Matematisk Statistik: Modelbaseret Inferens

Gamle eksamensopgaver

Jens Ledet Jensen

Sceneskift

Forår 2018, opgave 1

Forår 2018 opgave 1

Lad $X_1, \ldots, X_n$ i.i.d. stokastiske variable med værdier i $\{1, 2, 3\}$,

$$P(X_i = 1) = \theta, \quad P(X_i = 2) = 2\theta, \quad P(X_i = 3) = 1 - 3\theta,$$

hvor θ er en ukendt parameter, $0 \leq \theta \leq 1/3$.
(a) Brug moment metoden til at udlede en estimator for θ.

(b) Er din estimator unbiased? Giv en kort begrundelse for dit svar.

Forår 2018 (1a)

Momentmetoden går ud på at finde skøn ved at sætte den teoretiske middelværdi lig meed gennemsnit af data

$$E(X) = 1 \cdot \theta + 2 \cdot 2\theta + 3(1 - 3\theta) = 3 - 4\theta$$

$$3 - 4\hat{\theta} = \bar{x} \Rightarrow \hat{\theta} = (3 - \bar{x})/4$$

Forår 2018 (1b)

Estimator er unbiased:

$E((3-\bar{X})/4) = (3-(3-4\theta))/4 = \theta$

Indirekte argument: momentligning er lineær i θ

Sceneskift

Forår 2018, opgave 2

Forår 2018 opgave 2

Der er blevet simuleret en stikprøve af størrelse 200 fra hver af tre forskellige fordelinger. Nedenstående figur viser et normalfraktilplot for en af stikprøverne, samt boxplots for alle tre.

(a) Hvilket af de tre boxplots A, B, eller C hører til det viste normalfraktilplot? Giv en kort begrundelse for dit svar.

Forår 2018 (2a

I qqplot aflæses median af data ved at finde 0 på 1.aksen, se hvor denne skærer "kurven" og finde tilhørende værdi på 2.aksen.

Værdi aflæses til lige lidt over 2. Dette passer kun med medianen i boxplot A

Sceneskift

Forår 2018, opgave 3

Forår 2018 opgave 3

Wafers er siliciumskiver, der danner råmateriale til elektroniske chips. Wafers produceres i renrum, da hver mikropartikel, der lander på overfladen, mindsker udbyttet af chippene. Det kan dog ikke helt undgås, at wafers kontamineres af partikler.

Et firma, der producerer renluftanlæg, lover, at i deres anlæg sætter sig gennemsnitligt kun 10 partikler på overfladen af en wafer i løbet af en time. For at teste firmaets påstand tælles antallet X af partikler på 5 wafers der lå i renrummet i 24 timer. Hvis antallet overstiger 1260 partikler, forkaster vi påstanden om, at der kun sætter sig 10 partikler per time på overfladen af en wafer.

Der antages, at partiklerne sætter sig på wafers uafhængigt af hinanden, og at antallet af partikler, der sætter sig på en wafer i løbet af en time, følger en Poisson-fordeling med parameter μ.

Forår 2018 (3a)

(a) Gør rede for, at antallet X er Poisson-fordelt med parameter $\lambda = 120 \cdot \mu$.

$X = \sum_{i=1}^{5} \sum_{j=1}^{24} X_{ij} \sim \text{poisson}(2 \cdot 24 \cdot \mu)$

da de stokastiske variable er uafhængige og poisson(μ) (MSRR, theorem B.6.2)

Forår 2018 (3b)

(b) Formuler nulhypotesen og den alternative hypotese i testet.
(c) Angiv sandsynligheden for type 1 fejl.
(d) Angiv sandsynligheden for type 2 fejl, i tilfældet hvor den sande rate er $\mu = 11$ partikler per time og wafer.

Hypotesen er $\mu = 10$, rate på 10 partikler per time per wafer

Alternativet er en højere værdi: $\mu > 10$

Type 1 fejl: forkaster sand hypotese:

$$P(\text{type1}) = P_{\mu=10}(X > 1260) = 1 - \text{ppois}(1260, 120 * 10) = 0.04120406$$

Forår 2018 (3d)

(d) Angiv sandsynligheden for type 2 fejl, i tilfældet hvor den sande rate er $\mu = 11$ partikler per time og wafer.

Type 2 fejl: forkaster ikke falsk hypotese:

$$P(\text{type2}) = P_{\mu=11}(X \leq 1260) = \text{ppois}(1260, 120 * 11) = 0.04992297$$

Fordi vi observerer 5 wafers i 24 timer har vi så meget information at vi i de fleste tilfælde vil opdage at raten er 11

Sceneskift

Forår 2018, opgave 4

Forår 2018 opgave 4

Selvom revnede æg sorteres fra, inden æg lægges i æggebakkerne til privatkunder, får man alligevel indimellem knækæg (æg med revner). Bakker med to eller flere knækæg mindsker kundetilfredsheden betydeligt. For at skønne risikoen fik et stort supermarked gennemset en levering på 1000 æggebakker, hver med 10 æg. Under antagelsen om, at æg i en bakke får revner uafhængigt af hinanden, burde antal knækæg følge en binomialfordeling. Under denne antagelse estimeres sandsynligheden for at et æg er et knækæg, og det forventede antal bakker med $0, 1, \ldots, 10$ knækæg beregnes. Tabellen viser en opstilling af de observerede og forventede tal.

	antal knækæg i bakken						
	0	1	2	3	4	5	>5
observeret	766	194	25	7	5	3	0
forventet	737.42	228.07	31.74	2.62	0.14	0.01	0.00

(a) Gør rede for, at den forventede antal bakker uden knækæg må være 737.42.

Forår 2018 (4a)

Hvis $X_i \sim \text{binom}(10, p)$, $i = 1, \ldots, 1000$, uafhængige, så er

skøn over p: $\sum_i X_i \sim \text{binom}(10 \cdot 1000, p)$

$\hat{p} = \frac{\sum_i x_i}{10000} = \frac{194+50+21+20+15}{10000} = \frac{300}{10000} = 0.03$ (MSRR prop 6.1.1)

Forventede antal: 1000 gange ss for at falde i kasse "0" i binomial med $p = 0.03$: $1000 * \text{dbinom}(0, 10, phat) = 737.4241$

```
a=c(766,194,25,7,5,3); j=c(0:5)
x=sum(a*j)
phat=x/10000
c(x,phat,1000*dbinom(0,10,phat))
[1] 300.0000   0.0300 737.4241
```

Forår 2018 (4b)

(b) Lav et goodness of fit test for hypotesen, at antal knækæg er binomialfordelt.

Webbog Resutat 1.1:

$$(A_1 \ldots, A_7) \sim \text{multinom}(1000, (\pi_1 \ldots, \pi_7)),\ M_0 :\ \pi \text{ vilkårlig}$$

teste hypotesen at $\pi_j = \text{dbinom}(j-1, 10, p)$, $j = 1, \ldots, 6$,
$\pi_7 = 1 - \pi_1 - \cdots - \pi_6$

For at få alle de forventede >=5 må vi slå de sidste 5 kasser sammen

```
                    antal knækæg i bakken
                0       1    >=2
observeret    766     194     40
forventet 737.42 228.07 34.51
```

$G = 7.28$, p-værdi $= 0.0069$ fra halen af en $\chi^2(3-1-1)$-fordeling.

Forår 2018 (4b)

```
a=c(766,194,40)
ex=c(737.42,228.07,34.51)
G=2*sum(a*log(a/ex))
c(G,1-pchisq(G,3-1-1))
[1] 7.2878115 0.0069424
```

(c) Fortolk resultatet af hypotesetestet.

Data tyder ikke på en binomialfordeling. Når vi sammenligner de observerede med de forventede kan vi se at der er observeret "for få" med 1 knækæg.

Dette kan tyde på manglende uafhængighed: når først der er 1 knækæg, er der også en sandsynlighed for at denne hændelse fører til endnu et knækæg.

Forår 2018, opgave 5

Forår 2018 opgave 5

Skal cykler til mænd have en anden geometri end kvindecykler? Eller: kører man som kort mand ganske bekvemt på en stor kvindecykel? Hvis ja, så burde forholdet mellem kropslængde og benlængde være ens hos begge køn. Dette undersøges i nærværende opgave.

For 1986 kvindelige og 4082 mandlige medlemmer af US army blev d er i årene 2010/2011 taget mål. Datasættet *anthro* til denne opgave består af en tilfældig delstikprøve på 200 mænd og 200 kvinder. Det indeholder faktorvariablen *koen* samt de numeriske variabler *benlaengde* og *kropslaengde*. Længder er angivet i mm. Figur 1 nedenfor viser benlængde plottet mod kropslængde, med forskellige symboler afhængigt af køn. Tilsyneladende kan data godt beskrives ved en simpel lineær regressionsmodel. Regressionslinjer er indtegnet separat for mænd og kvinder.

Vi starter med modellen

$$M_0: \; Y_{ij} \sim N(\alpha_i + \gamma_i x_{ij}, \sigma_i^2), \; i = 1, 2, \; j = 1, \ldots, 200,$$

hvor Y_{ij} er benlængde og x_{ij} er kropslængde, og $i = 1, 2$ er køn ($i = 1$ kvinde, $i = 2$ mand).

Forår 2018 opgave 5

Til besvarelse af de følgende spørgsmål kan man bruge relevante resultater fra R-udskrifterne nedenfor.

(a) Gør rede for, at modellen M_0 passer til data.

(b) Vis ved et test, at der kan antages, at de to varianser i M_0 er ens.

(c) Opskriv modellerne M_1, M_2, M_{3a} og M_{3b} fra R-protokol 3 i samme form som M_0 er givet.

(d) Vis ved et test, at der kan antages at hældningerne er ens i modellen med ens varianser.

(e) Vis at modellen ikke kan reduceres yderligere.

I opgaverne (f) og (g) betragtes modellen med to regressionslinier med fælles hældning.

(f) Angiv et 95%-konfidensinterval for forskellen i skæringen med andenaksen (intercept) for regressionslinjerne for mænd og kvinder.

(g) Angiv et estimat for den gennemsnitlige benlængde for en kvinde og for en mand, som

Forår 2018 (5a)

Vi har figur 2 til rådighed til at validere modellen.

Data viser et lineært forløb, og der er ikke noget der tyder på systematiske afvigelser i residualplottene

Det virker også som om at der er samme variation omkring linjen for forskellig e værdier af kropslængde i residualplottene

QQplots af residualer passer fint med en normalfordeling: snor sig om en ret linje

Forår 2018 (5b)

Teste hypotesen $\sigma_1^2 = \sigma_2^2$

Fra R-protokol 2 har vi variansskønnene $s_1 = 23.6309$ baseret på 198 frihedsgrader og $s_1 = 22.0628$ baseret på 198 frihedsgrader.

