Eksamen i Matematisk Statistik, F 2020

Vejledende besvarelse

Vær opmærksom på, at der nogle gange findes alternative rigtige svar, og de er ikke alle taget med i denne besvarelse.

Opgave 1

(1.a)

Vores nulhypotese er: der er ingen forskel i middelværdien af HMF indhold i starten og efter to måneder. Vi tester mod alternativen: HMF indhold er højere efter to måneder.

Det drejer sig her ikke om to uafhængige stikprøver, men om en forbunden stikprøve, og vi bruger derfor en matched pairs permutationstest.

```
honning <- read.csv ("honning.csv")
Diff <- honning$HMFend - honning$HMFstart
observed <- mean (Diff)
N <- 10^5 - 1
result <- numeric (N)
for (i in 1 : N)
{
  Sign <- sample (c (- 1, 1), 14, replace=TRUE)
  Diff2 <- Sign * Diff
  result[i] <- mean (Diff2)
}
pval <- (sum (result >= observed) + 1) / (N + 1)
pval
```

```
## [1] 0.00012
```

Konklusion: p-værdien (0.00012) ligger betydeligt under 0.05. Det giver anledning til at tvivle på nulhypotesen, og vi må forkaste hypotesen at HMF koncentrationen er den samme efter to måneder, til fordel for alternativen, at koncentrationen steg.

Opgave 2

(2.a)

Lad Y_i vær den i'te måling, $i = 1, \ldots, 25$, lad Tid_i være den i'te tid, og lad T_i være den i'te værdi af en faktor dannet ud fra Tid.

$$\text{Model } M_0 : Y_i \sim N(\mu_{T_i}, \sigma^2_{T_i}).$$

(Bemærkning: alternativ kan bruges dobbelt indeks notation: lad Y_{ij} være den j-te måling til tid i. $i = 1, \ldots, 5$, $j = 1, \ldots, 5$. Model M_0: $Y_{ij} \sim N(\mu_i, \sigma_i^2)$.)

Hypotesen om samme varians: $\sigma^2_{T_i} = \sigma^2$ kan undersøges ved et Bartlett test.

```
dat <- read.csv("musling.csv", header=TRUE)
y <- dat$Protein
Tid <- dat$Tid
T <- factor(Tid)
bartlett.test(y ~ T)
```

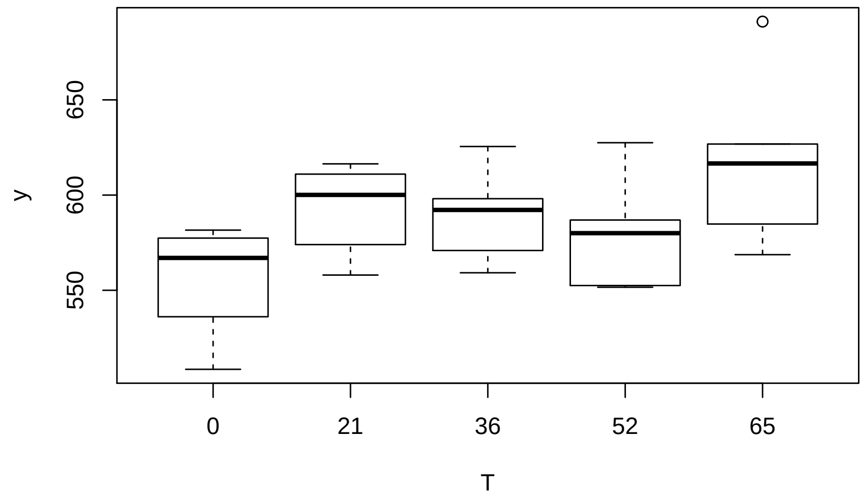
```
##
##  Bartlett test of homogeneity of variances
##
## data:  y by T
## Bartlett's K-squared = 2.1247, df = 4, p-value = 0.7128
```

Fra R-udskrift ses at p-værdien fra en $\chi^2(4)$-fordeling er 0.71, og vi konkluderer derfor at data ikke strider mod hypotesen om samme varians (p-værdi er langt over 0.05).

