes ses, at forskellen på  $-3 = 91 - 94$  ligger udenfor 95% prædiktionsintervallet. Udfra denne betragtning er den nye spiller dermed usædvanlig.

- (d) Lad  $y_1, \dots, y_5$  angive målingerne af forskellene for de 5 spillere med fejlstilling i venstre side, og lad  $z_1, z_2$  angive målingerne af forskellene for de 2 spillere med fejlstilling i højre side af bækkenet. Fra opgaveteksten vides, at vi kan bruge modellen for 2 uafhængige stikprøver med samme varians:

$$Y_1, \dots, Y_5 \text{ er uafhængige } N(\mu_Y, \sigma^2), \\ Z_1, Z_2 \text{ er uafhængige } N(\mu_Z, \sigma^2).$$

Udsagnet at effekten af fejlstilling ikke afhænger af hvilken side der er berørt kan formuleres som nul hypotesen  $H_0 : \mu_Y = -\mu_Z$ . Responsvariablen jo er givet som forskellen mellem fleksibiliteten i venstre og højre baglår, hvormed fortegnet skal ændres for de spillere der har fejlstillingen i højre side. Vi laver et T-test af denne hypotese, og starter med at finde den sammenvejede stikprøvespredning:

$$s = \sqrt{\frac{(5-1) \cdot s_y^2 + (2-1) \cdot s_z^2}{5+2-2}} \\ = \sqrt{\frac{4 \cdot 1.30384^2 + 1.414214^2}{5}} \\ = 1.32665.$$

Derefter beregnes teststørrelsen

$$t_{\text{obs}} = \frac{\bar{y} - (-\bar{z})}{s \sqrt{\frac{1}{5} + \frac{1}{2}}} = \frac{-10.3 + 10.5}{1.32665 \sqrt{\frac{7}{10}}} = 0.1802$$

som er t-fordelt med  $5+2-2 = 5$  frihedsgrader under nul hypotesen. Det ses umiddelbart at p-værdien er større end 5%. Den præcise p-værdi kan beregnes via R:

```
> 2*(1-pt(0.1802,df=5))
[1] 0.8640714
```

Vi accepterer altså nul hypotesen, at effekten af fejlstillingen er den sammen uanset om fejlstillingen er i venstre eller højre side af bækkenet.

- (e) I spørgsmål (d) fandt vi, at effekten af fejlstillingen ikke afhæng af hvilken side der havde fejlstilling. Lad nu  $u_1 = y_1, \dots, u_5 = y_5, u_6 = -z_1, u_7 = -z_2$  angiver forskellen mellem fleksibiliteten i det berørte ben og det andet ben for de  $5+2=7$  spillere med fejlstilling i bækket, og  $x_1, \dots, x_{10}$  angive forskellen mellem fleksibiliteten af venstre og højre ben for spillere uden fejlstilling som i spørgsmål (a). Vi betragter dette som 2 uafhængige stikprøver med samme varians, og bruger altså den statistiske model

$$U_1, \dots, U_7 \text{ er uafhængige } N(\mu_U, \sigma^2), \\ X_1, \dots, X_{10} \text{ er uafhængige } N(\mu_X, \sigma^2).$$

Vi ønsker at teste nul hypotesen at der ikke er effekt af fejlstillingen, altså  $H_0 : \mu_U = \mu_X$ . R koden

```
c(venstre,-højre)
```

laver netop vektoren  $u_1, \dots, u_7$ , hvormed kaldet til **t.test()** laver det ønskede T-test. Ligesom i (a) er testet stærkt signifikant ( $p=0.0000000001003$ ), og ud fra estimatorne konkluderes at fejlstillingen reducerer fleksibiliteten af baglåret i den berørte side.

## Opgave 2

- (a) Der er fundet 7 mænd med fejlstilling i bækkenet. Lad  $Y$  angive antallet af disse mænd, hvor fejlstillingen er i venstre side, og lad  $p$  angive proportionen af mænd som har fejlstillingen i venstre side i populationen af mænd med fejlstilling. Idet mændene antages at være uafhængige af hinanden er  $Y$  binomial fordelt med antalsparameter  $n = 7$  og sandsynlighedsparameter  $p$ . Dette er vores statistiske model. Den videnskabelige hypotesen at fejlstilling forekommer lige hyppigt i venstre og højre side kan formuleres som nul hypotesen  $H_0 : p = \frac{1}{2}$ . Vi tester dette med et binomial test, som giver

$$\begin{aligned} \text{p-værdi} &= \frac{1}{2^{7-1}} \sum_{y=0}^{\min\{5,7-5\}} \frac{7!}{y!(7-y)!} = \frac{1}{64} \sum_{y=0}^2 \frac{7!}{y!(7-y)!} \\ &= \frac{1 + 7 + \frac{7 \cdot 6}{2}}{64} = \frac{29}{64}. \end{aligned}$$

Stikprøven giver dermed ikke anledning til at tro at der forskel på hyppigheden af fejlstilling i henholdsvis venstre og højre side.

En alternativ (omend ikke helt så god) besvarelse fås ved at lave et 95% konfidensinterval for  $p$ . Bruges formlen for det forbedrede 95% konfidensinterval fås

$$\frac{7}{11} \pm 1.96 \cdot \sqrt{\frac{7 \cdot 4}{11^3}} = [0.3521 ; 0.9206].$$

Idet  $\frac{1}{2}$  ligger i dette interval fås samme konklusion som angivet ovenfor.

- (b) Der er 7 ud af 85 mænd i stikprøven som har fejlstilling i bækkenet. Vi bruger en binomial model med kendt antalsparameter  $n = 85$  og ukendt sandsynlighedspараметer  $p$ . Parameteren  $p$  beskriver andelen af mænd med fejlstilling i bækkenet, og det forbedrede 95% konfidensinterval for  $p$  er givet ved

$$\frac{9}{89} \pm 1.96 \cdot \sqrt{\frac{9 \cdot 80}{89^3}} = [0.0385 ; 0.1638].$$

### Opgave 3

- (a) Den beskrevede model er en ANCOVA (Analysis of COVariance) med responsvariabel `log(volumen)` og forklarende variable `vand` og `varighed`. R kode til estimation og validering af denne model kunne være:

```
# estimation
m1 <- lm(log(volumen) ~ vand + varighed + vand:varighed, data=baobab)
summary(m1)

# validering: residualplot
plot(predict(m1), rstandard(m1))
abline(0,0)

# validering: qq-plot
qqnorm(rstandard(m1))
abline(0,1)
```

- (b) Parametrene hørende til den kategoriske variable `vand` angiver forskellene i log-volumen mellem vettingsgrupperne til tid `varighed=0`, altså før vettingsbehandlingen er startet. I termer af parametrene i `m1` kan hypotesen, at der ikke er forskel på vettingsgrupperne før vettingsbehandlingen starter dermed formuleres som nul hypotesen:

$$H_0 : \text{vand50\%}=0 \text{ og } \text{vand75\%}=0$$

I R output findes p-værdierne for T-tests af de to separate nul hypoteser

$$H_{0a} : \text{vand50\%}=0 \text{ (T-test giver } p=0.615), \\ H_{0b} : \text{vand75\%}=0 \text{ (T-test giver } p=0.699).$$

Idet begge de separate nul hypotesen har meget store p-værdier konkluderes, at data er i overensstemmelse med  $H_0$ .

- (c) Parameteren  $\beta_{\text{vand}=50\%} - \beta_{\text{vand}=100\%}$ , der angiver forskellen mellem hældningen mht. den kontinuerte variable **varighed** for vandingsgruppen **vand=50%** i forhold til reference vandingsgruppen **vand=100%**, hedder **varighed:vand50%** i modellen **m2**. Fra R output har vi

$$\hat{\beta}_{\text{vand}=50\%} - \hat{\beta}_{\text{vand}=100\%} = -0.18522, \quad \text{SE}(\hat{\beta}_{\text{vand}=50\%} - \hat{\beta}_{\text{vand}=100\%}) = 0.01163.$$

I august måned er der gået 7 måneder siden forsøget startede i januar måned. Et 95% konfidensinterval for forskellen mellem log-volumen af et træ der har fået 50% vand i forhold til et træ der har fået 100% vand er dermed

$$7 \cdot (\hat{\beta}_{\text{vand}=50\%} - \hat{\beta}_{\text{vand}=100\%}) \pm t_{0.975, \text{df}=456} \cdot 7 \cdot \text{SE}(\hat{\beta}_{\text{vand}=50\%} - \hat{\beta}_{\text{vand}=100\%})$$

Fra appendix C.3 ses, at vi med god approksimation kan brugen normalfordelingsfraktilen i stedet for t-fordelingsfraktilen med 456 frihedsgrader. Dermed bliver 95% konfidensintervallet

$$7 \cdot (-0.18522) \pm 1.96 \cdot 7 \cdot 0.01163 = [-1.456104 ; -1.136976],$$

hvilket efter tilbagetransformation med eksponentialfunktionen bliver

$$[\exp(-1.456104) ; \exp(-1.136976)] = [0.2331 ; 0.3208].$$

Med 95% konfidens har vi dermed, at median volumen af en ung baobab træ, der har fået 50% vand i vækstperioden fra januar til august måned, er mellem 23% of 32% af median volumen for et tilsvarende træ, der har fået 100% vand.