Vi laver test for ens varianser som i afsnit 2.12 i webbogen:

$F = \frac{23.6309^2}{22.0628^2} = 1.147$ som vurderes i en $F(198, 198)$-fordeling

P-værdi er 2 gange halesandsynligheden = 0.335. Da denne er langt over 0.05 accepterer vi ens varianser

```
F=23.6309^2/22.0628^2
c(F,2*(1-pf(F,198,198)))
[1] 1.1472003 0.3347622
```

Forår 2018 (5c)

```
M1 <- lm ( benlaengde ~ kropslaengde * koen , data = anthro )
M2 <- lm ( benlaengde ~ kropslaengde + koen , data = anthro )
M3a <- lm ( benlaengde ~ kropslaengde , data = anthro )
M3b <- lm ( benlaengde ~ koen , data = anthro )
```

$$M_1 : \; Y_{ij} \sim N(\alpha_i + \gamma_i x_{ij}, \sigma^2), \; (\alpha_1, \alpha_2, \gamma_1, \gamma_2) \text{ varierer frit}$$

$$M_2 : \; Y_{ij} \sim N(\alpha_i + \gamma x_{ij}, \sigma^2), \; (\alpha_1, \alpha_2, \gamma) \text{ varierer frit}$$

$$M_{3a} : \; Y_{ij} \sim N(\alpha + \gamma x_{ij}, \sigma^2), \; (\alpha, \gamma) \text{ varierer frit}$$

$$M_{3b} : \; Y_{ij} \sim N(\alpha_i, \sigma^2), \; (\alpha_1, \alpha_2) \text{ varierer frit}$$

1: hver sin linje, 2: hver sin skæring, fælles hældning, 3: fælles skæring og fælles hældning, 4: to grupper med hver sin middelværdi

Forår 2018 (5d+e)

Teste $\gamma_1 = \gamma_2$ i model M_1

Fra summary(M1) har vi linjen

```
kropslaengde:koenm 0.00065 0.00332 0.019 0.984
```

Denne linje vedrører forskellen i hældning: $\gamma_2 - \gamma_1$: t-testet giver en p-værdi på 0.984, som er langt over 0.05 hvorfor vi accepterer hypotesen om fælles hældning

Fra summary(M2) har vi linjerne

```
kropslaengde 0.61261 0.01657 36.971 <2e-16 ***
koenm -13.95712 3.08352 -4.526 7.94e-06 ***
```

Den første linje indeholder et t-test for at den fælles hældning er nul, $\gamma = 0$. Da p-værdien er langt under 0.05 afvises hypotesen

Den anden linje indeholder et t-test for at de to skæringer er ens $\alpha_1 = \alpha_2$. Da p-værdien er langt under 0.05 afvises hypotesen

Forår 2018 (5d+e)

```
(Intercept) -215.87526 27.12834 -7.958 1.86e-14 ***
koenm -13.95712 3.08352 -4.526 7.94e-06 ***
kropslaengde 0.61261 0.01657 36.971 <2e-16 ***
Residual standard error :  22.8315 on 397 degrees of
freedom
```

KI for forskel i skæring:
$-13.95712 \pm t_{inv}(0.975, 397) \cdot 3.08352 = [-20.0, -7.9]$
(Resultat 4.6 i webbog)

Kvindelængde, mandelængde af ben ved højde 1700 (middelværdi):
$\theta_1 = \alpha_1 + 1700 \cdot \gamma, \ \theta_2 = \alpha_2 + 1700 \cdot \gamma$

Skøn: $\hat{\theta}_1 = -215.875 + 1700 \cdot 0.61261 = 825.56,$
Skøn: $\hat{\theta}_2 = -215.875 - 13.95712 + 1700 \cdot 0.61261 = 811.60$
Ved samme højde har mænd i middel kortere ben

Forår 2018 (5d+e)

```
-13.95712+c(-1,1)*qt(0.975,397)*3.08352
[1] -20.019189  -7.895051

-215.875+c(0,-13.95712)+1700*0.61261
[1] 825.5620 811.6049
```

Sceneskift

Reeksamen 2018, opgave 1

Reeksamen 2018, opgave 1

Lad $X_1, \ldots, X_n$ være i.i.d kontinuert fordelte stokastiske variable, med tæthed givet ved

$$f(x, \theta) = \begin{cases} \theta \cos(\theta x), & 0 \leq x \leq \pi/(2\theta) \\ 0 & \text{ellers} \end{cases},$$

hvor $\theta > 0$ er en ukendt parameter.

Reeksamen 2018 (1a)

(a) Vis, at $EX = (\pi/2 - 1)/\theta$.

$$E(X) = \int_0^{\pi/(2\theta)} x\theta\cos(\theta x)dx = \int_0^{\pi/2}(y/\theta)\cos(y)dy$$

$$= [(y/\theta)\sin(y)]_0^{\pi/2} - \int_0^{\pi/2}(1/\theta)\sin(y)dy$$

$$= \pi/(2\theta) - [(1/\theta)(-\cos(y))]_0^{\pi/2} = (\pi/2 - 1)/\theta$$

Reeksamen 2018 (1b)

(b) Brug moment metoden til at udlede en estimator for θ, og beregn så estimatet fra den observerede stikprøve $(x_1, \ldots, x_5) = (24, 32, 17, 22, 37)$.

Ved momentmetoden sætter man den teoretiske middelværdi lig med gennemsnittet af observationerne.

$$(\pi/2 - 1)/\theta = \bar{x} \Rightarrow \hat{\theta} = \frac{\pi/2-1}{\bar{x}} = 0.02162$$

```
x=c(24,32,17,22,37)
(pi/2-1)/mean(x)
[1] 0.02162107
```

(c) Er din estimator unbiased? Giv en kort begrundelse for dit svar.

Unbiased betyder $E(\hat{\theta}) = \theta$
I vores tilfælde skal vi altså undersøge, om $E((\pi/2 - 1)/\bar{X}) = \theta$?

Fra tætheden ser vi at $Y_i = \theta X_i$ har tæthed $\cos(y)$ for y i intervallet $[0, \pi/2]$. Vi skal derfor undersøge om $E(1/\bar{Y}) = 1/(\pi/2 - 1)$?

Jensens ulighed giver $E(1/\bar{Y}) > 1/E(\bar{Y})$ hvorfor $E(1/\bar{Y}) > 1/(\pi/2 - 1)$

Altså er $\hat{\theta}$ ikke unbiased

Sceneskift

Reeksamen 2018, opgave 2

Reeksamen 2018, opgave 2

Der er blevet simuleret en stikprøve af størrelse 200 fra hver af tre forskellige fordelinger. Figuren til højre viser normalfraktilplottet for en af stikprøverne; nedenstående figur viser den empiriske fordelingsfunktion for alle tre.

Reeksamen 2018 (2a)

(a) Hvilken af de tre fordelingsfunktioner A, B, eller C hører til den viste normalfraktilplot? Giv et kort argument for dit svar.

Medianen aflæses i fraktilplot til cirka 0.25 (ved at se hvor den lodrette linje gennem nul skærer data).

Data A og B kan stemme med dette (ser hvor vandrette linje gennem 0.5 skærer data).

Vi betragter nu de fire øverste punkter i fraktilplot. Ved aflæsning svarer disse cirka til data 0.8, 0.85, 0.9 og 1. Dette stemmer kun med data B.

Sceneskift

Reeksamen 2018, opgave 3

Reeksamen 2018, opgave 3

Hurtigløb på skøjter afvikles parvis. En af løberne starter på den indre bane, mens den anden starter på den ydre bane. Nogle mener, at det giver en fordel at starte på den ydre bane. I nærværende opgave skal denne påstand undersøges med data fra 1500m løb fra vinter OL 2002. I alt startede 23 par i denne disciplin. Af disse løb blev 15 vundet af den løber, der startede på den ydre bane. Nulhypotesen, at sandsynligheden for at vinde er det samme på begge startbaner, skal undersøges mod alternativet, at den ydre bane giver en fordel.

Reeksamen 2018 (3a)

(a) Formuler nul- og alternativhypotesen i en binomialmodel for antallet X af løb, der bliver vundet af løberen på yderbanen.

Model: $X \sim \text{Binom}(23, p)$, $0 \leq p \leq 1$, hvor p er sandsynligheden for at vinde ved start på yderbane.

Nulhypotesen om ingen fordel er $p = \frac{1}{2}$.

Alternativhypotesen om fordel er $p > \frac{1}{2}$.

Reeksamen 2018 (3b)

(b) Beregn p-værdien for det observerede antal $x = 15$.

P-værdien er sandsynligheden for at få en værdi større end eller lig med 15 i en Binom$(23, 0.5)$-fordeling

1-pbinom(14,23,0.5) giver 0.1050198.

Da p-værdien er større end 0.05 siger vi, at data ikke strider mod hypotesen om ingen fordel.

Reeksamen 2018 (3c+d+e)

(c) Angiv den kritiske værdi svarende til signifikansniveauet $\alpha = 0.01$, og angiv hvornår H_0 forkastes.

```
round(1-pbinom(c(15:23),23,0.5),4)
[1] 0.0466 0.0173 0.0053 0.0013 0.0002 0.0000 0.0000 0.0000 0.0000
```

Vi ser at vi skal forkaste for $X \geq 18$ for at sandsynligheden for at forkaste er ≤ 0.01, men så tæt på 0.01 som muligt (nemlig 0.0053).

Reeksamen 2018 (3d+e)

(d) Beregn sandsynligheden for en type 1 fejl, i testet der bygger på den kritiske værdifra (c).

Sandsynligheden for en type 1 fejl er sandsynligheden for at forkaste under nulhypotesen, som er 0.0053.

(e) Beregn sandsynligheden for en type 2 fejl, i testet der bygger på den kritiske værdi fra (c), når sandsynligheden for at vinde ved start på yderbanen er p = 0.75.

Sandsynligheden for en type 2 fejl er sandsynligheden for ikke at forkaste under alternativet $p = 0.75$.
Denne beregnes som $P(X \leq 17)$. I R pbinom(17,23,0.75) som giver 0.5315

Sceneskift

Reeksamen 2018, opgave 4 (regnet 27/4)

Reeksamen 2018, opgave 4

Wafers er siliciumskiver, der danner råmateriale til elektroniske chips. Wafers produceres i renrum, da hver mikropartikel, der lander på overfladen, mindsker udbyttet af chippene. Det kan dog ikke helt undgås, at wafers kontamineres af partikler. Når man betragter antallet af partikler på en wafer, antages tit, at denne er Poisson fordelt. Er denne antagelse altid rimelig? Tabellen nedenunder viser observerede partikelantal for 100 små kvadratiske (2cm x 2cm) udsnit på wafers, samt det forventede antal beregnet fra en Poisson fordeling fittet til data.

	antal partikler per kvadratisk udsnit										
	0	1	2	3	4	5	6	7	8	9	>=10
observeret	25	24	19	16	6	4	1	2	2	1	0
forventet	13.67	27.20	27.07	17.95	8.93	3.55	1.18	0.34	0.08	0.02	0.00

(a) Gør rede for, hvordan det forventede antal udsnit uden partikler er blevet estimeret til 13.67.