(2.b)

Jeg starter med et boxplot for at se på forholdet mellem de 5 grupper (tidspunkter)

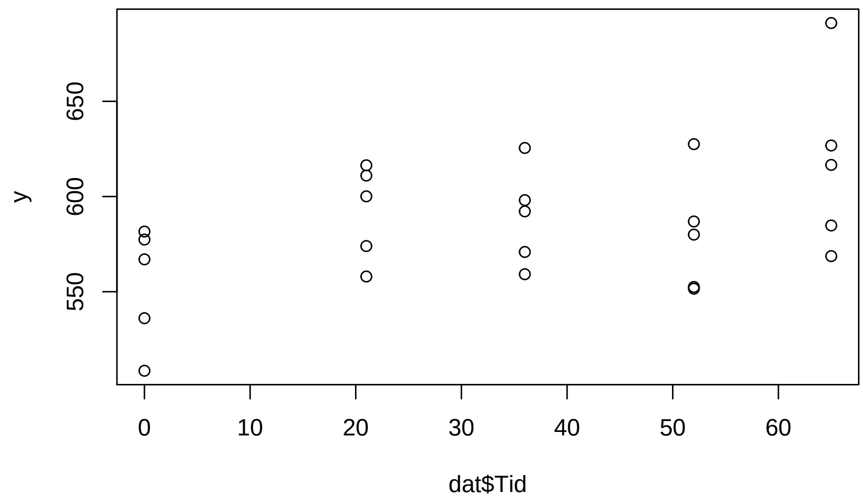
```
boxplot(y ~ T)
```



Man får en lille fornemmelse af at gruppe 1 ligger under de andre fire grupper, som til gengæld ser ret ens ud.

Men i vurderingen skal man huske at der kun er 5 observationer i hver gruppe. Så måske bedre at lave et plot af data direkte:

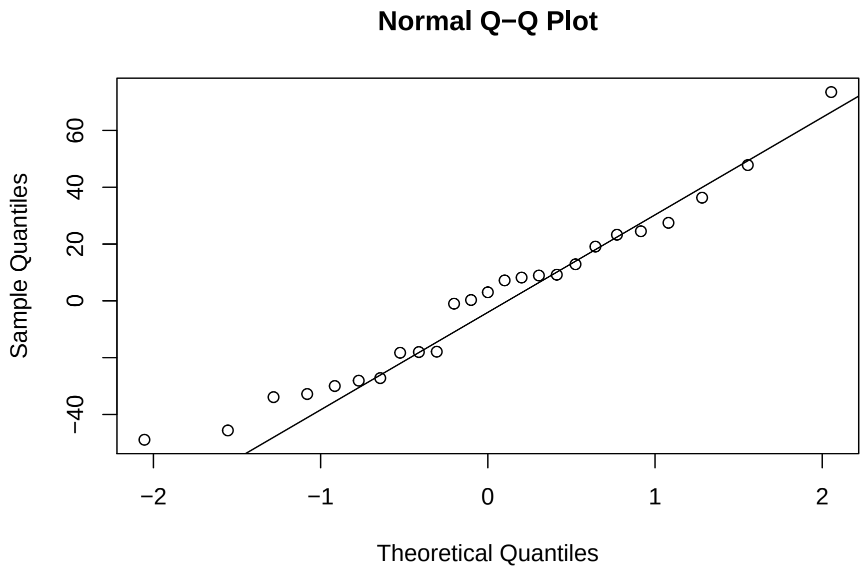
```
plot(dat$Tid, y)
```



Svært at afgøre om der er en stigende tendens eller at de alle har samme middelværdi.

Til sidst laver jeg et normalt-qqplot af residualerne fra modellen med fem grupper med hver sin middelværdi. Residualerne findes fra kørsel af lm.

```
r <- lm(y ~ T)$residuals
qqnorm(r)
qqline(r)
```



Fraktilsammenligningen giver ikke anledning til tvivl omkring normalitetsantagelsen.

(Bemærkning: det er nok at lave et boxplot og et normalt-qqplot, og det er det, de fleste har gjort. Der skal være kommentarer på plottene.)

(2.c)

Model M_1: $Y_i \sim N(\mu_{T_i}, \sigma^2)$.

Ved at køre confint(lm) får vi konfidensintervaller for forskel i middelværdi mellem en gruppe og den første gruppe. Med en valgte faktor hedderne niveauerne 0, 21, 36, 52 og 65.

```
confint(lm(y ~ T))
```

```
##                  2.5 %    97.5 %
## (Intercept) 523.297975 584.94203
## T21          -5.808926  81.36893
## T36          -8.528926  78.64893
## T52         -18.008926  69.16893
## T65          19.891074 107.06893
```

For forskel mellem den femte og den første gruppe $\mu_{65} - \mu_0$ aflæses konfidenintervallet til $[19.9, 107.1]$.

(2.d)

Model M_2: $Y_i \sim N(\alpha + \beta \cdot \text{Tid}_i, \sigma^2)$.

