Eksamen Matematisk Statistik F 2024

7. Juni 2024

Opgave 1

(1.a)

79 ud af 500 jordskælv opfylder kriteriet:

```
prop.test(79,500, correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  79 out of 500, null probability 0.5
## X-squared = 233.93, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1286527 0.1925624
## sample estimates:
##     p
## 0.158
```

Her aflæses et 95% CI for sandsynligheden for, at intervallet mellem to jordskælv overstiger 35 dage, til at være [0.1286527,0.1925624].

(1.b)

Vi gør det ret simpelt. Vi beregner et estimat af sandsynligheden $P(X \geq 35)$ og ganger med antal observationer:

```
(1 - pgamma(35,shape=0.8557,rate=0.0449)) * 500
```

```
## [1] 82.65862
```

Her aflæses den forventede værdi til at være 82.65862, som naturligvis kan afrundes til 82.7.

(1.c)

Vi laver et C-test af vores observerede data ud fra de forventede værdier givet, at data stammer fra en gammafordeling (hvor parametrene er MLE):

```
observed <- c(96,47,64,70,67,77,79)
expected <- c(89.3, 62.9, 65.9, 63.1, 64.0, 72.1, 82.7)

k <- length(observed)
n <- sum(observed)

Cstatistic <- sum((observed - expected)^2 / expected)
c(C = Cstatistic, pval_approx = 1- pchisq(Cstatistic, df = k-2-1))
```

```
##           C pval_approx
##   5.9703939   0.2013702
```

Ved C-test fås en p-værdi på 0.2013702, så vi kan godt acceptere, at data stammer fra en gammafordeling. Vi har 4 frihedsgrader, da vi har estimeret to parametre i gammafordelingen.

(1.d)

Vi beregner et CI for r (hvor vi anvender estimatet for r som middelværdi):

```
r <- 0.8557
SE_r <- 0.0618

q <- c(0.025, 0.975)
qnorm(q, mean = r, sd = SE_r)
```

```
## [1] 0.7345742 0.9768258
```

Et 95% CI for r aflæses til at være [0.7345742, 0.9768258]

Vi ved, at $Exponential(\lambda) = Gamma(1, \lambda)$. Det betyder, at det giver mening at modellere data ved eksponentialfordelingen, hvis og kun hvis r med rimelighed kan antages at være lig 1.

Tallet 1 ligger uden for vores 95% CI for r, så vi kan ikke med rimelighed modellere dataene ved eksponentialfordelingen.

Opgave 2

(2.a)

Vi opskriver log likelihood og finder maksimum:

$$logL(X_1, ..X_n; \theta)) = \sum_{i=1}^{n} log(2\theta^2 X_i^3 exp(-\theta X_i^2))$$

$$= \sum_{i=1}^{n} (log(2X_i^3) + 2log\theta - \theta X_i^2)$$

Dette differentieres for at finde maksimum:

$$0 = \frac{d}{d\theta} logL(X_1, ...X_n; \theta) = \sum_{i=1}^{n} (\frac{2}{\theta} - X_i^2) = \frac{2n}{\theta} - \sum_{i=1}^{n} X_i^2$$

$$\Rightarrow \hat{\theta}_{mle} = \frac{2n}{\sum_{i=1}^{n} X_i^2}$$

Som er det vi skulle vise.

(2.b)

$\hat{\theta}_{mle}$ er biased. For at være unbiased, skulle $E[\hat{\theta}_{mle}] = \theta$ for alle $\theta > 0$. Men $E[\hat{\theta}_{mle}] = E[\frac{2n}{\sum_{i=1}^{n} X_i^2}]$, hvilket ikke nødvendigvis er lig θ.

For eksempel i tilfældet $\theta = 1, n = 1$ vil $E[\frac{2n}{\sum_{i=1}^{1} X_i^2}] = 2 \cdot \int_0^\infty \frac{1}{x^2} x \cdot 2x^3 e^{-x^2} dx = \sqrt{\pi} \neq 1 = \theta$ (integralet blev beregnet i Wolfram Alpha).

Dog er det værd at nævne, at jf. kap. 8.2.4 i ISS-bogen er MLE altid asymptotisk unbiased.