(b) Lav et goodness of fit test for hypotesen, at antal partikler per kvadrat er Poisson fordelt. Benyt signifikansniveauet $\alpha = 0.05$.

(c) Fortolk resultatet af hypotesetestet. Tyder data på, at partiklerne sætter sig på overfladen af en wafer uafængigt af hinanden? Hvis nej, hvordan afviger observationerne fra det, man ville forvente?

Reeksamen 2018 (4a)

Lad $(A_1, \ldots, A_{11})$ være antallene i de 11 kasser. Jeg bruger modellen (M_0)

$(A_1, \ldots, A_{11}) \sim \text{multinom}(100, (\pi_1, ldots, \pi_{11})$

hvor $\pi_j \geq 0$ og $\pi_1 + \cdots + \pi_{11} = 1$.

Jeg vil teste hypotesen $\pi_j = \text{dpois}(j-1, \lambda)$, $j = 1, \ldots, 10$,

$\pi_{11} = 1 - \pi_1 - \cdots - \pi_{10}$, $\lambda \geq 0$.

(a) Som skøn over λ benytter jeg gennemsnittet af de 100 værdier (MSRR porposition 6.1.2). Denne beregnes nedenfor til $\hat{\lambda} = 1.99$. Det forventede antal i kasse 1 er derfor $100 \cdot \text{dpois}(0, 1.99) = 13.6695$

Reeksamen 2018 (4b+c)

(b) For at lave goodness of fit test, og have alle de forventede ≥ 5, slår jeg de sidste 6 kasser sammen. Den observerede værdi for disse bliver nu 10 og den forventede værdi bliver 5.18.

G-teststørrelsen fra Resultat 1.1 beregnes til 15.44, og p-værdien fra en $\chi^2(6-1-1)$-fordeling er 0.0039.

Da pværdien er langt under signifikansniveauet forkaster jeg hypotesen om at data stammer fra en poissonfordeling.

(c) Hvis partiklerne sætter sig tilfældigt uafhængigt af hinanden burde antallet i kvadraterne være poissonfordelt. Da vi forkaster poissonfordelingen holder denne antagelse ikke.

Den mest markante afvigelse er at der er for mange områder uden nogen partikler i forhold til hvad vi forventer. Der ser også ud til at være lidt for mange områder med rigtig mange partikler.

Reeksamen 2018 (4c)

```
a=c(25, 24, 19, 16, 6, 4, 1, 2, 2, 1, 0)
n=sum(a)
lambdahat=sum(a*c(0:10))/n

forventet=n*dpois(c(0:9),lambdahat)
forventet=c(forventet,n-sum(forventet))

forventet
 [1] 13.669542545 27.202389664 27.066377715 17.954030551  8.932130199
 [6]  3.554987819  1.179070960  0.335193030  0.083379266  0.018436082
[11]  0.004462168

lambdahat
[1] 1.99
```

Reeksamen 2018 (4c)

```
a1=c(a[1:5],sum(a[6:11]))
ex1=c(forventet[1:5],sum(forventet[6:11]))

rbind(a1,round(ex1,2))
    [,1] [,2]  [,3]  [,4] [,5]  [,6]
a1 25.00 24.0 19.00 16.00 6.00 10.00
   13.67 27.2 27.07 17.95 8.93  5.18

G=2*sum(a1*log(a1/ex1))

c(G1-pchisq(G,6-1-1))
[1] 15.437703709  0.003874401
```

Sceneskift

Reeksamen 2018 opgave 5

Reeksamen 2018 opgave 5

Påvirker grydens materiale jernindholdet af den tilberedte mad? Det er et vigtigt spørgsmål i afrikanske lande, hvor jernmangelanæmi, som den hyppigste form for fejlernæring, rammer ca. 50% af børn og kvinder, og 25% af mænd. Aluminiumsgryder, som er billige og lette, er ved at fortrænge de traditionelle jerngryder. Forskere har undersøgt, om jernindholdet i to forskellige traditionelle etiopiske måltider påvirkes af grydens materiale. Data indeholder målinger af jernindholdet i milligram per 100 gram mad, med 8 målinger for hver kombination af faktoren `food`, med niveauerne `meat` og `vegetables`, og faktoren `pot` med niveauerne `alu`, `clay` og `iron`. Data i denne opgave er simulerede baseret på informationen i artiklen *Effect of consumption of food cooked in iron pots on iron status and growth of young children: a randomised trial*, The Lancet 353, 712-716. (A.A. Adish et al. (1999)) Som udgangspunkt kan I antage at data kan beskrives med modellen

$$M_1: \; Y_{ijk} \sim N(\mu_{ij}, \sigma^2), \; i = 1, 2, \; j = 1, 2, 3, \; k = 1, \ldots, 8,$$

hvor $i = 1, 2$ står for maden (meat, vegetables) og $j = 1, 2, 3$ står for grydematerialet (alu, clay, iron), og Y_{ijk} er det målte jernindhold. Tabellen nedenunder gengiver gennemsnittene x_{ij} og stikprøvervarianserne s_{ij}^2, samt summen over begge dele.

Reeksamen 2018 opgave 5

mad	meat			vegetables			sum
gryde	alu	clay	iron	alu	clay	iron	
gennemsnit	2.43	2.64	4.34	1.44	1.48	3.10	15.46
varians	0.1983	0.1997	0.2271	0.1478	0.2193	0.1836	1.1758

Vink: delopgaverne (c) til (g) kan løses uden (a) og (b). Vi ser først på variansen.

(a) Gør rede for (giv et matematisk argument for), at $\mathrm{SSD}_E/\sigma^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} 7 \cdot s_{ij}^2/\sigma^2$ under model M_1 er et udfald af en stokastisk variabel som er χ^2-fordelt med 42 frihedsgrader.

Reeksamen 2018 (5a)

Vi antager at alle Y_{ijk} er uafhængige.

Fra MSRR B.10.5 (eller webbog afsnit 2.3) har vi at

$$7 \cdot s_{ij}^2/\sigma^2 \sim \chi^2(7) \text{ for alle grupper } i, j$$

Og fra MSRR B.10.2 har vi at summen over i og j af disse led følger en

$$\chi^2(6 \cdot 7)\text{-fordeling.}$$

Reeksamen 2018 (5b)

(b) Beregn et 95% konfidensinterval for σ^2.

Konfidensinterval for variansen står i webbog afsnit 2.6:

$$[\frac{df\,s^2}{\chi^2_{\text{inv}}(0.975,df)}, \frac{df\,s^2}{\chi^2_{\text{inv}}(0.025,df)}]$$

Vi benytter denne på $s^2 = \sum_{ij} 7s_{ij}^2/42 \sim \sigma^2\chi^2(42)/42$.

Beregningen i R giver intervallet $[0.133, 0.317]$

```
s2=7*1.1758/42
42*s2/c(qchisq(0.975,42),qchisq(0.025,42))
[1] 0.1332313 0.3165778
```

Reeksamen 2018 (5c)

(c) Undersøg ved en tegning, om den additive model M_2, givet ved

$$M_2: \ Y_{ijk} \sim N(\mu_{ij}, \sigma^2), \ \mu_{ij} = \mu + \beta_i^{\text{food}} + \beta_j^{\text{pot}}, \ \beta_1^{\text{food}} = 0, \ \beta_1^{\text{pot}} = 0$$

passer til data.

Vi bruger de opgivne gennemsnit til at lave et interactionplot. Figurerne viser kurver, der er tæt på at være parallelle. Dette tyder på at den additive model $\mu_{ij} = \eta_i + \zeta_j$ kan beskrive data.

```
me=c(2.43, 2.64, 4.34, 1.44, 1.48, 3.10)
par(mfrow=c(2,1))
plot(c(1,2,3,1,2,3),me)
lines(c(1,2,3),me[1:3])
lines(c(1,2,3),me[4:6],col=2)
plot(c(1,1,1,2,2,2),me)
lines(c(1,2),me[c(1,4)])
lines(c(1,2),me[c(2,5)],col=2)
lines(c(1,2),me[c(3,6)],col=4)
```

Reeksamen 2018 (5d)

(d) Til besvarelse af de følgende spørgsmål kan man bruge relevante resultater fra R-udskriften nedenfor. Data er gemt som en dataframe `irondata` med variablerne `iron`, `food` og `pot`.

Angiv formlen for μ_{ij} i modellerne M_{3a} og M_{3b} fra R-udskriften i samme form som givet for M_2 i delopgave (c). Angiv dernæst dimensionerne d_1, d_2, d_{3a} og d_{3b} af middelværdirummet i modellerne M_1, M_2, M_{3a} og M_{3b}.

Model M_{3a} siger, at middelværdien kun afhænger af hvilken foodgruppe man tilhører. Food svarer til index i i opgaven. Vi kan derfor skrive

$$\mu_{ij} = \mu + \beta_i^{\text{food}}, \; \beta_1^{\text{food}} = 0$$

Dimensionen er $d_{3a} = 2$ (2 niveauer for food)

Reeksamen 2018 (5d)

Model M_{3b} siger, at middelværdien kun afhænger af hvilken potgruppe man tilhører. Pot svarer til index j i opgaven. Vi kan derfor skrive

$$\mu_{ij} = \mu + \beta_j^{\text{pot}}, \; \beta_1^{\text{pot}} = 0$$

Dimensionen er $d_{3b} = 3$ (3 niveauer for pot)

Reeksamen 2018 (5d)

Model M_1 siger at alle 6 undergrupper givet ved food*pot har sin egen middelværdi, hvorfor $d_1 = 6$

Model M_2 er den additive model, hvorfor $d_1 = 1 + (2-1) + (3-1) = 4$, jævnfør parametriseringen i spørgsmål (c)

Reeksamen 2018 (5e)

(e) Vis ved et test, at model M_1 kan reduceres til model M_2.

Her må vi selv konstruere F-testet ud fra output i opgaven.

Formlen er givet i webbog afsnit 4.7:

(SSD2-SSD1)/(df2-df1)/(SSD1/df1)

Og p-værdien findes fra en

F(df2-df1,df1)-fordeling (store værdier er kritiske).