- (d) Lad  $Y$  angive log-volumen af et baobab træ i maj måned når det har fået 75% vand siden januar måned. Idet der er gået 4 måneder fra januar til maj gælder

$$Y \sim N(4.8 + 0.22 \cdot 4, 0.7^2) = N(5.68, 0.7^2)$$

Ved standardisering og opslag i Appendix C.2 fås, at sandsynligheden for at volumen er mindre end 150 er givet ved

$$\begin{aligned} P(Y \leq \log(150)) &= P\left(\frac{Y - 5.68}{0.7} \leq \frac{\log(150) - 5.68}{0.7}\right) \\ &= \Phi(-0.9562353) \\ &\approx \Phi(-0.96) \\ &= 0.169 \end{aligned}$$

- (e) Lad  $Y$  og  $X$  angive log-volumen af to baobab træer i august måned, der har fået henholdsvis 50% og 100% vand siden januar måned. Idet der er gået 7 måneder fra januar til august gælder

$$Y \sim N(4.8 + 0.15 \cdot 7, 0.7^2) = N(5.85, 0.7^2)$$

$$X \sim N(4.8 + 0.33 \cdot 7, 0.7^2) = N(7.11, 0.7^2)$$

Og idet  $Y$  og  $X$  er uafhængige gælder

$$X - Y \sim N(7.11 - 5.85, 0.7^2 + 0.7^2) = N(1.26, 0.98)$$

At træet der har fået 50% vand har den største volumen er det samme som  $X - Y < 0$ , og ved standardisering og opslag i Appendix C.2 fås sandsynligheden

$$\begin{aligned} P(X - Y < 0) &= P\left(\frac{X - Y - 1.26}{\sqrt{0.98}} \leq \frac{-1.26}{\sqrt{0.98}}\right) \\ &= \Phi(-1.272792) \\ &\approx \Phi(-1.27) \\ &= 0.102 \end{aligned}$$

Slut på opgavesættet.

# Besvarelse til reeksamen i Statistisk Dataanalyse 1, januar 2015

## Opgave 1

- (a) Hvis der findes en lineær sammenhæng mellem SWL og indkomstskat, så vil hældningskoeffienter mht. indkomstskat være forskellig fra 0. I den simple lineære regression af SWL på indkomstskat, som findes i R modellen  $m1$ , er hældningskoefficienten estimeret til 1.9397 med en standard error på 0.3301. Hypotesen at hældningskoefficienten er 0 forkastes med  $p=0.000000285$ . Der er altså en stærkt signifikant positiv sammenhæng mellem SWL og indkomstskat. Populært sagt; lande med højere indkomstskat har befolkninger, som er mere lykkelige.
- (b) Der skal laves et 95% konfidensinterval for prædiktionen af SWL i modellen  $m1$  når indkomst=59. Vi bruger formel (7.2) på side 197 i bogen med  $n = 55$ :

$$\begin{aligned} 95\%-KI &= \hat{\alpha} + \hat{\beta} \cdot 59 \pm t_{0.975, df=55-2} \cdot s \cdot \sqrt{\frac{1}{55} + \frac{(59 - \bar{x})^2}{SS_x}} \\ &= 145.7468 + 1.9397 \cdot 59 \pm 2.005746 \cdot 28.45 \cdot \sqrt{\frac{1}{55} + \frac{(59 - 34.28182)^2}{7428.382}} \\ &= [242.1051; 278.2731] \end{aligned}$$

- (c) Svaret på dette spørgsmål er givet ved et 95% prædiktionsinterval for SWL i modellen  $m1$  når indkomst=35. Vi bruger formel (7.3) på side 197 i bogen med  $n = 55$ :

$$\begin{aligned} 95\%-PI &= \hat{\alpha} + \hat{\beta} \cdot 35 \pm t_{0.975, df=55-2} \cdot s \cdot \sqrt{1 + \frac{1}{55} + \frac{(35 - \bar{x})^2}{SS_x}} \\ &= 145.7468 + 1.9397 \cdot 35 \pm 2.005746 \cdot 28.45 \cdot \sqrt{1 + \frac{1}{55} + \frac{(35 - 34.28182)^2}{7428.382}} \\ &= [156.0544; 271.2182] \end{aligned}$$

- (d) Modellen  $m2$  er en multilineær regression, som beskriver SWL via de forklarende variable selvkab, indkomst og VAT. Vi ser først på fortegnet af de estimerede hældningskoefficienter: Der er en negativ sammenhæng med selvkabsskat og med VAT, mens der er en positiv sammenhæng med indkomstskat. Lande med lavere selvkabsskat har dermed befolkninger, som er mere lykkelige. Tilsvarende med lavere omsætningsafgift, og med højere indkomstskat. Der er dog kun sammenhængene med omsætningsafgift ( $p=0.0223$ ) og med indkomstskat ( $p=0.0000000806$ ), som er signifikante, mens sammenhængen med selvkabsskat slet ikke er signifikant ( $p=0.2669$ ).

## Opgave 2

- (a) Vi skal undersøge om proportionen af kvinder med forhøjet blodtryk afhænger af om de har fået aspirin eller ej. Det gøres ved at teste nul hypotesen om homogenitet mellem de 2 søjler i 2x2-tabellen. Vi ønsker at gøre dette ved et chi-kvadrat test, og kontrollerer først om *tommelfingerregler* er opfyldt:

$$\frac{15 * 31}{65} = 7.153846 > 5$$

Idet den mindste forventede værdi er større end 5 kan vi stole på chi-kvadrat testet. Teststørrelsen er givet ved

$$X^2 = \frac{65 * (4 * 20 - 30 * 11)^2}{15 * 50 * 34 * 31} = 5.139152$$

p-værdien fås ved at evaluere  $X^2$  i en chi-kvadrat fordeling med 1 frihedsgrad. Fra R findes p-værdien

```
> 1-pchisq(5.139152,df=1)
[1] 0.02339207
```

Dermed forkastes nul hypotesen på et 5% signifikansniveau. Der er altså evidens for en sammenhæng mellem aspirin og blodtryk. Idet andelen af kvinder med forhøjet blodtryk er lavere i aspirin gruppen konkluderes dermed, at aspirin kan bruges til at forebygge forhøjet blodtryk hos gravide kvinder.

Bemærk: Aspirin har andre virkninger. Generelt set frarådes det at tage aspirin under graviditeten.

- (b) Man kan ikke benytte chi-kvadrat testet på 2x6-tabellen idet *tommenfingerreglen* ikke er opfyldt. Den minste forventede værdi,

$$\frac{22 * 43}{351} = 2.695157,$$

er nemlig mindre end 5.

- (c) Ved at slå kolonnerne sammen to-og-to fås følgende 2x3-tabel af observationer:

Kejsersnit	skostørrelse			Total
	$\leq 37$	37.5 eller 38	$\geq 39$	
Ja	12	13	18	43
Nej	45	77	186	308
Total	57	90	204	351

Den tilsvarende 2x3-tabel af forventede værdier under hypotesen om homogenitet mellem de to rækker er:

Kejsersnit	skostørrelse			Total
	$\leq 37$	37.5 eller 38	$\geq 39$	
Ja	6.982906	11.025641	24.991453	43
Nej	50.01709	78.97436	179.00855	308
Total	57	90	204	351

Dermed bliver teststørrelsen

$$\begin{aligned} X^2 &= \frac{(12 - 6.982906)^2}{6.982906} + \dots + \frac{(186 - 179.00855)^2}{179.00855} \\ &= 6.7398 \end{aligned}$$

Denne teststørrelse skal evalueres i en chi-kvadrat fordeling med  $(2-1) \times (3-1) = 2$  frihedsgrader. Vha. R fås p-værdien

```
> 1-pchisq(6.7398, df=2)
[1] 0.03439308
```

Der er altså en signifikant sammenhæng mellem skostørrelse og risiko for at få kejsersnit ( $p=0.0344$ ). Estimater for risikoen for kejsersnit findes ved at tage proportionerne indenfor søjlerne:

	skostørrelse			Total
	$\leq 37$	37.5 eller 38	$\geq 39$	
Risiko for kejsersnit	21.1%	14.4%	8.8%	12.3%

Det ses, at risikoen for kejsersnit aftager med skostørrelsen. En mulig forklaring er, at skostørrelsen kan bruges som et mål for en kvindes generelle størrelse. Og at mindre kvinder har større risiko for at få kejsersnit.

- (d) Vi skal lave et 95% konfidensinterval for proportionen i søjlen " $\leq 37$ ". Formel (10.6) på side 283 i bogen giver

$$95\%-KI = \frac{12}{57} \pm 1.96 * \sqrt{\frac{12 * 45}{57^3}} = [0.1071; 0.3140]$$

- (e) Den efterspurgte sandsynlighed kan findes ved at bruge *loven om total sandsynlighed*, som findes i Infobox 9.5 på side 264 i bogen:

$$\begin{aligned} \hat{p} &= \frac{12}{57} * \frac{2}{10} + \frac{13}{90} * \frac{3}{10} + \frac{18}{204} * \frac{5}{10} \\ &= 0.1295562 \end{aligned}$$

Vores estimat for at sandsynligheden for at en tilfældig valgt kvinde blandt de 10 kvinder på fødegangen for kejsersnit er altså ca. 13%.