F-testet for reduktion fra model M_1 til M_2 laves med anova i R.

```
anova(lm(y ~ Tid), lm(y ~ T))
```

```
## Analysis of Variance Table
##
## Model 1: y ~ Tid
## Model 2: y ~ T
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     23 25571
## 2     20 21833  3    3738.5 1.1415 0.3564
```

Fra output ser vi at $F = 1.1415$ som skal vurderes i en $F(3, 20)$-fordeling. Store værdier er kritiske og p-værdien aflæses til 0.36. Data strider således ikke mod antagelsen om en lineær sammenhæng (p-værdi er over 0.05).

(2.e)

I model M_2 er middelværdien $\alpha + \beta \cdot \text{Tid}$ og forskel mellem to tidspunkter med afstand 65 bliver derfor $\beta \cdot 65$. Vi skal derfor lave et konfidensinterval for 65β, hvilket fås som 65 ganget med konfideninterallet for β. Dette findes igen med confint i R.

```
confint(lm(y ~ Tid))
```

```
##                  2.5 %     97.5 %
## (Intercept) 536.337101 586.604109
## Tid           0.115525   1.322946
```

Fra R-udskrift ses at det ønskede konfidensinterval er $[7.5, 86.0]$. Vi ser, at forskellen er meget ubestemt (bredt konfidensinterval), men dog ligger nul ikke i intervallet.

Konfidensintervallet er rykket lidt nedad og er en anelse kortere her i forhold til konfidensintervallet fra spørgsmål (c) ($[19.9, 107.1]$).

Opgave 3

(3.a)

Data (29,21,21,129) kan opfattes som er udfald fra en Multinom$(200, (\pi_1, \pi_2, \pi_3, \pi_4))$-fordeling.

Model M_0 er modellen, hvor $(\pi_1, \pi_2, \pi_3, \pi_4)$ kan variere frit: $\pi_j \geq 0$, $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$.

(3.b)

De fire antal betegnes med x_j, $j = 1, 2, 3, 4$.

Likelihoodfunktionen er

$$
\begin{aligned}
L(\theta; c_1, x_2, x_3, x_4) &= \binom{200}{x} \theta^{x_1} ((1-\theta)\theta)^{x_2} ((1-\theta)^2\theta)^{x_3} (1-\theta)^{3x_4} \\
&= \binom{200}{x} \theta^{x_1+x_2+x_3} (1-\theta)^{x_2+2x_3+3x_4}.
\end{aligned}
$$

Da dette ligner likelihoodfunktionen for en binomialfordeling med $x = x_1+x_2+x_3$ og $n = x_1 + x_2 + x_3 + x_2 + 2x_3 + 3x_4$ får vi fra øverst side 152 (MSRR) at $\hat{\theta} = (x_1 + x_2 + x_3)/(x_1 + 2x_2 + 3x_3 + 3x_4)$.

For vores data bliver dette 0.1363 som ses af R-udskrift.

```
x <- c(29, 21, 21, 129)
thetahat <- (x[1] + x[2] + x[3]) / (x[1] + 2*x[2] + 3*x[3] + 3*x[4])
thetahat
```

```
## [1] 0.1362764
```

(3.c)

Vi tester nu hypotesen $\pi_1 = \theta, \pi_2 = (1-\theta)\theta, \pi_3 = (1-\theta)^2\theta, \pi_4 = (1-\theta)^3$ under model M_0.

De forventede antal e_j findes ved at indsætte $\hat{\theta}$ og gange med 200. De forventede kan ses i R-udskrift.

Alle de forventede er større end 5 (den mindste er 20.3), hvorfor vi bruger χ^2-approksimationen til fordelingen af G-teststørrelsen, $G = 2\sum_j x_j \log(x_j/e_j)$.

Antallet af frihedsgrader er $4 - 1 - 1 = 2$ og store værdier af G er kirtiske. Da p-værdien er langt over 0.05 (pværdi = 0.81) strider data ikke mod fast-rate-hypotesen.

```
ex = 200*c(thetahat,(1-thetahat)*thetahat,
(1-thetahat)^2*thetahat,(1-thetahat)^3)
ex
```

```
## [1]  27.25528  23.54103  20.33294 128.87075
```

```
G = 2 * sum(x * log(x / ex))
c(G, 1 - pchisq(G, 4 - 1 - 1))
```

```
## [1] 0.4158731 0.8122586
```

Opgave 4

(4.a)

Hældningen af qqline i qqplot af en normalfordelt stikprøve er cirka éns med standardafvigelsen af den tilbundsliggende normalfordeling. Fra tegningen aflæses hældningen at være circa 0.7, altså $\sigma \approx 0.7$. Derfor passer svar (C) bedst, $\sigma^2 = 0.5$.

(4.b)

Sandsynligheder for fejlene skønnes som

```
type1fejl <- sum(T0 <= crit) / Nsim
type2fejl <- sum(T1 > crit) / Nsim
```