Vi finder SSD(M) ud fra s(M) i output:

$\text{SSD(M)}=\text{df(M)}*\text{s(M)}^2$.

Reeksamen 2018 (5e)

Fra R-beregningen finder vi F=0.283 og p-værdien er 0.755.

Da p-værdien er langt over 0.05 strider data ikke mod hypotesen om additivitet, det vil sige vi kan lave reduktionen fra model M_1 til model M_2.

```
df1=42
SSD1=0.442699^2*df1
df2=44
SSD2=0.435428^2*df2
F=(SSD2-SSD1)/(df2-df1)/(SSD1/df1)
c(F,1-pf(F,df2-df1,df1))
[1] 0.2832675 0.7547460
```

Reeksamen 2018 (5f)

(f) Undersøg (under model M_2), om grydematerialet har en indflydelse på jernindholdet af maden. Formuler en nulhypotese og lav et test.

Under modellen $\mu_{ij} = \mu + \beta_i^{\text{food}} + \beta_j^{\text{pot}}$ vil vi teste hypotesen

$$\beta_1^{\text{pot}} = \beta_2^{\text{pot}} = \beta_3^{\text{pot}} = 0.$$

Igen kan vi lave et F-test.

R-kørslen viser en meget lav p-værdi, hvorfor data strider mod ingen effekt af gryde.

```
df3a=46
SSD3a=0.941094^2*df3a
F=(SSD3a-SSD2)/(df3a-df2)/(SSD2/df2)
c(F,1-pf(F,df3a-df2,df2))
[1] 8.543880e+01 6.661338e-16
```

Reeksamen 2018 (5g)

(g) Beregn (under M_2) et 95% konfidensinterval for stigningen i jernindholdet i en ret, når den er kogt i jerngryde frem for aluminiumsgryde.

Vi aflæser skøn over $\mu_{\text{iron}} - \mu_{\text{alu}}$ til 1.8018 (linjen potiron under M2) og den tilhørende standard error er 0.1539.

95%-Konfidensintervallet beregnes fra t-fordelingen med 44 frihedsgrader som

$$1.8018 \pm t_0 \cdot 0.1539, \text{ hvor } t_0 \text{ er 97.5\%-fraktilen i en t(44)-fordeling.}$$

Dette giver intervallet $[1.49, 2.11]$. Forskel afhænger ikke af niveau for food-faktoren.

```
1.8018+c(-1,1)*qt(0.975,44)*0.1539
[1] 1.491635 2.111965
```

Sceneskift

Forår 2019 opgave 1 (regnet 27/4)

Forår 2019 opgave 1

Efter det sidste Europavalg i 2014 var 5 af de valgte medlemmer fra Danmark kvinder, og 8 var mænd; i Sverige var 11 ud af 20 medlemmer kvinder og i Finland var det 7 ud af 13. Undersøg ved et test, om kvindernes chancer for at blive medlem i Europaparlamentet er den samme i de tre lande. Giv korte svar:

(a) Hvilket test bruger du, og hvilken fordeling antages for teststørrelsen under nul hypotesen?

(b) Angiv teststørrelsen og p-værdien.

(c) Hvad er den faglige konklusion?

Besvarelse (a)

Lad K_{DK}, K_S og K_F være antal valgte kvinder i de tre lande. Jeg bruger modellen

$$\begin{aligned} K_{DK} &\sim \text{binom}(13, p_{DK}) \\ K_S &\sim \text{binom}(20, p_S) \\ K_F &\sim \text{binom}(13, p_F) \end{aligned}$$

og vil teste hypotesen $p_{DK} = p_S = p_F$

For at bruge Resultat 1.4 betragter jeg multinomialfordelinger:

$$(K_{DK}, 13 - K_{DK}) \sim \text{binom}(13, (p_{DK}, 1 - p_{DK}))$$

med tilsvarende for de to andre lande

(a) Som teststørrelse bruger jeg G fra Resultat 1.4, og den approksimative fordeling er $\chi^2((3-1)(2-1)) = \chi^2(2)$ under forudsætning om at alle forventede er ≥ 5

Besvarelse (b)

(b) Her kommer R-kørsel:

```
obs=rbind(c(5,8),c(11,9),c(7,6))
ex=outer(rowSums(obs),colSums(obs))/sum(obs)
gTest=2*sum(obs*log(obs/ex))
pval=1-pchisq(gTest,(dim(obs)[1]-1)*(dim(obs)[2]-1))
list(Forventede=ex,G=gTest,Pvaerdi=pval)

$Forventede
     [,1] [,2]
[1,]  6.5  6.5
[2,] 10.0 10.0
[3,]  6.5  6.5

$G
[1] 0.975921

$Pvaerdi
[1] 0.6138771
```

Besvarelse (b+c)

(b) Teststørrelsen er $G = 0.976$ og p-værdien er 0.62

(c) Data strider ikke mod at antage samme andel af kinder i de tre lande

Besvarelse: bonus

Bonusspørgsmål: Angiv, under antagelse af samme sandsynlighed for kvinde i de tre lande et 95%-konfidensinterval for sandsynligheden

Mode $K \sim \text{binom}(46, p)$, K er antal kvinder i de tre lande, observeret til 23
Formel for konfidensintervallet står i afsnit 1.2 (skjulte punkt) i webbogen og kan i R beregnes som

```
prop.test(23,46)$conf.int
```

```
[1] 0.3611894 0.6388106
attr(,"conf.level")
[1] 0.95
```

Konfidensintervallet er således $[0.36, 0.64]$. Intervallet er bredt, da vi kun har observeret $n = 46$. Intervallet indeholder værdien 0.5 svarende til at der er lige stor andel af mænd som kvinder.

Sceneskift

Forår 2019 opgave 2

Trykstyrke er en af de vigtigste egenskaber ved beton. Datasættet betonstyrke.csv, som skal analyseres i denne opgave, stammer fra et forsøg vedrørende trykstyrke af en betonblanding. Ved forsøget blev der anvendt tre forskellige blandemaskiner og to forskellige trykstyrkemålere. For hver kombination af blandemaskine og trykstyrkemåler blev fem betonblokke fremstillet med den pågældende blandemaskine, og trykstyrken blev målt med den pågældende trykstyrkemåler. Variablerne indeholdt i datasættet er
styrke - trykstyrke, målt i N/mm^2,
blander - angiver de tre blandemaskiner,
maaler - angiver de to trykstyrkemalere, og
gruppe - angiver de seks grupper, dannet ved kombination af blandemaskine og trykstyrkemåler.

I det følgende kan det antages, at for hver af de seks grupper er trykstyrken normalfordelt. Obs: Opskriv undervejs formelt de modeller du bruger.

Forår 2019 (2a)

Opstil en statistisk model for disse trykstyrker. Vis ved et test, at det kan antages, at variansen af trykstyrken ikke afhænger af gruppen.

a) Vi har inddelt data efter to faktorer *B=blander* (B1/B2/B3) og *M=maaler* (a/b). Respons er *styrke*.

Model M_0: $\text{Styrke}_i \sim N(\mu_{B_i,M_i}, \sigma^2_{B_i,M_i})$, idet vi ifølge opgaven kan antage normalfordelte data.

Vi betragter hypotesen at alle varianserne er ens, det vil sige at $\sigma^2_{B1,a} = \sigma^2_{B1,b} = \sigma^2_{B2,a} = \sigma^2_{B2,b} = \sigma^2_{B3,a} = \sigma^2_{B3,b}$.

Til dette bruger vi Bartletts test (webbog afsnit 4.5).
Vi inddeler efter både B og M ved at benytte faktoren *gruppe* i datasættet.

Forår 2019 (2a)

Da der er 6 grupper bruges en χ^2-fordeling med 5 frihedsgrader til beregning af p-værdien. Store værdier er kritiske.

Fra R aflæses p-værdien til 0.23, hvorfor vi siger at data ikke strider mod hypotesen om samme varians.

```
dat=read.csv("betonstyrke.csv",header=TRUE)
styrke=dat$styrke
B=dat$blander
M=dat$maaler
gr=dat$gruppe

bartlett.test(styrke~gr)
Bartlett's K-squared = 6.8635, df = 5, p-value = 0.231
```

Forår 2019 (2b)

Vis ved et test, at det kan antages, at der er additiv virkning af blandemaskine og trykstyrkemåler på trykstyrken.

b) Vores model er nu M_1: $\text{styrke}_i \sim N(\mu_{B_i,M_i}, \sigma^2)$.
Inden for denne model ønsker vi at teste additivitet,
det vil sige en reduktion til modellen M_2: $\text{styrke}_i \sim N(\eta_{B_i} + \zeta_{M_i}, \sigma^2)$.

Til dette kan vi både lave grafisk kontrol og et F-test.

```
par(mfrow=c(2,2))
interaction.plot(B,M,styrke)
interaction.plot(M,B,styrke)
```

Kontrolplottene viser nogenlunde parallelle kurver, hvilket tyder på additivitet.

Forår 2019 (2b)

F-test (som i afsnit 4.7 i webbog):

```
anova(lm(styrke~B+M),lm(styrke~B*M))

Model 1: styrke ~ B + M
Model 2: styrke ~ B * M
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 1035.1
2     42 1027.4  2    7.7867 0.1592 0.8534
```

F-testet giver en p-værdi på 0.85 (fra en $F(2,42)$-fordeling, store værdier er kritiske), hvorfor vi accepterer hypotesen om additivitet.

Forår 2019 (2c)

Tag udgangspunkt i modellen med additiv virkning af blandemaskine og trykstyrkemåler - den model du endte med i delopgave (b). Vis ved et test, at der kan antages, at der ikke er virkning af blandemaskiner på trykstyrken.

Vi ønsker at teste reduktion til model M_3: styrke$_i \sim N(\zeta_{M_i}, \sigma^2)$, hvor blandemaskine ikke indgår i beskrivelsen af middelværdien.

Dette test udføres igen ved et F-test under brug af *anova* i R.

```
anova(lm(styrke~M),lm(styrke~B+M))

Model 1: styrke ~ M
Model 2: styrke ~ B + M
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     46 1065.6
2     44 1035.1  2    30.455 0.6473 0.5284
```

Da p-værdien fra en $F(2, 44)$-fordeling er 0.53 strider data ikke mod hypotesen om ingen effekt af blandemaskine.