## Opgave 3

- (a) Den statistiske model hørende til `m1` er en 2-sidet variansanalyse med vekselvirkning. Altså

$$\text{koncentration}_i = \alpha(\text{temperatur}_i, \text{tilsætning}_i) + \epsilon_i \quad \text{for } i = 1, \dots, 24,$$

hvor fejlleddene  $\epsilon_1, \dots, \epsilon_{24}$  er indbyrdes uafhængige og normalfordelte med mid-delværdi 0 og den samme varians  $\sigma^2$ . I `drop1()` testes hypotesen om ingen vekselvirkning. Altså om modellen kan reduceres til den additive model

$$\text{koncentration}_i = \beta(\text{temperatur}_i) + \gamma(\text{tilsætning}_i) + \epsilon_i \quad \text{for } i = 1, \dots, 24.$$

Denne hypotese accepteres med  $p=0.2953$ . Der er altså ingen evidens for en vekselvirkning mellem temperatur og mængden af tilsætningsstoffet.

- (b) Reference gruppen i `m2` er givet ved `temperatur=I` og `tilsætning=A`. Dermed er estimat og standard error for den tilhørende vitamin koncentration givet ved (`Intercept`) parameteren i `m2`. Vi finder konfidensinterval som estimat  $\pm$  t-fraktil gange standard error, hvor antal frihedsgrader er det samme som for residual spredningen, dvs.  $df=18$ . Vi har  $t_{0.975, df=18} = 2.101$ , og dermed

$$95\%-KI = 39.083 \pm 2.101 * 2.666 = [33.482; 44.684]$$

- (c) Idet der er lige mange observationer for hver kombination af temperatur og tilsætning, gælder

$$\text{SE}(\hat{\gamma}(D) - \hat{\gamma}(C)) = \text{SE}(\hat{\gamma}(B) - \hat{\gamma}(A)) = 3.078$$

Dermed er t-teststørrelsen for nul hypotesen  $\gamma(D) = \gamma(C)$  givet ved

$$t_{\text{obs}} = \frac{\hat{\gamma}(D) - \hat{\gamma}(C)}{\text{SE}(\hat{\gamma}(D) - \hat{\gamma}(C))} = \frac{30.667 - 27.667}{3.078} = 0.9746589$$

Denne teststørrelse skal evalueres i en t-fordeling med 18 frihedsgrader. Via R findes p-værdien

```
> 2*(1-pt(0.9746589, df=18))
[1] 0.3426469
```

Vi accepterer dermed nul hypotesen  $\gamma(D) = \gamma(C)$ . Dermed er data i overensstemmelse med producentens forhåndsforventning om, at der findes et niveau for mængden af tilsætningsstoffet, således at der ikke er yderligere effekt af at til sætte mere af tilsætningsstoffet ud over dette niveau. Dette niveau kunne nemlig være niveau  $C$ .

Slut på opgavesættet.

# Besvarelse til eksamen i Statistisk Dataanalyse 1, november 2015

## Opgave 1

- (a) Der er 4 antagelser på fejlleddene i en lineær regression. Disse antagelser kan vurderes via modellernes residualer:

1. Residualerne skal have middelværdi 0 uanset værdien af den prædikterede værdi. For at dette er opfyldt skal gennemsnittet på y-aksen være cirka 0 omkring alle værdier på x-aksen i residualplottet (figurerne i højre kolonne). Dette gælder for log(pris)-modellen, men ikke for pris-modellen.
2. Residualerne skal have samme spredning uanset værdien af den prædikterede værdi. For at dette er opfyldt skal bredden på y-aksen være cirka den samme for alle værdier på x-aksen i residualplottet. Dette gælder for log(pris)-modellen, men ikke for pris-modellen.
3. Residualerne skal være normalfordelte, hvilket ses som en ret linie i qq-plottet (figurerne i venstre kolonne). Dette gælder for log(pris)-modellen, men ikke for pris-modellen.
4. De skal være uafhængige. Generelt set kan dette ikke vurderes udfra residualerne. Men udfra beskrivelsen af datasættet kan vi antage uafhængighed mellem observationerne.

Altså konkluderes, at det kun er log(pris) modellen, som er statistisk valid.

- (b) Vi vælger altså en lineær regression for logaritmen af prisen mod bilens alder. Den statistiske model er

$$\log(\text{pris}_i) = \alpha + \beta * \text{alder}_i + \epsilon_i, \quad i = 1, \dots, 121,$$

hvor  $\epsilon_1, \dots, \epsilon_{121}$  er uafhængige og  $N(0, \sigma^2)$  fordelte. Parametrene er intercept  $\alpha$ , hældning  $\beta$  og spredning  $\sigma > 0$ . Estimaterne aflæses fra `summary(linRegLog)`:

$$\hat{\alpha} = 11.760901, \quad \hat{\beta} = -0.164965, \quad \hat{\sigma} = 0.3365.$$

Idet der er angivet standard errors for intercept og hældnings parametrene kan vi også finde 95% konfidensintervaller for disse parametre. Vi skal bruge en t-fordeling med 119 frihedsgrader. Denne findes ikke i Appendix C.3, så vi bruger i stedet for 97.5% fraktilen for en t-fordeling med 100 frihedsgrader. Denne er 1.98, hvilket giver 95% konfidensinterval for  $\alpha$ :

$$11.760901 \pm 1.98 * 0.059744 = [11.64261; 11.87919]$$

Og 95% konfidensinterval for  $\beta$ :

$$-0.164965 \pm 1.98 * 0.006727 = [-0.1782845; -0.1516455]$$

- (c) Vi starter med at lave et 95% konfidensinterval for  $\log(\text{pris})$  når alder=5. Dette er

$$\begin{aligned}\hat{\alpha} + 5 * \hat{\beta} &\pm 1.98 * \hat{\sigma} * \sqrt{\frac{1}{n} + \frac{(5 - \text{mean(alder)})^2}{(n - 1) * \text{sd(alder)}^2}} \\ &= 11.760901 - 5 * 0.164965 \pm 1.98 * 0.3365 * \sqrt{\frac{1}{121} + \frac{(5 - 7.628099)^2}{120 * 4.566421^2}} \\ &= [10.86612; 11.00603].\end{aligned}$$

For at få 95% konfidensintervallet for prisen tilbagetransformeres med eksponentialekfunktionen. Dette giver intervallet:

$$[\exp(10.86612); \exp(11.00603)] = [52371.61; 60236.27]$$

- (d) For at afgøre om prisen på en brugt bil er usædvanlig skal vi sammenligne denne med populationen. Altså med et prædiktionsinterval. Vi vælger at lave et 95% prædiktionsinterval. Først for  $\log(\text{prisen})$ :

$$\begin{aligned}\hat{\alpha} + 5 * \hat{\beta} &\pm 1.98 * \hat{\sigma} * \sqrt{1 + \frac{1}{n} + \frac{(5 - \text{mean(alder)})^2}{(n - 1) * \text{sd(alder)}^2}} \\ &= 11.760901 - 5 * 0.164965 \pm 1.98 * 0.3365 * \sqrt{1 + \frac{1}{121} + \frac{(5 - 7.628099)^2}{120 * 4.566421^2}} \\ &= [10.26614; 11.60601],\end{aligned}$$

som ved tilbagetransformation giver

$$[\exp(10.26614); \exp(11.60601)] = [28742.73; 109755.45]$$

Idet 80000 ligger indenfor 95% prædiktionsintervallet for prisen på en 5 år gammel bil vælger vi at sige, at en pris på 80000 ikke er usædvanlig.

- (e) Den forventede pris på en 0 år gammel brugt bil er  $\exp(\hat{\alpha})$ . Alderen  $x$  (i år) hvor den forventede pris er halveret er dermed givet ved ligningen:

$$\exp(\hat{\alpha} + \hat{\beta} * x) = \frac{1}{2} * \exp(\hat{\alpha})$$

Løses denne ligning i  $x$  fås:

$$x = \frac{\log(\frac{1}{2})}{\hat{\beta}} = \frac{\log(2)}{-\hat{\beta}} = \frac{0.6931472}{0.164965} = 4.201783$$

## Opgave 2

- (a) Der er brugt en 1-vejs ANOVA. Den statistiske model er givet ved:

$$\text{frekevens}_i = \alpha(\text{ojenfarve}_i) + \epsilon_i, \quad i = 1, \dots, 19,$$

hvor  $\epsilon_1, \dots, \epsilon_{19}$  er uafhængige og  $N(0, \sigma^2)$  fordelte. Nul hypotesen

$$H_0: \alpha(\text{Blå}) = \alpha(\text{Brun}) = \alpha(\text{Grøn})$$

betyder, at der ikke er nogen sammenhæng mellem øjenfarve og den kritiske frekvens. Fra R output ses, at  $H_0$  afvises med p-værdien 0.02325. Vi konkluderer dermed, at øjenfarven har betydning for den kritiske frekvens.

- (b) I den parametrisering af 1-vejs ANOVA'en som R vælger som standard svarer interceptet til den forventede kritiske frekvens for personer med blå øjne. Idet er bliver spurgt til fordelingen af population af personer med blå øjne skal vi beregne et 95% prædiktionsinterval for at besvare spørgsmålet. Dette er givet ved:

$$\begin{aligned}\hat{\alpha}(\text{Blå}) \pm t_{0.975, \text{df}=16} * s * \sqrt{1 + \frac{1}{n_{\text{Blå}}}} &= 28.1667 \pm 2.120 * 1.547 * \sqrt{1 + \frac{1}{6}} \\ &= [24.62428; 31.70912]\end{aligned}$$

- (c) Udfra `summary(model)` findes estimatet

$$\hat{\alpha}(\text{Brun}) - \hat{\alpha}(\text{Grøn}) = -2.5792 - (-1.2467) = -2.5792 + 1.2467 = -1.3325$$

Standard error er givet ved:

$$\begin{aligned}\text{SE}(\hat{\alpha}(\text{Brun}) - \hat{\alpha}(\text{Grøn})) &= s * \sqrt{\frac{1}{n_{\text{Brun}}} + \frac{1}{n_{\text{Grøn}}}} = 1.547 * \sqrt{\frac{1}{8} + \frac{1}{5}} \\ &= 1.547 * \sqrt{\frac{13}{40}} = 0.8819\end{aligned}$$

Dermed fås 95% konfidensinterval:

$$-1.3325 \pm 2.120 * 1.547 * \sqrt{\frac{13}{40}} = [-3.2022; 0.5372]$$

- (d) I R output fra `summary(model)` aflæses p-værdierne for nul hypoteserne

$$H_1: \alpha(\text{Blå}) = \alpha(\text{Brun}), \quad H_2: \alpha(\text{Blå}) = \alpha(\text{Grøn}).$$

Disse er henholdsvis  $p_1 = 0.00708$  og  $p_2 = 0.20200$ . Fra spørgsmål (c) vides, at 0 ligger i 95% konfidensintervallet for  $\alpha(\text{Brun}) - \alpha(\text{Grøn})$ . Dette betyder, at p-værdien for nul hypotesen

$$H_3: \alpha(\text{Brun}) = \alpha(\text{Grøn})$$

er større end 0.05. Ved en Bonferroni korrektion skal man gange med antallet af tests, hvilket her er 3, før man sammenligner med signifikansniveauet. Vi konkluderer dermed, at der kun er signifikant forskel (på 5% signifikansniveau) mellem øjenfarverne blå og brun efter en Bonferroni korrektion for multipel testing.

### Opgave 3

- (a) Hvis de studerende løber netop når de har slæt krone, så vil  $p_{løb} = 0.5$ . Men hvis nogle studerende vælger ikke at rette sig efter møntkastet, så vil der generelt set gælde  $p_{løb} \neq 0.5$ . For at undersøge om nogle af de studerende “snyder” er det derfor relevant at teste nul hypotesen  $H_0: p_{løb} = 0.5$ .

(b) 2,3

- (c) Beregningsmæssigt er det tungt at lave et binomialtest for  $p = 0.5$  med  $y = 7$  og  $n = 7+18 = 25$  uden direkte brug af R. Så vi vælger at lave det beregningsmæssige lettere goodness-of-fit test. Teststørrelsen er:

$$X^2 = \frac{(7 - 0.5 * 25)^2}{0.5 * 25} + \frac{(18 - 0.5 * 25)^2}{0.5 * 25} = 4.84,$$

som skal evalueres i en chi-kvadrat fordeling med  $2-1=1$  frihedsgrader. Idet 95% fraktilen for denne chi-kvadrat fordeling er 3.841 fås, at p-værdien er under 0.05. Idet den estimerede sandsynlighed  $\hat{p}_{løb} = 7/25$  er mindre end 0.5, er der altså noget der tyder på, at nogle studerende snyder og bliver siddende selv om de slog krone.

- (d) Vi har to uparrede stikprøver af pulsændringer for henholdsvis *løbe* og *hvile* grupperne. Vi mangler derfor blot at afgøre om vi vil antage ens varians eller ej. Der er to grunde til ikke at antage ens varianser. Dels er stikprøve spredningen for *løbe* gruppen ( $sd=20.02498$ ) meget større end stikprøve spredningen for *hvile* gruppen ( $sd=3.704011$ ). Dels giver det også biologisk mening, at *løbe* gruppen har en større spredning end *hvile* gruppen. For *løbe* gruppen må man nemlig antage, at pulsændringen afhænger af de studerendes fysiske form etc, mens for *hvile* gruppen vil man forvente af pulsændringerne ligger tæt på 0 (idet der jo ikke er nogen væsentlig ændring fra før til efter for denne gruppe). Vi vælger altså modellen for to uparrede stikprøver med forskellig varians.

- (e) Den statistiske model valgt i (d) er givet ved:

$$x_1, \dots, x_9 \text{ er uafhængige og } N(\mu_{løb}, \sigma_{løb}^2) \text{ fordelte,}$$
$$y_1, \dots, y_{12} \text{ er uafhængige og } N(\mu_{hvile}, \sigma_{hvile}^2) \text{ fordelte.}$$

Fra R output fra `t.test(x,y,var.equal=FALSE)` aflæses 95% konfidensinterval for  $\mu_{løb} - \mu_{hvile}$  til at være

$$[46.12637; 77.04029]$$

Slut på opgavesættet.

# Besvarelse til reeksamen i Statistisk Dataanalyse 1, 3. februar 2016

I de fleste af svarerne nedenfor angives lige så mange decimaler, som der kommer ud af R. Dette er praktisk i forhold til sammenligning af resultater, men normalt vil man ikke bruge så mange decimaler.

## Opgave 1

- (a) For at undersøge om risikoen for lav fødselsvægt er den samme blandt rygere og ikke rygere laves et homogenitetstest mellem kolonnerne i  $2 \times 2$ -tabellen:

Fødselsvægt	Ryger		Total
	nej	ja	
$\leq 2500$ gram	29	30	59
$> 2500$ gram	345	175	520
Total	374	205	579

Med Yates' kontinuitetskorrektion fås teststørrelsen

$$X^2 = \frac{579 * (|29 * 175 - 345 * 30| - \frac{579}{2})^2}{59 * 520 * 374 * 205} = 6.11808$$

Ved opslag i en  $\chi^2$ -fordeling med 1 frihedsgrad fås en p-værdi på 0.0134:

```
> 1-pchisq(6.11808,df=1)
[1] 0.01338061
```

Vi forkaster altså nulhypotesen om homogenitet. Og idet risikoen for lav fødselsvægt er højere blandt rygere ( $\hat{p}_{\text{ryger}} = 30/205 = 14.6\%$ ) end blandt ikke-rygere ( $\hat{p}_{\text{ikke ryger}} = 29/374 = 7.8\%$ ) konkluderes, at rygning medfører en signifikant forøget risiko for lav fødselsvægt.

- (b) Ved brug af formel (11.12) i bogen fås følgende 95% konfidensinterval for forskellen  $\hat{p}_{\text{ryger}} - \hat{p}_{\text{ikke ryger}}$ :

$$\frac{30}{205} - \frac{29}{374} \pm 1.96 * \sqrt{\frac{30 * 175}{205^3} + \frac{29 * 345}{374^3}} = [0.01334189; 0.12426083]$$

- (c) Under antagelse af nulhypotesen, at der ikke er nogen sammenhæng mellem fødselsvægt og antallet af jordemoderbesøg under graviditeten, er det forventede antal kvinder givet ved rækemarginal gange søjlemarginal divideret med det totale antal. Det mindste forventede antal, som er for cellen ( $\leq 2500$  gram, 3+), er således givet ved

$$\frac{59 * 31}{579} = 3.16$$

Og hele tabellen over de forventede antal observationer er givet ved

Fødselsvægt	Jordemoderbesøg				Total
	0	1	2	3+	
$\leq 2500$ gram	29.86	16	9.99	3.16	59
$> 2500$ gram	263.14	141	88.01	27.84	520
Total	293	157	98	31	579

- (d) For at få krydstabuleringen af fødselsvægt mod antal jordemoderbesøg, hvor den sidste variabel grupperes i 0, 1 og 2+, lægges de to sidste kolonner i  $2 \times 4$ -tabellen sammen. Dette giver

Fødselsvægt	Jordemoderbesøg			Total
	0	1	2+	
$\leq 2500$ gram	36	11	12	59
$> 2500$ gram	257	146	117	520
Total	293	157	129	579

Fra spørgsmål (c) vides at der er en celle i  $2 \times 4$ -tabellen, som har forventet antal observationer lavere end 5. Dermed kan der være problemer med den  $\chi^2$ -approksimation der bruges i beregning af p-værdien, jvf. *tommelfingerreglen*. For øvrigt fås også en advarsel i R:

```
> chisq.test(jordemoder)
Pearson's Chi-squared test

data: jordemoder
X-squared = 5.3353, df = 3, p-value = 0.1488

Warning message:
In chisq.test(jordemoder) : Chi-squared approximation may be incorrect
```

Dette problem forsvinder hvis de to sidste kolonner slås sammen. Hvis man vil bruge  $\chi^2$ -approksimationen i beregningen af p-værdien, så bør man derfor lave det statistiske test på  $2 \times 3$ -tabellen.