Forår 2019 (2d)

Tag udgangspunkt i modellen du endte med i delopgave (c). Find estimat og 95%-konfidensinterval for forskellen i middelværdien af trykstyrken mellem de to trykstyrkemålere. Kan der antages, at der ikke er forskel på de to trykstyrkemålere?

d) Estimater i modellen M_3: $\text{styrke}_i \sim N(\zeta_{M_i}, \sigma^2)$ findes med *summary* og konfidensintervaller med *confint*.

```
lmUD=lm(styrke~M)
summary(lmUD)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.0458     0.9825  55.011  < 2e-16 ***
Mb           -5.8667     1.3894  -4.222 0.000113 ***
Residual standard error: 4.813 on 46 degrees of freedom
```

Skøn over forskel $\zeta_b - \zeta_a$ aflæses til -5.8667

Forår 2019 (2d)

```
confint(lmUD)
                 2.5 %    97.5 %
(Intercept) 52.068255 56.023411
Mb          -8.663384 -3.069949
```

Skøn over forskel $\zeta_b - \zeta_a$ aflæses til -5.8667, og det tilhørende 95%-konfidensinterval er $[-8.7, -3.1]$.

Dette interval ligger langt fra nul, hvorfor et test af hypotesen om $\zeta_b - \zeta_a = 0$ vil give en p-værdi under 5% (fra summary tabellen aflæses p-værdien til 0.000113).

Find 95%-konfidensinterval for residualvariansen i modellen du endte med i delopgave (c).

e) Fra summary for modellen M_3: styrke$_i \sim N(\zeta_{M_i}, \sigma^2)$ aflæses skøn over spredning til 4.813 og frihedsgrader til 46. Skøn over varians er derfor 4.813^2

Dette er et udfald fra en $\sigma^2\chi^2(46)/46$-fordeling med 46 frihedsgrader.

Konfidensinterval for variansen beregnes ud fra formlen i webbogen afsnit 2.6

$$[\frac{df\,s^2}{\chi^2_{\text{inv}}(0.975,df)}, \frac{df\,s^2}{\chi^2_{\text{inv}}(0.025,df)}],$$

hvor df er frihedsgraderne for variansskønet s^2.

Forår 2019 (2e)

```
46*4.813^2/c(qchisq(0.975,46),qchisq(0.025,46))
[1] 15.99586 36.54275
```

Fra R-udskrift ser vi at 95%-konfidensintervallet for variansen bliver $[16.0, 36.5]$.

Sceneskift

Eksamen 2019 opgave 3

(a) Nogle gange logaritmetransformeres observerede data for at disse kan beskrives med en normalfordeling. Figuren nedenunder viser til venstre et normalfraktilplot, som er lavet pa logaritmen af varigheden i dage, af sygehusophold for 55 borgere, der opsøgte skadestuen med akutte mavesmerter. Til højre ses boxplots af tre datasæt på 55 observationer hver. Et af de tre boxplots A, B eller C er for de oprindelige varighedsdata før logaritmetransformation. Angiv, hvilket boxplot der svarer til de oprindelige varighedsdata, og giv en kort begrundelse for dit valg.

Forår 2019 (3a)

Median for log-data aflæses til cirka 1.9 (svarer til 0 på førsteaksen).

Dette giver 6.7 for de ikke-transformerede data.

Dermed er det enten B elle C der er det rigtige boxplot.

Da 75%-fraktilen i en standard normalfordeling er 0.67, aflæser vi at 75%-fraktilen for log-data er cirka 2.4,

hvilket giver cirka 11 på ikke-transformerede data.

Dette viser at C er det rigtige boxplot.

Forår 2019 (3b)

Firmaet Trinamix udvikler 3D kameraer baseret pa afstandsmåling under brug af infrarødt lys. For hvert 3D punkt i billedet tages n målinger; så beregnes gennemsnittet $\bar{d}$ af afstanden. Spredningen pa en enkeltmåling er 0.34 mm. Hvor stor må n være for at opnå en spredning SE[$\bar{d}$] på maksimalt 0.05 mm?

Med SE menes spredning. Spredning på gennemsnit er $\sigma/\sqrt{n}$, hvor σ er spredning på en enkeltmåling.

Vi har $\sigma = 0.34$ og ønsker $\sigma/\sqrt{n} < 0.05$.

Dette giver $n > (0.34/0.05)^2 = 46.24$.

Det vil sige, at vi må forlange at n er større end 46.

Forår 2019 (3c)

Ved klimaopvarmning frygtes, at vi i fremtiden vil få flere voldsomme storme også i Danmark. For at undersøge, om dette allerede er ved at ske, har en klimaforsker optalt antallet af voldsomme storme i Danmark med middelvind på mindst 26.4 m/s i perioden 1999-2008 (X1) og i perioden 2009-2018 (X2) . Angiv en statistisk model for data, og formuler en nulhypotese og en alternativhypotese der kan bruges til at besvare den ovenfor omtalte frygt. Angiv så den faglige konklusion, når man ved et test af denne hypotese får en p-værdi på $p_{\mathrm{obs}} = 0.97$.

Vi betragter tilfældige ankomster i tid, hvorfor vi benytter modellen

$$X_1 \sim \mathrm{Poisson}(\lambda_1), \quad X_2 \sim \mathrm{Poisson}(\lambda_2)$$

Nulhypotesen er, at der ikke er sket en udvikling, $\lambda_1 = \lambda_2$.

Den alternative hypotese, i forhold til klimaforskerens frygt, er $\lambda_2 > \lambda_1$.

Ved en p-værdi på 0.97 er der ingen grund til at forkaste nulhypotesen (hvis

Sceneskift

Reeksamen 2019 opgave 1

Reeksamen 2019 opgave 1

Hvordan kan strømmen af cyklister pa en cykelsti modelleres? Byrådsmedlem Bent J. ville gerne finde ud af det og registrerede derfor med et stopur tiden, der gik mellem forbisusende cyklister nedad Langelandsgade ved busstoppestedet "Universitet" en lørdag morgen mellem kl. 10:00 og 10:10. I alt kom der 49 cyklister forbi, og den gennemsnitlige tid mellem to cyklister var 12.2 sekunder. I litteraturen anføres ofte, at eksponentialfordelingen er en god model til disse ventetider mellem cyklister. For at undersøge dette nærmere har Bent grupperet tiderne som vist i følgende tabel. Tabellen indeholder ogsa de forventede antal under eksponentialfordelingen med middelværdi 12.2.

Tabel 1: Antal ventetider som falder i et tidsinterval (sekunder)

Interval	0-2	2-4	4-6	6-10	10-15	15-25	>25
Observeret antal	6	9	7	9	6	7	5
Forventet antal	7.4	6.3	5.3	8.4	7.3	8.0	6.3

Reeksamen 2019 (1a)

(a) Gør rede for, at det forventede antal ventetider af en længde mellem 2 og 4 sekunder er 6.3, under modellen hvor tiderne følger en eksponentialfordeling. (JLJ: underforstået eksponentialfordeling med middelværdi 12.2)

a) Eksponentialfordelingen har tæthed $\lambda e^{-\lambda x}$, $x > 0$, hvor middelværdien er $1/\lambda$ og fordelingsfunktionen er $P(X \leq x) = 1 - e^{-\lambda x}$ (MSRR side 495).

Det forventede antal mellem 2 og 4 med $\hat{\lambda} = 1/12.2$ er

$$49 \cdot \{(1 - e^{-4/12.2}) - (1 - e^{-2/12.2})\} = 6.288669.$$

Reeksamen 2019 (1b)

(b) Lav et test for nulhypotesen, at data stammer fra en eksponentialfordeling. Beregn p-værdien, og angiv konklusionen.

b) Vi laver et goodness of fit test. De observerede antal i de 7 kasser betragtes som udfald fra en Multinom$(49, (\pi_1, \ldots, \pi_7))$ fordeling, med $\pi_j \geq 0$ og $\sum \pi_j = 1$.

Vi ønsker at teste hypotesen

$$\pi_1 = 1 - e^{-\lambda 2}, \quad \pi_2 = e^{-\lambda 2} - e^{-\lambda 4}, \quad \pi_3 = e^{-\lambda 4} - e^{-\lambda 6}, \quad \pi_4 = e^{-\lambda 6} - e^{-\lambda 10}$$
$$\pi_5 = e^{-\lambda 10} - e^{-\lambda 15}, \quad \pi_6 = e^{-\lambda 15} - e^{-\lambda 25}, \quad \pi_7 = e^{-\lambda 25},$$

hvor $\lambda > 0$ er en fri parameter.

Reeksamen 2019 (1b)

```
obs=c(6, 9, 7, 9, 6, 7, 5)
ex=c(7.4, 6.3, 5.3, 8.4, 7.3, 8.0, 6.3)
G=2*sum(obs*log(obs/ex))
c(G,1-pchisq(G,7-1-1))
[1] 2.5062857 0.7755481
```

Da alle de forventede er større end 5, bruger vi χ^2-approksimationen til fordelingen af $G = 2\sum_j x_j \ln(x_j/e_j)$, hvor x_j er det observerede antal i kasse j og e_j er det forventede antal. Metoden er beskrevet i Resultat 1.1 i webbogen.

Frihedsgraderne er antal kasser minus 1 minus antal frie parametre: 7-1-1. P-værdien er 0.775, og da denne er langt over 0.05, strider data ikke mod en beskrivelse med en eksponentialfordeling.

Sceneskift

Reeksamen 2019 opgave 2

Reeksamen 2019 opgave 2

Bageriet Sundbrød producerer rugbrødet "Motion Mette" i 1000 g pakker. Pakkernes vægt antages at være normalfordelt med middelværdi 1000 g og en standardafvigelse pa 20 g. Firmaet er sikker på, at middelværdien overholdes, men er bekymret for, at der er for meget variation i vægten - dette kunne give anledning til reklamationer. Derfor tages jævnligt tilfældige stikprøver pa 100 pakker, og den empiriske varians s^2 bestemmes for stikprøven. Hvis s^2 overstiger 500 g^2 tvivler bageriet på, at standardafvigelsen ligger på 20 g.

Reeksamen 2019 (2a)

(a) Opskriv modellen til pakkernes vægt formelt, og formuler nul- og alternativhypotesen i bageriets problemstilling. Angiv teststørrelsens fordeling under nulhypotesen.

a) Lad V_i være vægt af den i-te pakke, $i = 1, \dots, 100$. Vi benytter modellen $V_i \sim N(\mu, \sigma^2)$, uafhængige, μ of σ kan variere frit. Bageriet er interesseret i størrelsen af spredningen σ.

Nulhypotesen er $\sigma = 20$ og alternativet er $\sigma > 20$. Som teststørrelse bruges s^2, og bageriet vil gribe ind (forkaster hypotesen) hvis $s^2 > 500$.

Under hypotesen $\sigma = 20$ har vi $99 \cdot s^2/20^2 \sim \chi^2(99)$, (MSRR Theorem B.10.5, webbog Resultat 2.2)

Reeksamen 2019 (2b)

(b) Beregn sandsynligheden for type 1 fejl i firmaets test.

b) En type 1 fejl består i at forkaste når nulhypotesen er sand (MSRR Definition 8.1).