Alternativt kan man argumentere for, at  $2 \times 4$ -tabellen indeholder mere detaljeret information vedrørende antal jordemoderbesøg. For at dette kan godkendes som et korrekt svar skal man dog betone, at p-værdien så ikke bør beregnes via  $\chi^2$ -approksimationen. F.eks. kan man beregne p-værdien via simulering. Dette ikke er blevet diskuteret på SD1, men i praksis er det let at gøre i R:

```
> chisq.test(jordemoder, simulate.p.value = TRUE)
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: jordemoder
X-squared = 5.3353, df = NA, p-value = 0.1574
```

- (e) *Loven om total sandsynlighed* (Infobox 10.5) giver, at sandsynligheden for lav fødselsvægt hos de 20 børn er

$$\begin{aligned}\hat{p} &= P(\leq 2500 \text{ gram} | \text{ryger}) * P(\text{ryger}) + P(\leq 2500 \text{ gram} | \text{ikke-ryger}) * P(\text{ikke-ryger}) \\ &= \frac{30}{205} * \frac{7}{20} + \frac{29}{374} * \frac{13}{20}\end{aligned}$$

Og idet der er 20 børn i alt, så er det forventede antal børn med lav fødselsvægt givet ved

$$20 * \hat{p} = \frac{30}{205} * 7 + \frac{29}{374} * 13 = 2.03$$

## Opgave 2

- (a) Den statistiske model er

$$\log(\text{hjernevægt}_{\text{art}}) = \alpha + \beta * \log(\text{kropsvægt}_{\text{art}}) + \epsilon_{\text{art}}, \quad \text{art} \in \{\text{Africanelephant}, \dots, \text{Yellow-belliedmarmot}\},$$

hvor  $\epsilon_{\text{Africanelephant}}, \dots, \epsilon_{\text{Yellow-belliedmarmot}}$  er uafhængige og  $N(0, \sigma^2)$  fordelte. Alternativt kan man naturligvis (som sædvanligt) indicerer de øvrige pattedyrsarter med index  $i = 1, \dots, 61$ . Parametrene er intercept  $\alpha$ , hældning  $\beta$  og spredning  $\sigma > 0$ . Estimaterne aflæses fra `summary(linRegLog)`:

$$\hat{\alpha} = 2.11500, \quad \hat{\beta} = 0.74228, \quad \hat{\sigma} = 0.6511.$$

Idet der er angivet standard errors for intercept og hældnings parametrene kan vi også finde 95% konfidensintervaller for disse parametre. Vi skal bruge en t-fordeling med 59 frihedsgrader. Denne findes ikke i Appendix C.3, så vi bruger i stedet for 97.5% fraktilen for en t-fordeling med 60 frihedsgrader. Denne er 2.00, hvilket giver 95% konfidensinterval for  $\alpha$ :

$$2.11500 \pm 2.00 * 0.09030 = [1.9344; 2.2956]$$

Og 95% konfidensinterval for  $\beta$ :

$$0.74228 \pm 2.00 * 0.02687 = [0.68854; 0.79602]$$

- (b) Vi starter med at lave et 95% konfidensinterval for  $\log(\text{hjernevægt})$  når kropsvægt=62. Dette er

$$\begin{aligned}\hat{\alpha} + \hat{\beta} * \log(62) &\pm 2.00 * \hat{\sigma} * \sqrt{\frac{1}{n} + \frac{(\log(62) - \text{mean}(\log(\text{kropsvægt})))^2}{(n-1) * \text{sd}(\log(\text{kropsvægt}))^2}} \\ &= 2.11500 + 0.74228 * \log(62) \pm 2.00 * 0.6511 * \sqrt{\frac{1}{61} + \frac{(\log(62) - 1.291808)^2}{60 * 3.128045^2}} \\ &= [4.952616; 5.404363].\end{aligned}$$

For at få 95% konfidensintervallet for hjernevægten tilbagetransformeres med eksponentialfunktionen. Dette giver intervallet:

$$[\exp(4.952616); \exp(5.404363)] = [141.5447; 222.3745]$$

- (c) For at afgøre om *mennesket* har en usædvanlig stor hjernevægt skal vi sammenligne med populationen af *pattedyrsarter*. Altså med et prædiktionsinterval. Vi vælger at lave et 95% prædiktionsinterval. Først for  $\log(\text{hjernevægt})$ :

$$\begin{aligned}\hat{\alpha} + \hat{\beta} * \log(62) &\pm 2.00 * \hat{\sigma} * \sqrt{1 + \frac{1}{n} + \frac{(\log(62) - \text{mean}(\log(\text{kropsvægt})))^2}{(n-1) * \text{sd}(\log(\text{kropsvægt}))^2}} \\ &= 2.11500 + 0.74228 * \log(62) \pm 2.00 * 0.6511 * \sqrt{1 + \frac{1}{61} + \frac{(\log(62) - 1.291808)^2}{60 * 3.128045^2}} \\ &= [3.856845; 6.500134],\end{aligned}$$

som ved tilbagetransformation giver

$$[\exp(3.856845); \exp(6.500134)] = [47.31583; 665.23052]$$

Idet *menneskets* hjernevægt på 1320 gram er cirka dobbelt så stor som den øvre grænse i 95% prædiktionsintervallet, så kan man sige, at *mennesket* har en usædvanlig stor hjerne i forhold til de øvrige pattedyrsarter.

- (d) Den prædikterede hjernevægt for *mennesket* og *rhesusaben* er

$$\begin{aligned}\log(\widehat{\text{hjernevægt}}_{\text{mennesket}}) &= \alpha + \beta * \log(62), \\ \log(\widehat{\text{hjernevægt}}_{\text{rhesusabe}}) &= \alpha + \beta * \log(6.8).\end{aligned}$$

Dermed fås

$$\begin{aligned}\log\left(\frac{\widehat{\text{hjernevægt}}_{\text{mennesket}}}{\widehat{\text{hjernevægt}}_{\text{rhesusabe}}}\right) &= \log(\widehat{\text{hjernevægt}}_{\text{mennesket}}) - \log(\widehat{\text{hjernevægt}}_{\text{rhesusabe}}) \\ &= \alpha + \beta * \log(62) - \alpha - \beta * \log(6.8) \\ &= \beta * \log\left(\frac{62}{6.8}\right) \\ &= 2.210212 * \beta.\end{aligned}$$

Ifølge spørgsmål (a) er estimat og 95% konfidensinterval for dette givet ved

$$\text{Estimat: } 2.210212 * 0.74228 = 1.640596,$$

$$95\% \text{ konfidensinterval: } [2.210212 * 0.68854; 2.210212 * 0.79602] = [1.521819; 1.759373].$$

Ved tilbagetransformation med eksponentialfunktionen fås estimat og 95% konfidensinterval for kvotienten mellem hjernevægten for *mennesket* og *rhesusaben*

$$\text{Estimat: } \exp(1.640596) = 5.158244,$$

$$95\% \text{ konfidensinterval: } [\exp(1.521819); \exp(1.759373)] = [4.580551; 5.808794].$$

Til sammenligning kan det oplyses (dette står ikke i opgaveteksten, men er en bonus information), at en rhesusabehjerne vejer 179 gram. Den observerede kvoteint mellem hjernevægten for *mennesket* og *rhesusaben* er dermed

$$\frac{1320}{179} = 7.37$$

Menneskehjernerne er altså også relativ stor i forhold til rhesusabehjernen.

## Opgave 3

- (a) Der er 4 antagelser på fejlleddene i en lineær normal model. Disse antagelser undersøges via modellens residualer. Idet forsøgsdesignet er balanceret har alle residualerne samme varians, hvormed det er lige meget om man bruger de *rå residualer* eller de *standardiserede residualer*.
1. Residualerne skal have middelværdi 0 uanset størrelsen af den prædikterede værdi. For at dette er opfyldt skal gennemsnittet på y-aksen være cirka 0 omkring alle værdier på x-aksen i residualplottet (figuren til venstre). Dette ses at være tilfældet.
  2. Residualerne skal have samme spredning uanset størrelsen af den prædikterede værdi. For at dette er opfyldt skal bredden på y-aksen være cirka den samme for alle værdier på x-aksen i residualplottet. Dette ses at være tilfældet.
  3. Residualerne skal være normalfordelte, hvilket ses som en ret linie i qq-plottet (figuren til højre). Dette ses at være tilfældet.
  4. Residualerne skal være uafhængige. Mange gange kan dette konkluderes ud fra beskrivelsen af datasættet. Dette er dog ikke tilfældet i dette datasæt, hvor observationer har en tidsmæssig ordning. Men at undersøge uafhængigheden nærmere ligger ud over pensum på *Statistisk Dataanalyse 1*, hvormed det er ok at antage uafhængighed.

Altså konkluderes, at den additive model er statistisk valid.

- (b) Man kan *ikke* bruge en 2-vejs ANOVA med vekselvirkning mellem *måned* og *fuldmåne*. En sådan model ville nemlig have 36 middelværdi parametre, hvormed der vil være  $n - 36 = 0$  frihedsgraden til at bestemme variansparameteren. Og uden frihedsgrader til at bestemme variansparameteren kan man ikke lave statistik.
- (c) Den additive model model er givet ved:

$$\text{antal}_i = \alpha + \beta(\text{måned}_i) + \gamma(\text{fuldmåne}_i) + \epsilon_i, \quad i = 1, \dots, 36,$$

hvor  $\epsilon_1, \dots, \epsilon_{36}$  er uafhængige og  $N(0, \sigma^2)$  fordelte. Nulhypotesen at der ikke er nogen sammenhæng mellem antal henvendelser og om det er fuldmåne formuleres som

$$H_0: \gamma(\text{Ja}) = \gamma(\text{Nej.efter}) = \gamma(\text{Nej.før})$$

R output fra `drop1(additiv.model,test="F")` giver, at  $H_0$  afvises med p-værdien 0.04533. Vi konkluderer dermed, at fuldmåne har en statistisk signifikant betydning for antal henvendelser. Parameterestimaterne i R output fra `summary(additiv.model)` viser videre, at der er flere henvendelser under fuldmåne.

- (d) I den parameterisering som R bruger er  $\hat{\gamma}(\text{Ja}) = 0$ . Og fra R output aflæses dermed

$$\begin{aligned}\hat{\gamma}(\text{Ja}) - \hat{\gamma}(\text{Nej.før}) &= -\hat{\gamma}(\text{Nej.før}) = 2.5, \\ \text{SE}(\hat{\gamma}(\text{Ja}) - \hat{\gamma}(\text{Nej.før})) &= \text{SE}(\hat{\gamma}(\text{Nej.før})) = 0.984\end{aligned}$$

Idet variansparameteren estimeres med 22 frihedsgrader og 97.5%-fraktilen den tilhørende t-fordeling er 2.073873 fås dermed, at 95% konfidensintervallet for forskellen mellem antal henvendelser per dag under og før fuldmåne er

$$2.5 \pm 2.073873 * 0.984 = [0.459309; 4.540691]$$

- (e) For at undersøge om der er signifikant forskel på antallet af henvendelser per dag *før* og *efter* fuldmåne opstilles nulhypotesen

$$H_0: \hat{\gamma}(\text{Nej.før}) = \hat{\gamma}(\text{Nej.efter})$$

T-teststørrelsen for testet af denne nulhypotese er givet ved

$$\begin{aligned}T_{\text{obs}} &= \frac{\hat{\gamma}(\text{Nej.før}) - \hat{\gamma}(\text{Nej.efter})}{\text{SE}(\hat{\gamma}(\text{Nej.før}) - \hat{\gamma}(\text{Nej.efter}))} \\ &= \frac{\hat{\gamma}(\text{Nej.før}) - \hat{\gamma}(\text{Nej.efter})}{\text{SE}(\hat{\gamma}(\text{Ja}) - \hat{\gamma}(\text{Nej.efter}))} \\ &= \frac{\hat{\gamma}(\text{Nej.før}) - \hat{\gamma}(\text{Nej.efter})}{\text{SE}(\hat{\gamma}(\text{Nej.efter}))} \\ &= \frac{-2.500 + 1.958}{0.984} \\ &= -0.550813,\end{aligned}$$

hvor vi har brugt at designet er balanceret til at omskrive standard error, således at denne kan aflæses fra R output. Dermed fås p-værdien via R beregningen

```
> 2*(1-pt(0.550813,df=22))
[1] 0.5873124
```

Nulhypotesen accepteres således med en p-værdi på 58%. Der er altså ikke noget der tyder på, at der er forskel i antallet af henvendelser per dag før og efter fuldmåne.

Slut på opgavesættet.

# Besvarelse til eksamen i Statistisk Dataanalyse 1, november 2016

I den vejledende besvarelse nedenfor er der flere steder angivet flere decimaler end nødvendigt. Grunden til dette er, at tallerne er indsat via copy-paste fra beregninger lavet i R.

## Opgave 1

- (a) Lad  $X \sim N(168.3448, 40.21111)$  være højden af en tilfældigt udvalgt kvinde. Gennemsnitshøjden for en mand er 183, hvormed vi ved standardisering beregner

$$\begin{aligned} P(X > 183) &= P\left(\frac{X - 168.3448}{\sqrt{40.21111}} > \frac{183 - 168.3448}{\sqrt{40.21111}}\right) \\ &= P(Z > 2.3111) = 1 - P(Z \leq 2.3111) \\ &= 0.01041367. \end{aligned}$$

Ovenfor har  $Z$  en standard normalfordeling.

- (b) Lad  $X \sim N(168.3448, 40.21111)$  og  $Y \sim N(183, 37.11765)$  være højden af henholdsvis kvinden og manden i venneparret. Hvis man ikke vælger sine venner udfra deres højde, så kan vi antage at  $X$  og  $Y$  er uafhængige. Forskellen mellem kvindens og mandens højde er dermed

$$X - Y \sim N(168.3448 - 183, 40.21111 + 37.11765) = N(-14.6552, 77.32876).$$

Ved standardisering kan vi dermed beregne sandsynligheden for at kvinden er højere end manden

$$\begin{aligned} P(X - Y > 0) &= P\left(\frac{X - Y + 14.6552}{\sqrt{77.32876}} > \frac{14.6552}{\sqrt{77.32876}}\right) \\ &= P(Z > 1.666561) = 1 - P(Z \leq 1.666561) \\ &= 0.04780086. \end{aligned}$$

Ovenfor har  $Z$  en standard normalfordeling.

- (c) Den sammenvejede stikprøvespredning findes ved at vægte varianserne med deres frihedsgrader, hvorefter man tager kvadratroden for at få spredningen:

$$s = \sqrt{\frac{86 * 40.21111 + 34 * 37.11765}{86 + 34}} = 6.271733.$$

I modellen for to uafhængige stikprøver med ens varians, dvs.

$$\begin{aligned} x_1, \dots, x_m &\sim N(\mu_x, \sigma^2), \\ y_1, \dots, y_n &\sim N(\mu_y, \sigma^2), \end{aligned}$$

er den sammenvejede stikprøvespredning estimatet for den fælles spredning  $\sigma$ .

- (d) Vi bruger modellen for to uafhængige stikprøver med ens varians, der er opskrevet i besvarelsen af spørgsmål (c). Vi tester nulhypotesen at de to bagvedliggende populationer har den sammen middelværdi, altså  $H_0: \mu_x = \mu_y$ . T-teststørrelsen for denne nulhypotese er

$$T_{\text{obs}} = \frac{\bar{y} - \bar{x}}{s \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{168.3448 - 183}{6.271733 \sqrt{1/87 + 1/35}} = -11.67395$$

Denne teststørrelse skal evalueres i en t-fordeling med  $87 + 35 - 2 = 120$  frihedsgrader. Dette giver p-værdien

```
> 2*(1-pt(11.67395,df=120))
[1] 0
```

p-værdien er altså 0, hvormed nulhypotesen klart afvises. Gennemsnitshøjden af mænd er altså signifikant større end gennemsnitshøjden af kvinder.

- (e) Et 90%-konfidensinterval for forskellen mellem mænds og kvinders gennemsnithøjde findes ved

$$\begin{aligned} \bar{y} - \bar{x} &\pm t_{0.95,\text{df}=120} * s * \sqrt{\frac{1}{m} + \frac{1}{n}} \\ &= 168.3448 - 183 \pm 1.657651 * 6.271733 * \sqrt{1/87 + 1/35} \\ &= [-16.73618; -12.57422] \end{aligned}$$

Her er t-fraktilen fundet ved opslag i R. Bemærk, at vi bruger en 95%-fraktil idet der skal laves et 90%-konfidensinterval.

- (f) Vi tester hypotesen om homogenitet mellem kvinder og mænd for sandsynligheden for at gættet på underviserens højde er for lavt. Først tilføjes række- og søjlemarginalerne til tabellen:

Køn af den studerende	Højdegæt var for lavt	Højdegæt var for højt	Sum
kvinde	71	16	87
mand	22	13	35
Sum	93	29	122

Testet kan laves som et  $\chi^2$ -test da tommelfingerreglen er opfyldet idet  $\frac{35*29}{122} = 8.319672 > 5$ . Med Yates' kontinuitets korrektion fås teststørrelsen

$$\begin{aligned} X^2 &= \frac{n * (|a * d - b * c| - \frac{n}{2})^2}{(a + b) * (c + d) * (a + c) * (b + d)} \\ &= \frac{122 * (|71 * 13 - 22 * 16| - \frac{122}{2})^2}{87 * 35 * 93 * 29} \\ &= 3.863954 \end{aligned}$$

og p-værdien

```
> 1-pchisq(3.863954,df=1)
[1] 0.04933394
```

Hypotesen om homogenitet afvises dermed på det klassiske 5% signifikansniveau. Der er altså evidens for en sammenhæng mellem de studerendes køn og hvorvidt de gætter for lavt eller for højt på underviserens højde. Estimateet for sandsynligheden for at gætte for lavt er

$$\hat{p}_{kvinde, \text{for lavt}} = \frac{71}{87} = 0.816092, \quad \hat{p}_{mand, \text{for lavt}} = \frac{22}{35} = 0.6285714.$$

Studerende som er meget gode til at gætte underviserens højde burde gætte for lavt (og for højt) med sandsynlighed 50%. Vi konkluderer dermed, at de mandlige studerende er signifikant bedre end de kvindelige studerende til at gætte underviserens højde. En mulig forklaring på dette er, at underviseren er en mand af almindelig højde, og at man er bedre til at gætte højden på en person som er cirka ligeså høj som man selv er.