I vores tilfælde er det

$$P(99 \cdot \frac{s^2}{20^2} > \frac{99 \cdot 500}{20^2}) = 1 - \chi^2_{\mathrm{cdf}}(123.75, 99) = 0.046765$$

(beregnes i R som: 1-pchisq(99*500/400,99))

Sceneskift

Reeksamen 2019 opgave 3

Reeksamen 2019 opgave 3

I et eksperiment lavet pa en skole skulle undersøges, om lysfarven påvirker plantevækst. Atten bønner blev sat i potter med en bønne per potte, og potterne tilfældigt delt i to lige store grupper. Den ene gruppe blev udsat for rødt lys, og den anden for grønt lys. To uger efter spiring blev skudhøjden målt. Datasættet beans.csv, som udleveredes med digital eksamen, indeholder variablen growth med skudhøjden i cm, og variablen color med lysfarven.

Reeksamen 2019 (3a)

(a) Undersøg ved hjælp af et permutationstest, om der er forskel mellem de to grupper. Brug forskellen mellem medianerne som teststørrelse. Husk at angive den faglige konklusion.

Two sample permutation test er beskrevet side 54 i MSRR.

Blandt de 18 observationer vælges tilfældigt 9 som udgør den ene gruppe, og de resterende 9 udgør den anden gruppe.

Herefter beregnes den ønskede teststørrelse.

Dette gentages N gange, og det undersøges, hvor ofte man har fået noget der er større end eller lig med teststørrelsen baseret på de oprindelige data.

Da der i opgaven ikke bliver bedt om et ensidet test, bruger vi som teststørrelse den numeriske forskel mellem medianerne i de to grupper.

Reeksamen 2019 (3a)

Først indlæses data og teststørrelse baseret på data beregnes

```
dat=read.csv("beans.csv",header=TRUE)
farv=dat$color
vaekst=dat$growth
testval=abs(median(vaekst[farv=="red"])-
median(vaekst[farv=="green"]))
testval
[1] 8.8
```

Dernæst laver vi $N = 99999$ permutationer og beregner teststørrelsen hver gang.

```
N=10^5-1
res=numeric(N)
for (i in 1:N){
 index=sample(18,size=9,replace=FALSE)
 res[i]=abs(median(vaekst[index])-median(vaekst[-index]))
}
```

Reeksamen 2019 (3a)

Til sidst beregner vi permutations p-værdien, idet vi medtager de oprindelige data i beregningen

```
(sum(res>=testval)+1)/(N+1)
[1] 0.07874
```

Da vi får en permutations p-værdi på 0.079 siger vi, at data, baseret på det udførte test, ikke strider mod hypotesen om samme middelværdi i de to lysgrupper (med et signifikansniveau på 0.05).

Sceneskift

Reeksamen 2019 opgave 4

Reeksamen 2019 opgave 4

Når der skal bygges motorveje eller landeveje i Florida, inviterer Department of Transportation (DOT) forskellige byggeentreprenører til at afgive tilbud. Entreprenørene skal sende deres tilbud i en forseglet konvolut, og dem der byder den laveste pris vinder anlægskontrakten. Ingeniører som er ansat i DOT beregner altid et estimat på omkostningerne i forvejen, og i de fleste tilfælde passer den fint med den endelige pris. I 1970erne og '80erne kunne der dog registreres flere tilfælde, hvor byggeentreprenørerne havde manipuleret prisen opad ved hemmelige prisaftaler (price-fixing). Formålet med denne opgave er at finde en model, der forudsiger den endelige pris ud fra manipulationsforhold og ingeniørernes estimat. Det udleverede datasæta FloridaRoads.csv er indsamlet af Floridas justitsministerium i 1980erne og indeholder 235 observationer med fem variabler:

- cost: pris i 1000 US dollars,
- logcost: logaritmen til cost,
- DOTestimate: prisen som DOTs ingeniører har estimeret (i 1000 US$),
- logDOTestimate: logaritmen til DOTestimate,
- bid: en faktorvariabel med niveauer competitive og fixed. Niveauet fixed betyder at prisen var manipuleret. I alt var 50 af de 235 observationer manipuleret, mens 185 overholder reglerne.

Reeksamen 2019 opgave 4

Lad Y_{gj} betegne den endelige pris, og x_{gj} estimatet som ingeniørene beregnede, hvor $g = 1, 2$ betegner gruppen ifølge `bid` (1: competitive, 2: fixed), og $j = 1, \ldots, n_g$ nummererer observationen indenfor gruppen, $n_1 = 185$ og $n_2 = 50$. Som udgangspunkt betragter vi to modeller, henholdsvis for de oprindelige og for de log-transformerede data:

$$\tilde{M}_1 : \; Y_{gj} \sim N(\tilde{\alpha}_g + \tilde{\gamma}_g x_{gj}, \sigma^2)$$
$$M_1 : \; log(Ygj) \sim N(\alpha_g + \gamma_g \log(x_{gj}), \sigma^2)$$

Reeksamen 2019 (4a)

(a) Indlæs datasættet og fit modellen $\tilde{M}_1$. Beregn residualerne, og undersøg ved en grafisk analyse, om de er normalfordelt og om variansen af Y_{gj} afhænger af x_{gj}. Gør det samme med modellen M_1. Hvilken af de to modeller passer bedst til data?

De to modeller er angivet i opgaveteksten.

Der er tale om to grupper der har hver sin lineære sammenhæng mellem respons og forklarende variabel.

Modelformlen til denne type model er på formen $<$
"faktor+faktor*x", hvor x er den forklarende variabel.

Nedenfor analyseres de to modeller med "lm", residualerne aflæses og der laves henholdsvis residualplot og normal qqplot

Reeksamen 2019 (4a)

```
dat=read.csv("FloridaRoads.csv",header=TRUE)
lmTildeUD=lm(cost~bid+bid*DOTestimate,data=dat)
lmUD=lm(logcost~bid+bid*logDOTestimate,data=dat)

par(mfrow=c(2,2))
plot(dat$DOTestimate,lmTildeUD$residuals)
qqnorm(lmTildeUD$residuals,datax=TRUE)
plot(dat$logDOTestimate,lmUD$residuals)
qqnorm(lmUD$residuals,datax=TRUE)
```

Det er tydeligt, at når der ikke tages logaritme (model $\tilde{M}_1$, 2 øverste plots), så stiger variansen med DOTestimate, og residualerne ser ikke normalfordelte ud fra qqplottet (store udsving fra at de snor sig om en ret linje).

Omvendt, for de logaritmetransformerede data viser residualplot hverken systmatiske afvigelser eller ikke-konstant varians. Desuden ser residualerne normalfordelt ud fra qqplottet.

Reeksamen 2019 (4b)

(b) Opstil model M_2 for $\log(Y_{gj})$, hvor hældningen ikke afhænger af, om buddet var manipuleret (angiv som formel). Eftervis ved et test, at modellen M_1 kan reduceres til M_2. Fit modellen M_2 og angiv alle estimerede parameter eksplicit, inklusive variansen, med de betegnelser du brugte i formlen.

Model M_1 som modelformel: logcost~bid+bid*logDOTestimate.

Model M_2 som modelformel: logcost~bid+logDOTestimate,

det vil sige $\log(Y_{gj}) \sim N(\alpha_g + \beta \log(x_{gj}), \sigma^2)$.

Vi laver et F-test (webbog afsnit 4.7) for reduktion fra model M_1 til model M_2:

Reeksamen 2019 (4b)

```
lmFra=lm(logcost~bid+bid*logDOTestimate,data=dat)
lmTil=lm(logcost~bid+logDOTestimate,data=dat)
anova(lmTil,lmFra)
```

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    232 5.5562
2    231 5.5405  1  0.015641 0.6521 0.4202
```

F-teststørrelsen er 0.6521 og p-værdien fra en $F(1, 231)$-fordeling er 0.4202

Da p-værdien er større end 0.05 strider data ikke mod hypotesen om samme hældning i de to grupper.

(Samme test kan aflæses som t-test i summary(lmFra))

Reeksamen 2019 (4b)

Vi betragter modellen $\log(Y_{gj}) \sim N(\alpha_g + \beta \log(x_{gj}), \sigma^2)$ og skal lave parameterestimater for de parametre, der optræder.

Vi benytter summary og tilføjer "-1" i kaldet for at få estimaterne for de to skæringer.

```
lmUD=lm(logcost~bid+logDOTestimate-1,data=dat)
summary(lmUD)
                Estimate Std. Error t value Pr(>|t|)
bidcompetitive -0.146721   0.048456  -3.028  0.00274 **
bidfixed        0.069901   0.050107   1.395  0.16434
logDOTestimate  1.005407   0.007422 135.462  < 2e-16 ***
Residual standard error: 0.1548 on 232 degrees of freedom
```

Skøn over skæringerne er $\hat{\alpha}_1 = -0.1467$ og $\hat{\alpha}_2 = 0.0699$, hvor 1 er competitive og 2 er fixed.
Skøn over den fælles hældning er $\hat{\beta} = 1.0054$ og
skøn over spredningen σ er $\hat{\sigma} = 0.1548$ og dermed $\hat{\sigma}^2 = 0.1548^2 = 0.02396$.

Reeksamen 2019 (4c)

(c) Eftervis, at modellen M_2 ikke kan reduceres yderligere.

Vi betragter hypotesen at de to skæringer er ens under model M_2: $\alpha_1 = \alpha_2$.

```
lmTil=lm(logcost~bid+logDOTestimate,data=dat)
summary(lmTil)

               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.146721   0.048456  -3.028  0.00274 **
bidfixed        0.216622   0.024750   8.753 4.38e-16 ***
logDOTestimate  1.005407   0.007422 135.462  < 2e-16 ***
Residual standard error: 0.1548 on 232 degrees of freedom
```

Fra linjen *bidfixed* ser vi, at p-værdien i t-testet for $\alpha_1 = \alpha_2$ er meget lille (4.38e-16).
Vi kan derfor ikke reducere model M_2.
Vi ser også fra tabellen, at vi ikke kan sætte hældningen lig med nul (p-værdi mindre end 2e-16).

Reeksamen 2019 (4d)

(d) Beregn et 95%-konfidensinterval for den fælles hældning i M_2, og test, om hældningen kan antages at være lig med 1.

Konfidensintervaller for parametre i middelværdien beregnes ved hjælp af *confint* i R. Dette baserer sig på t-fordelingen.