Om ønsket kan man gå lidt dybere i analysen af disse tal — dette er dog ikke nødvendigt for at lave en fyldestgørende besvarelse af spørgsmål 1.f: Hvis man undersøger hypotesen for at sandsynligheden for at gætte for lavt er 50%, så fås følgende for kvinderne:

```
> prop.test(71,87)

1-sample proportions test with continuity correction

data: 71 out of 87, null probability 0.5
X-squared = 33.517, df = 1, p-value = 7.064e-09
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.7155435 0.8880851
sample estimates:
      p 
0.816092
```

Og følgende for mændene:

```
> prop.test(22,35)

1-sample proportions test with continuity correction

data: 22 out of 35, null probability 0.5
X-squared = 1.8286, df = 1, p-value = 0.1763
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
```

0.4494860 0.7800773

sample estimates:

p  
0.6285714

Kvindernes sandsynlighed for at gætte for lavt er således signifikant højere end 50%, mens vi med p-værdi 0.1763 accepterer nulhypotesen at mænd gætter for lavt med sandsynlighed 50%.

## Opgave 2

- (a) Der er lavet en lineær regression af  $Y$  på pH i jorden efter godskning med aske. Den statistiske model er

$$Y_i = \alpha + \beta * \text{pH}_i + \epsilon_i, \quad i = 1, \dots, 15,$$

hvor  $\epsilon_1, \dots, \epsilon_{15}$  er uafhængige og  $N(0, \sigma^2)$  fordelte. Parametrene er intercept  $\alpha$ , hældning  $\beta$ , og spredning  $\sigma > 0$ . Estimaterne aflæses fra `summary(m1)`:

$$\hat{\alpha} = 1.71715, \quad \hat{\beta} = 0.05010, \quad \hat{\sigma} = 0.1703.$$

Idet der er angivet standard errors for intercept og hældnings parametrene kan vi også finde 95% konfidensintervaller for disse parametre. Vi skal bruge en t-fordeling med 13 frihedsgrader. Ved opslag i Appendix C.3 fås 97.5% fraktilen til at være 2.16, hvilket giver 95% konfidensinterval for  $\alpha$ :

$$1.71715 \pm 2.16 * 0.19477 = [1.296447; 2.137853]$$

Og 95% konfidensinterval for  $\beta$ :

$$0.05010 \pm 2.16 * 0.03118 = [-0.0172488; 0.1174488]$$

- (b) Et 95% konfidensinterval for middelværdien af  $Y$  når  $\text{pH} = 7.0$  er

$$\begin{aligned} \hat{\alpha} + 7.0 * \hat{\beta} \pm 2.16 * \hat{\sigma} * \sqrt{\frac{1}{n} + \frac{(7.0 - \text{mean}(\text{pH}))^2}{(n-1) * \text{sd}(\text{pH})^2}} \\ = 1.71715 + 7.0 * 0.05010 \pm 2.16 * 0.1703 * \sqrt{\frac{1}{15} + \frac{(7.0 - 6.085333)^2}{14 * 1.460034^2}} \\ = [1.954651; 2.181049]. \end{aligned}$$

- (c) Der bliver spurgt til et 95% prædiktionsinterval for  $Y$  når  $\text{pH} = 5.6$ . Dette er givet ved

$$\begin{aligned} \hat{\alpha} + 5.6 * \hat{\beta} \pm 2.16 * \hat{\sigma} * \sqrt{1 + \frac{1}{n} + \frac{(5.6 - \text{mean}(\text{pH}))^2}{(n-1) * \text{sd}(\text{pH})^2}} \\ = 1.71715 + 5.6 * 0.05010 \pm 2.16 * 0.1703 * \sqrt{1 + \frac{1}{15} + \frac{(5.6 - 6.085333)^2}{14 * 1.460034^2}} \\ = [1.616395; 2.379025]. \end{aligned}$$

- (d) Hvis der laves et forsøg med 4 gange så mange potter, og dermed 20 potter for hver af de 3 aske mængder, så kan man forvente at  $\bar{pH}$  bliver cirka det samme som  $pH$  for det nuværende forsøg, og at

$$SS_{pH} = \sum_{i=1}^{60} (pH_i - \bar{pH})^2$$

bliver cirka 4 gange så stor som  $SS_{pH}$  for det nuværende forsøg. Idet vi forventer at få cirka samme estimat for  $\sigma$ , så vil vi dermed forvente at

$$SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{SS_{pH}}}$$

bliver cirka halvt så stor i det nye forsøg. Og idet vi forventer at få cirka samme estimat for  $\beta$ , så forventer vi at t-teststørrelsen for  $H_0: \beta = 0$  bliver cirka dobbelt så stor i det nye forsøg, altså cirka  $2 * 1.607 = 3.214$ . Denne teststørrelse skal nu evalueres i en t-fordeling med  $60 - 2 = 58$  frihedsgrader, hvilket givet p-værdien

```
> 2*(1-pt(2*1.607,df=58))
[1] 0.00213943
```

Hvis den sande hældning er  $\beta = 0.05010$ , så vil vi altså forvente at finde en stærkt signifikant effekt af pH i det nye forsøg.

- (e) Der er lavet en kvadratisk regression af  $Y$  på pH i jorden efter godtning med kalk. Den statistiske model er

$$Y_i = \alpha + \beta * pH_i + \gamma * pH_i^2 + \epsilon_i, \quad i = 1, \dots, 15,$$

hvor  $\epsilon_1, \dots, \epsilon_{15}$  er uafhængige og  $N(0, \sigma^2)$  fordelte. Parametrene er intercept  $\alpha$ , den lineære hældning  $\beta$ , den kvadratiske hældning  $\gamma$ , og spredning  $\sigma > 0$ . Estimaterne aflæses fra `summary(m2)`:

$$\hat{\alpha} = -8.55709, \quad \hat{\beta} = 3.62922, \quad \hat{\gamma} = -0.31892, \quad \hat{\sigma} = 0.1357.$$

Idet der er angivet standard errors for middelværdiparametrene kan vi også finde 95% konfidensintervaller for disse parametre. Vi skal bruge en t-fordeling med 12 frihedsgrader. Ved opslag i Appendix C.3 fås 97.5% fraktilen til at være 2.18, hvilket giver 95% konfidensinterval for  $\alpha$ :

$$-8.55709 \pm 2.18 * 1.08363 = [-10.919403; -6.194777]$$

95% konfidensinterval for  $\beta$ :

$$3.62922 \pm 2.18 * 0.36340 = [2.837008; 4.421432]$$

Og 95% konfidensinterval for  $\gamma$ :

$$-0.31892 \pm 2.18 * 0.02931 = [-0.3828158; -0.2550242]$$

- (f) Middelværdikurven hørende til aske er en ret linje (1'ste grads polynomium), og middelværdikurven hørende til kalk er en parabel (2'den grads polynomium). Forskellen mellem disse to kurver er dermed en parabel  $f(\text{pH}) = a*\text{pH}^2 + b*\text{pH} + c$ , hvor

$$a = -\gamma_{\text{kalk}} = 0.31892,$$

$$b = \beta_{\text{aske}} - \beta_{\text{kalk}} = 0.05010 - 3.62922 = -3.57912,$$

$$c = \alpha_{\text{aske}} - \alpha_{\text{kalk}} = 1.71715 + 8.55709 = 10.27424.$$

Denne parabel antager sit minimum i toppunktet, altså når

$$\text{pH} = \frac{-b}{2 * a} = \frac{3.57912}{2 * 0.31892} = 5.611313$$

Og den minimale afstand mellem middelværdikurven for aske og for kalk er

$$f(5.611313) = 0.31892 * 5.611313^2 - 3.57912 * 5.611313 + 10.27424 = 0.2324584.$$

Slut på opgavesættet.

# Besvarelse til eksamen i Statistisk Dataanalyse 1, februar 2017

I den vejledende besvarelse nedenfor er der flere steder angivet flere decimaler end nødvendigt. Grunden til dette er, at tallerne er indsat via copy-paste fra beregninger lavet i R.

## Opgave 1

- (a) Den observerede  $2 \times 3$ -tabel er

Søgealgoritme:	nuværende	A	B	Total
successful søgning	3511	1749	1818	7078
ikke-sucessful søgning	1489	751	682	2922
Total	5000	2500	2500	10000

De forventede antal brugere under nulhypotesen fås ved at gange række- og søjlemarginaler og dividere med det totale antal. F.eks. er øverste venstre hjørne i tabellen nedenfor givet ved

$$\frac{7078 * 5000}{10000} = 3539$$

Hele  $2 \times 3$ -tabellen af de forventede antal brugere er således

Søgealgoritme:	nuværende	A	B	Total
successful søgning	3539	1769,5	1769,5	7078
ikke-sucessful søgning	1461	730,5	730,5	2922
Total	5000	2500	2500	10000

- (b) For at undersøge om de tre søgealgoritmer er lige gode laves et test for homogenitet mellem søgerne. De forventede antal under denne nulhypotese er beregnet i spørgsmål (a). Vi kan dermed beregne chi-kvadrat teststørrelsen:

$$X^2 = \frac{(3511 - 3539)^2}{3539} + \dots + \frac{(682 - 730,5)^2}{730,5} \\ = 6,1203$$

P-værdien fås ved at finde sandsynligheden i højre hale i en chi-kvadrat fordeling med  $(2 - 1) * (3 - 1) = 2$  frihedsgrader. Denne sandsynlighed er

```
> 1-pchisq(6.1203,df=2)
[1] 0.04688066
```

Vi bemærker, at der ikke er nogen problemer med at bruge chi-kvadrat testet idet de forventede antal alle er langt større end 5.

Alternativt kan man finde p-værdien ved at bruge `chisq.test()` funktionen i R:

```
> chisq.test(matrix(c(3511, 1489, 1749, 751, 1818, 682), 2, 3))
```

Pearson's Chi-squared test

```
data: matrix(c(3511, 1489, 1749, 751, 1818, 682), 2, 3)
X-squared = 6.1203, df = 2, p-value = 0.04688
```

Idet p-værdien (=0,04688) er mindre en 5% konkluderes, at der er en signifikant forskel på de tre søgealgoritmer. Der er således evidens for, at algoritmerne ikke er lige gode.

- (c) For den *nuværende* søgealgoritme er et 95% konfidensinterval for sandsynligheden for en *succesfuld søgning* givet ved

$$\frac{3511}{5000} \pm 1.96 * \sqrt{\frac{3511 * 1489}{5000^3}} = [0,6895245; 0,7148755]$$

For søgealgoritme A er et 95% konfidensinterval for sandsynligheden for en *succesfuld søgning* givet ved

$$\frac{1749}{2500} \pm 1.96 * \sqrt{\frac{1749 * 751}{2500^3}} = [0,6816295; 0,7175705]$$

For søgealgoritme B er et 95% konfidensinterval for sandsynligheden for en *succesfuld søgning* givet ved

$$\frac{1818}{2500} \pm 1.96 * \sqrt{\frac{1818 * 682}{2500^3}} = [0,7097404; 0,7446596]$$

Af konfidensintervallerne fremgår det, at der ikke er forskel på den *nuværende* søgealgoritme og søgealgoritme A (dette kan bekræftes ved et hypotese test, men det er ikke en del af eksamensopgaven), mens søgealgoritme B har en større sandsynlighed for at give en *succesfuld søgning*. Søgealgoritme B er dermed bedst.

- (d) Vi bruger *loven om total sandsynlighed* (se Infobox 10.5):

$$\begin{aligned} P(\text{succesfuld } \text{søgning}) &= P(\text{succesfuld } \text{søgning} \mid \text{algoritme B}) * P(\text{algoritme B}) \\ &\quad + P(\text{succesfuld } \text{søgning} \mid \text{nuværende algoritme}) * P(\text{nuværende algoritme}) \\ &= \frac{1818}{2500} * 0,6 + \frac{3511}{5000} * 0,4 \\ &= 0,7172 \end{aligned}$$

(e) Vi skal beregne to sandsynligheder:

- (1) Der var 1818 brugere der foretog en *succesfuld søgning* med søgealgoritme  $B$  ud af 10000 brugere i alt. Dermed er svaret:

$$P(\text{algoritme } B, \text{succesfuld søger}) = \frac{1818}{10000} = 0,1818$$

- (2) Her skal vi kun betragte de 7078 brugere, der foretog en *succesfuld søger*. Heraf var der 1818 brugere, der foretog søgeringen med søgealgoritme  $B$ . Dermed er svaret:

$$P(\text{algoritme } B | \text{succesfuld søger}) = \frac{1818}{7078} = 0,2570701$$

## Opgave 2

- (a) De statistiske modeller, som svarer til R koderne er

$$\text{m1: } \text{height.guess}_i = \alpha_{\text{sex}_i} + \beta_{\text{sex}_i} * \text{height}_i + e_i$$

$$\text{m2: } \text{height.guess}_i = \alpha_{\text{sex}_i} + \beta * \text{height}_i + e_i$$

$$\text{m3: } \text{height.guess}_i = \alpha + \beta * \text{height}_i + e_i$$

For alle tre modeller gælder, at  $e_1, \dots, e_{122}$  er indbyrdes uafhængige og normalfordelte med middelværdi 0 og den samme varians.

Modellen m1 er også kendt som en ANCOVA, mens modellen m3 er en simpel lineær regression.

- (b) Ved modelreduktion fjernes de ikke-signifikante effekter en efter en. Vi starter med at teste vekselvirkningen i modellen m1. Dette findes som testet på parameteren `sexMand:height` i output fra `summary(m1)`. Vekselvirkningen er ikke-signifikant ( $p=0,650$ ), hvorefter m1 reduceres til modellen m2. I modellen m2 kan man teste både hovedvirkningen af køn og af højde. Disse tests findes som tests på parametrene `sexMand` og `height` i output fra `summary(m2)`. Her er hovedvirkningen af køn mindste signifikant ( $p=0,765$ ). Vi fjerner denne effekt, og reducerer m2 til modellen m3. Modellen m3 er en simpel lineær regression af højdegaættet på højde. I denne model er effekten af højde signifikant ( $p=0,0481$ ). Modellen m3 er dermed slutmodellen.

- (c) Parametrene i m3 er intercept  $\alpha$ , hældning  $\beta$  og residual spredning  $\sigma$ . Parametreestimaterne aflæses fra `summary(m3)`:

$$\hat{\alpha} = 166,26466, \quad \hat{\beta} = 0,08579, \quad \hat{\sigma} = 4,297.$$

Parametrene der beskriver middelværdien er  $\alpha$  og  $\beta$ . I output fra `summary(m3)` aflæses tilhørende standard errors:

$$SE(\hat{\alpha}) = 7,42161, \quad SE(\hat{\beta}) = 0,04296.$$

Desuden aflæses det, at der skal bruges en t-fordeling med 120 frihedsgrader ved beregningen af konfidensintervaller. For at lave 95% konfidensintervaller bruges 97,5%-fraktilen:

```
> qt(0.975, df=120)
[1] 1.97993
```

Dette giver 95% konfidensintervallet for  $\alpha$ :

$$166,26466 \pm 1,97993 * 7,42161 = [151,5704; 180,9589]$$

Og 95% konfidensintervallet for  $\beta$ :

$$0,08579 \pm 1,97993 * 0,04296 = [0,0007322072; 0,1708477928]$$

- (d) Et 95% konfidensinterval for middelværdien af højdegættet når  $height = 168$  er

$$\begin{aligned} \hat{\alpha} + 168 * \hat{\beta} &\pm 1,97993 * \hat{\sigma} * \sqrt{\frac{1}{n} + \frac{(168 - \text{mean}(height))^2}{(n-1) * \text{sd}(height)^2}} \\ &= 166,26466 + 168 * 0,08579 \pm 1,97993 * 4,297 * \sqrt{\frac{1}{122} + \frac{(168 - 172,5164)^2}{121 * 9,093167^2}} \\ &= [179,8166; 181,5381] \end{aligned}$$

- (e) Der bliver spurgt til et 95% prædiktionsinterval for højdegættet når  $height = 183$ . Dette er givet ved

$$\begin{aligned} \hat{\alpha} + 183 * \hat{\beta} &\pm 1,97993 * \hat{\sigma} * \sqrt{1 + \frac{1}{n} + \frac{(183 - \text{mean}(height))^2}{(n-1) * \text{sd}(height)^2}} \\ &= 166,26466 + 183 * 0,08579 \pm 1,97993 * 4,297 * \sqrt{1 + \frac{1}{122} + \frac{(183 - 172,5164)^2}{121 * 9,093167^2}} \\ &= [173,3753; 190,5532] \end{aligned}$$

- (f) Forskellen på den gennemsnitlige højde af mænd og af kvinder er  $183 - 168 = 15$ . Forskellen på det gennemsnitlige højdegæt af mænd og af kvinder er dermed

$$(\alpha + 183 * \beta) - (\alpha + 168 * \beta) = 15 * \beta$$

Et 95% konfidensinterval for denne forskel er dermed givet via 95% konfidensintervallet for  $\beta$  beregning i spørgsmål (c):

$$[15 * 0,0007322072; 15 * 0,1708477928] = [0,01098311; 2,56271689]$$

- (g) Ifølge Infobox 7.2 er de to vigtigste modelvalideringsplot 1) scatterplot af de standardiserede residuals mod de prædikterede værdier, 2) qq-plot af de standar-diserede residualer. Disse kan laves med følgende R kode, hvor `abline()` koderne indtegner referencelinjer:

```
plot(fitted(m3),rstandard(m3))
abline(h=0)
qqnorm(rstandard(m3))
abline(0,1)
```

Alternativt kan man få fire modelvalideringsplots (heraf to andre varianter af residualplottet, og også qq-plottet) fra R ved at “plotte” modellen:

```
plot(m3)
```

Slut på opgavesættet.