```
lmTil=lm(logcost~bid+logDOTestimate,data=dat)
confint(lmTil)
```

```
                    2.5 %      97.5 %
(Intercept)    -0.2421900 -0.05125113
bidfixed        0.1678592  0.26538438
logDOTestimate  0.9907840  1.02003057
```

Konfidensintervallet for den fælles hældning findes under *logDOTestimate* i tabellen: $[0.991, 1.020]$.

Da værdien 1 ligger i dette interval strider data ikke mod hypotesen om at hældningen er 1 (proportionalitet mellem cost og DOTestimate på den oprindelige skala).

Reeksamen 2019 (4e)

(e) Brug modellen M_2 til at beregne et 90% prædiktionsinterval for prisen Y (i 1000 US\$), når ingeniørerne forudsiger omkostninger til at være 3 millioner US\$ og prisen ikke er manipuleret. (Vink: 3 mio = $3000 \cdot 1000$)

Prædiktionsintervaller beregnes med predict i R.

Man skal angive interval="prediction", og angive værdier for bid og logDOTestimate.

Man skal efterfølgende huske at prædiktionsintervallet er på en log-skala, og regne tilbage til oprindelige skala.

Reeksamen 2019 (4e)

```
lmTil=lm(logcost~bid+logDOTestimate,data=dat)
udP=predict(lmTil,data.frame(bid="competitive",logDOTestimate=log(3000)),
interval="prediction",level=0.90)

udP
      fit      lwr      upr
1 7.90294 7.645874 8.160006

exp(udP)
       fit      lwr      upr
1 2705.224 2091.997 3498.206
```

Prædiktionsintervallet (90%) på ikke-log skala er således $[2092, 3498]$

Sceneskift

Forår 2020 opgave 1

Forår 2020 opgave 1

Hydroxymethylfuran (HMF) er en substans, der opstår ved nedbrydning af fruktose. HMF findes blandt andet i honning; i frisk bihonning er koncentrationen forholdsvis lav. Erfaringen siger, at HMF koncentrationen stiger med lagring, og når honning varmes. Kunstig honning indeholder store mængder af HMF, derfor kan HMF koncentrationen bruges både som indikator af honningkvalitet og friskhed, og for at opspore honning, der sælges under falsk betegnelse. Forsker i fødevarekemi vil gerne finde ud af, om man allerede efter to måneder ser en stigning i HMF koncentrationen. Derfor målte de HMF koncentrationen i fjorten glas honning, frisk fra biavleren, og igen efter to måeders opbevaring ved stuetemperatur. Data er i filen honning.csv med søjlerne honeysample, der betegner honningglasset, HMFstart: koncentrationen i mg/kg ved første måling og HMFend: koncentrationen efter to måneder

Forår 2020 (1a)

(a) Lav et passende permutationstest for nulhypotesen, at den gennemsnitlige HMF koncentration ikke ændrer sig. Angiv verbalt den alternative hypotese, du tester imod. Husk at angive den faglige konklusion.

Lad X_{ij}, $i = 1, \ldots, 14$ $j = 1, 2$ være HMF, hvor j er tidspunkt.

Vi ønsker at se på forskel mellem de to tidspunkter og lader
$$D_i = X_{i2} - X_{i1} \text{ ("end" minus "start").}$$
Lad $\mu = E(D_i)$.

Vi ønsker at teste hypotesen $\mu = 0$ mod alternativet at $\mu > 0$.

Der er tale om en matched pair situation og vi laver permutationstest som på side 69 i MSRR.

Forår 2020 (1a)

Vi bruger gennemsnit af differenser som teststørrelse, og simulerer nye målingerne ved at bytte rundt på start og end (ganger plus eller minus 1 på)

Resultatet af R-kørsel giver en simuleret p-værdi på 0.0003.

Denne er meget lille (langt under 0.05):

data strider mod hypotesen om ingen forskel.

Forår 2020 (1a)

```
dat=read.csv("honning.csv",header=TRUE)
diff=dat[,3]-dat[,2]
observed=mean(diff)

N=10^4-1
res=rep(0,N)

for (i in 1:N){
 res[i]=mean(diff*sample(c(-1,1),14,replace=TRUE))
 }

c(observed,(1+sum(res>=observed))/(1+N))
[1] 0.9214286 0.0003000
```

Sceneskift

Forår 2020 opgave 2

Forår 2020 opgave 2

I en biologisk undersøgelse i det nordvestlige Grønland har man på 5 tidspunkter henover sommeren indsamlet i alt 25 muslinger (5 muslinger til hvert tidspunkt) og målt indholdet af protein (mg per gram tørvægt). Data er i filen musling.csv, der har tre søjler: Dato der angiver de fem tidspunkter, Tid der angiver tidspunkt som antal dage fra det første tidspunkt, og Protein, der angiver proteinindholdet.

Forår 2020 (2a)

(a) Opskriv modellen M_0, hvor proteinindholdet er normalfordelt, og hver gruppe (hvert tidspunkt) har sin egen middelværdi og varians på proteinindholdet. Undersøg, om det kan antages, at varianserne er ens.

Lad P_i være proteinindhold, D_i være dato
betegnes med 1,2,3,4,5 nedenfor
Lad T_i være tiden fra første dato, $i = 1 \ldots, 25$.

Model M_0 kan skrives som
$M_0: \; P_i \sim N(\mu_{D_i}, \sigma^2_{D_i})$
$(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5)$ kan variere frit.

Vi tester hypotesen
$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$
ved hjælp af et Bartletts test (webbog afsnit 4.5).

Forår 2020 (2a)

Teststørrelsen er Ba=2.1247 og p-værdien er 0.7128
beregnes fra en $\chi^2(5-1)$-fordeling store værdier er kritiske
Da p-værdi>0.05: data ikke strider mod hypotesen om samme varians.

```
dat=read.csv("musling.csv",header=TRUE)
D=dat[,1]
T=dat[,2]
P=dat[,3]

bartlett.test(P~D)
Bartlett's K-squared = 2.1247, df = 4, p-value = 0.7128

par(mfrow=c(1,2))
plot(T,P)
qqnorm(lm(P~D)$residuals) # D er en faktor
```

Forår 2020 (2b)

(b) Lav grafiske undersøgelser for at se på forholdet mellem de fem grupper og for at se, om det er rimeligt at beskrive proteinindholdet med en normalfordeling, hvor hver gruppe har sin egen middelværdi af proteinindholdet og varianserne er ens.

Forklar dit valg af grafiske metoder.

Figur med protein afsat mod tid, et qqplot af residualer fra model med fælles varians.

Kun 5 observationer i hver gruppe

- boxplot og qqplot for hver gruppe: lidt "overkill"
- qqplot af residualer fra model med fælles varians giver mening.

P mod T viser måske en svag stigning over tid.

QQplot ser fint ud således at det er rimeligt at sige at residualerne er normalfordelt.

Forår 2020 (2c)

(c) Opskriv modellen M_1, hvor proteinindholdet er normalfordelt, hver gruppe har sin egen middelværdi og alle grupperne har samme varians. Lav, under model M_1, et konfidensinterval for forskel i middelværdi mellem gruppe 1 og gruppe 5 (svarende til de to tidspunkter 6/7 og 9/9).

$$M_1: \; P_i \sim N(\mu_{D_i}, \sigma^2), \; i = 1, \ldots, 25,$$
$$(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \sigma) \text{ varierer frit}$$

Konfidensinteral for $\mu_5 - \mu_1$:

Konfidenintervallet baserer sig på et t-test som beskrevet i afsnit 4.8.
- Beregnes i R via confint
- må lave "Jul06" til det første niveau i faktoren D.

Fra R får vi 95%-intervallet $[19.891074, 107.06893] \approx [20, 107]$
- meget bredt
- tyder på forskel mellem første og sidste dato

Forår 2020 (2a)

```
D=relevel(D,"Jul06")
confint(lm(P~D))

                 2.5 %    97.5 %
(Intercept) 523.297975 584.94203
DAug11       -8.528926  78.64893
DAug27      -18.008926  69.16893
DJul27       -5.808926  81.36893
DSep09       19.891074 107.06893

# F-test for reduktion fra M1 til M2
anova(lm(P~T),lm(P~D))
Model 1: P ~ T
Model 2: P ~ D
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     23 25571
2     20 21833  3    3738.5 1.1415 0.3564
```

Forår 2020 (2d)

(d) Opskriv modellen M_2, hvor proteinindhold er normalfordelt og middelværdien afhænger lineært af tiden. Lav et F-test for reduktion fra model M_1 til model M_2.

$$M_2: \; P_i \sim N(\alpha + \beta T_i, \sigma^2), \; i = 1, \dots, 25,$$
(α, β, σ) kan variere frit.

Det generelle F-test: afsnit 4.7 i webbogen

fra anova i R: $F = 1.1415$ med p-værdi=0.3564

p-værdi fra højre hale af en $F(3, 20)$-fordeling

Da p-værdien er langt over 0.05:

data strider ikke mod en lineær sammenhæng med tid

Forår 2020 (2e)

(e) Lav, under model M_2, et konfidensinterval for forskel i middelværdi mellem to tidspunkter, der ligger 65 dage fra hinanden. Sammenlign med konfidensintervallet i spørgsmål (c)

Jeg kører confint på output fra lm for model M_2.
giver konfidensinteral for hældningen β
ønsker: konfideninterval for 65β
ganger grænserne fra konfidensinterval for β med 65.

95%-konfidenintervallet for β: fra R: $[0.1155, 1.3229]$
95%-konfidensintervallet for 65β: $[7.509, 85.991]$
Intervallet er meget bredt, uklart hvor stor stigningen er

Konfidensintervallet er rykket lidt nedad og er en anelse kortere her i forhold til konfidensintervallet fra spørgsmål (c) ([20,107]).

Forår 2020 (2a)

```
summary(lm(P~T))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 561.4706    12.1497  46.213   <2e-16 ***
T             0.7192     0.2918   2.465   0.0216 *
---
Residual standard error: 33.34 on 23 degrees of freedom
Multiple R-squared:  0.2089,    Adjusted R-squared:  0.1745
F-statistic: 6.074 on 1 and 23 DF,  p-value: 0.02162

confint(lm(P~T))
                 2.5 %     97.5 %
(Intercept) 536.337101 586.604109
T             0.115525   1.322946

65*c(0.115525,1.322946)
[1]  7.509125 85.991490
```

Sceneskift

Forår 2020 opgave 3 (regnet 29/4)

Forår 2020 opgave 3

Data i tabellen nedenfor viser for 200 kvinder, hvor mange der er blevet gravide efter første, andet og tredje forsøg med ivf-behandling, og hvor mange der ikke er blevet gravide i de tre forsøg (data er opdigtede).

	Første	Andet	Tredje	Ikke-gravide
Antal	29	21	21	129

(a) Opskriv en statistisk model for antallet af kvinder i de fire kategorier Første, Andet, Tredje og Ikke-gravide.
(b) Betragt hypotesen (fast-rate-hypotesen), at sandsynligheden for at blive gravid er den samme i hvert forsøg. Hvis θ er sandsynligheden for at blive gravid i det enkelte forsøg, er sandsynligehden for at blive gravid i det første forsøg ?, gravid i det andet forsøg $(1-\theta)\theta$, gravid i det tredje forsøg $(1-\theta)^2\theta$, og sandsynligheden for ikke at blive gravid $(1-\theta)^3$. Opstil likelihoodfunktionen under hypotesen og find et udtryk for $\hat{\theta}$. Eftervis, at for data i tabellen ovenfor er $\hat{\theta} = 0.1363$.

(c) Lav et test for fast-rate-hypotesen.

Forår 2020 (3a+b)

(a) Lad (a_1, a_2, a_3, a_4) være antallene i de fire kasser. Vi benytter modellen

$$(A_1, A_2, A_3, A_4) \sim \text{multinom}(200, (\pi_1, \pi_2, \pi_3, \pi_4))$$
$$\pi_j \geq 0, \ \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

(b) Likelihoodfunktionen under hypotesen er

$$L(\theta) = \binom{200}{a}\theta^{a_1}((1-\theta)\theta)^{a_2}((1-\theta)^2\theta)^{a_3}((1-\theta)^3)^{a_4}$$
$$= \binom{200}{a}\theta^{a_1+a_2+a_3}(1-\theta)^{a_2+2a_3+3a_4}$$

Denne har samme strukur som likelihoodfunktionen i en binomialmodel og fra bevis af Propsition 6.1.1 i MSRR får vi

$$\hat{\theta} = \frac{a_1+a_2+a_3}{(a_1+a_2+a_3)+(a_2+2a_3+3a_4)}$$

Med data indsat giver dette $\hat{\theta} = 0.1362764$ (se R-kørsel)

Forår 2020 (3c)

(c) Vi benytter G-testet fra Resultat 1.1 i webbogen. Fra R-kørsel får vi

> forventede: $27.26, 23.54, 20.33, 128.87$, alle ≥ 5, så vi kan bruge $\chi^2(4-1-1)$-fordelingen til beregning af p-værdi
>
> $G = 0.4159$ og p-værdien $= 0.8123$

Da p-værdien er langt over 0.05 siger vi, at data ikke strider mod hypotesen om fast rate.

Forår 2020 (3c)

```
a=c(29,21,21,129)
thetahat=sum(a[-4])/sum(a*c(1,2,3,3))
thetahat
[1] 0.1362764

ex=200*c(thetahat,thetahat*(1-thetahat),thetahat*(1-thetahat)^2,(1-thetahat
round(ex,2)
[1]  27.26  23.54  20.33 128.87

G=2*sum(a*log(a/ex))
c(G,1-pchisq(G,4-1-1))
[1] 0.4158731 0.8122586
```

Sceneskift

Forår 2020 opgave 4

Forår 2020 opgave 4a

Figuren nedenunder viser normalfraktilplot af en stikprøve $x_1, \ldots, x_{35}$ fra en normalfordeling $N(\mu, \sigma^2)$, samt indtegnet linje via qqline i R. Forklar kort, hvordan (eller om) du kan finde variansen σ^2 ud fra linjen i figuren. Angiv, på denne baggrund, hvilket udsagn bedst passer til plottet:

(A) $\sigma^2 = 1.0$
(B) $\sigma^2 = 0.7$
(C) $\sigma^2 = 0.5$
(D) $\sigma^2 = 2.0$
(E) $\sigma^2 = 4.0$
(F) ingen af de ovensående værdier passer

(G) det er umuligt at skønne variansen ud fra linjen i figuren

Forår 2020 opgave 4a

Webbog afsnit 2.2: punkterne snor sig i et qqplot omkring en ret linje

linje har hældning σ (hvis data er normalfordelt)

blot aflæse hældningen i figuren og kvadrere denne.

Jeg aflæser punterne (-2,0.5) og (2,3.3).

hældning $= frac 3.3 - 0.52 - (-2) = 0.7$

$0.7^2 = 0.49$. Jeg satser derfor på (C)

Forår 2020 opgave 4b

En statistiker undersøger et kompliceret hypotesetest, baseret på teststørrelsen T, ved hjælp af simulation. Testet forkaster nulhypotesen når T er mindre end eller lig med en kritisk værdi C. Statistikeren har gemt Nsim udfald af T, som var simuleret under nulmodellen, i en vektor i R med navn T0. Dernæst har statistikeren simuleret lige sa mange udfald under en bestemt alternativhypotese H_A og gemt disse i vektoren T1. Variablen crit indeholder den kritiske værdi. Angiv R-programlinjer der beregner skøn over sandsynligheden for type 1 fejl og type 2 fejl fra de simulerede data, og gemmer værdierne henholdsvis som type1fejl og type2fejl.

Forår 2020 opgave 4b

Type 1 fejl er når vi forkaster hypotesen, når hypotesen er sand.

Vi forkaster små værdier ifølge opgavetekst:

```
type1fejl=sum(T0<=crit)/Nsim
```

Type 2 fejl er når vi ikke forkaster hypotesen, når hypotesen er falsk

Vi forkaster små værdier ifølge opgavetekst:

```
type2fejl=sum(T1>crit)/Nsim
```

Sceneskift

MSRR 10.10 (regnet 29/4)

MSRR 10.10

I general survey (CSS2002) er personer spurgt om hvor glade de er og hvem de har stemt på til præsidentvalg i 2000.

	Bush	Gore	Nader
Not too happy	29	46	5
Pretty happy	245	233	17
Very happy	164	123	7

a) State the hypothesis of interest.
b) Perform the test using the chisq.test command in R.
c) R returns a warning message. Compute the expected counts for each cell to see why.
(JLJ b+c) Gennemfør G-test for hypotesen
d) Perform a permutation test of independence.
e) Perform the test using Fisher's exact test.

f) Compare the P-values for the three approaches. Would you come to the same conclusion regardless of which test you used?

MSRR 10.10 (a)

869 personer er blevet spurgt om deres grad af happiness (Not too happy/Pretty happy/Very happy = 1/2/3) og om hvem de har stemt på (Bush/Gore/Nader = 1/2/3). Lad

$$a_{ij} \text{ være antal blandt de 869 med svarene } (i,j),\ i,j = 1,2,3$$

Model: $(A_{11}, A_{12}, A_{13}, A_{21}, A_{22}, A_{23}, A_{31}, A_{32}, A_{33}) \sim$ multinom$(137, (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{31}, \pi_{32}), \pi_{33}))$, $\pi_{ij} \geq 0\ \sum_{ij} \pi_{ij} = 1$

(a) Vi ønsker at teste om svar på de to spørgsmål er uafhængige. Hypotesen er, at der eksisterer α, β således at

$$\pi_{ij} = \alpha_i \beta_j \ i,j = 1,2,3$$

hvor $\alpha_1 + \alpha_2 + \alpha_3 = 1$ og $\beta_1 + \beta_2 + \beta_3 = 1$

MSRR 10.10 (b+c)

(b+c)

Fra R-output set at C-teststørrelsen er 11.3 med approksimativ p-værdi på 0.023. Vi er dog usikre på denne p-værdi da en af cellerne har en forventet værdi på 2.7 under hypotesen om uafhængighed.

Vi vil derfor også være usikre på den approksimative pværdi for G-testet. G-teststørrelsen fra Resultat 1.4 er 11.15 og den approksimative p-værdi fra en $\chi^2(4)$-fordeling er 0.025

MSRR 10.10 (c)

```
obs=rbind(c(29,46,5),c(245,233,17),c(164,123,7))

# C-test
chisq.test(obs)
data:  obs
X-squared = 11.3, df = 4, p-value = 0.02339
Advarselsbesked:
I chisq.test(obs) : Chi-squared approximation may be incorrect

chisq.test(obs)$expected
          [,1]      [,2]      [,3]
[1,]  40.32221  37.00806  2.669735
[2,] 249.49367 228.98734 16.518987
[3,] 148.18412 136.00460  9.811277
```

MSRR 10.10 (c)

```
# G-test
ex=outer(rowSums(obs),colSums(obs))/sum(obs)
gTest=2*sum(obs*log(obs/ex))
pval=1-pchisq(gTest,(dim(obs)[1]-1)*(dim(obs)[2]-1))
list(Forventede=ex,G=gTest,Pvaerdi=pval)

$Forventede
          [,1]      [,2]      [,3]
[1,]  40.32221  37.00806  2.669735
[2,] 249.49367 228.98734 16.518987
[3,] 148.18412 136.00460  9.811277

$G
[1] 11.14512

$Pvaerdi
[1] 0.02498055
```

MSRR 10.10 (d+e+f)

(d) Vi bruger kode fra afsnit 1.8 i webbogen til beregning af simulerede (betingede) p-værdi. Denne bliver 0.030.

(e) Fra R-kørsel ses at Fishers eksakte test giver en p-værdi på 0.01934

(f) P-værdien fra Fishers eksakte test ligger lidt under den simulerede værdi på 0.030 (95%-konfideninterval 0.027-0.034) (bygger på samme betingede fordeling, men bruger forskellig teststørrelse). De to approksimative værdier ligner hinanden og er nærmest midt mellem den eksakte og den simulerede værdi. Alle værdier fortæller cirka den samme historie.

Med et signifikansniveau på 0.05 vil de forskellige test give samme resultat for data her.

MSRR 10.10 (f)

```
obs=rbind(c(29,46,5),c(245,233,17),c(164,123,7))

r=dim(obs)[1]; k=dim(obs)[2]
rs=rowSums(obs)
cs=colSums(obs)
h=rep(c(1:r),rs)
m=rep(c(1:k),cs)

Gfct=function(Amat){
ex=outer(rowSums(Amat),colSums(Amat))/sum(Amat)
A1=ifelse(Amat==0,1,Amat)
return(2*sum(Amat*log(A1/ex)))
}
Gobs=Gfct(obs)

nSim=10^4-1
tval=rep(0,nSim)
```

MSRR 10.10 (f)

```
for (i in 1:nSim){
 msamp=sample(m)
 tabelsamp=table(h,msamp)
 tval[i]=Gfct(tabelsamp)
}

pval=(1+sum(tval>=Gobs))/(1+nSim)
c(G=Gobs,pværdi=pval)
       G   pværdi
11.14512  0.03000

# Fishers test
fisher.test(obs)

data:  obs
p-value = 0.01934
```

Sceneskift

Opgave