(2.c)

Vi løser ligningen $\bar{X} = EX$:

$$\bar{X} = EX = \frac{3\pi}{4\sqrt{\theta}} \Rightarrow \hat{\theta}_{mom} = \frac{9\pi^2}{16 \cdot (\bar{X})^2}$$

Så har vi fundet vores estimat. ### (2.d)

```
set.seed(123)
theta <- 5
n <- 10
Nsim <- 10000

shape <- 2
rate <- theta
```

```r
mle_estimates <-  numeric(Nsim)
moment_estimates <- numeric(Nsim)

mle_bias_estimates <-  numeric(Nsim)
moment_bias_estimates <- numeric(Nsim)


for (i in 1:Nsim) {
  Z_i <- rgamma(n, shape = shape, rate = rate)
  X_i <- sqrt(Z_i)

  mle_est_for_current_n <- 2 * n / (sum(X_i ^ 2))
  moment_est_for_current_n <- 9 * pi^2 / (16 * (mean(X_i) ^2))

  mle_estimates[i] <- mle_est_for_current_n
  moment_estimates[i] <- moment_est_for_current_n

  mle_bias_estimates[i] <- mle_est_for_current_n - theta
  moment_bias_estimates[i] <- moment_est_for_current_n - theta
}
simulated_mle_bias_est <- mean(mle_bias_estimates)
simulated_moment_bias_est <- mean(moment_bias_estimates)

simulated_mle_bias_est
```

```
## [1] 0.2485681
```

```r
simulated_moment_bias_est
```

```
## [1] 11.30892
```

```r
var(mle_bias_estimates) + simulated_mle_bias_est ^2
```

```
## [1] 1.54741
```

```r
var(moment_bias_estimates) + simulated_moment_bias_est ^2
```

```
## [1] 142.9015
```

Bias for MLE-estimatet er 0.2485681 og MLE-estimatets MSE er 1.54741. Moment-estimatets bias er 11.30892 og moment-estimatets MSE er 142.9015. Dvs. der er kæmpe forskel på de to estimater ved n=10.

```r
set.seed(123)
theta <- 5
n <- 100
Nsim <- 10000

shape <- 2
rate <- theta

mle_estimates <-  numeric(Nsim)
moment_estimates <- numeric(Nsim)

mle_bias_estimates <-  numeric(Nsim)
moment_bias_estimates <- numeric(Nsim)


for (i in 1:Nsim) {
  Z_i <- rgamma(n, shape = shape, rate = rate)
  X_i <- sqrt(Z_i)

  mle_est_for_current_n <- 2 * n / (sum(X_i ^ 2))
  moment_est_for_current_n <- 9 * pi^2 / (16 * (mean(X_i) ^2))

  mle_estimates[i] <- mle_est_for_current_n
  moment_estimates[i] <- moment_est_for_current_n

  mle_bias_estimates[i] <- mle_est_for_current_n - theta
  moment_bias_estimates[i] <- moment_est_for_current_n - theta
}

simulated_mle_bias_est <- mean(mle_bias_estimates)
simulated_moment_bias_est <- mean(moment_bias_estimates)

simulated_mle_bias_est
```

```
## [1] 0.02781073
```

```r
simulated_moment_bias_est
```

```
## [1] 10.78011
```

```r
var(mle_bias_estimates) + simulated_mle_bias_est ^2
```

```
## [1] 0.1301978
```

```r
var(moment_bias_estimates) + simulated_moment_bias_est ^2
```

```
## [1] 117.5585
```

Her er for MLE: bias 0.02791073 og MSE 0.1301978. For momentmetoden er bias 10.78011 og MSE er 117.5585. Det ser ud til, at for MLE konvergerer bias mod 0 for n gående mod uendelig (hvilket vi ved fra ISS-bogen er korrekt). Desuden er MSE også meget tættere på 0 nu.

Momentmetoden blev lidt bedre men giver stadig meget dårlige resultater.

Opgave 3

(3.a)

F-test forudsætter at vores populationer er normalfordelte jf. Chihara-bogen s. 428. Den antagelse er ikke opfyldt, da vi tester t-fordelt data.

Der simuleres fra nulhypotesen, da begge stikprøver har en varians på 3 (dvs. samme varians).

Ekstra note: Hvis man erstattede rt med rnorm i scriptet, ville man få en graf, der ville vise uniformt fordelte p-værdier. Grafen i opgaven er ikke helt uniformt fordelt, fordi vores antagelse om normalfordeling ikke er opfyldt.

(3.b)

(iii) ca. 0.25. $F(x)$ fortæller os andelen af p-værdier, der er mindre end x. Vi aflæser derfor F(0.05) som er ca. 0.25.

Der er derfor stor risiko for type 1 fejl ved denne test, hvilket skyldes, at antagelsen om normalfordeling ikke er opfyldt.

Opgave 4

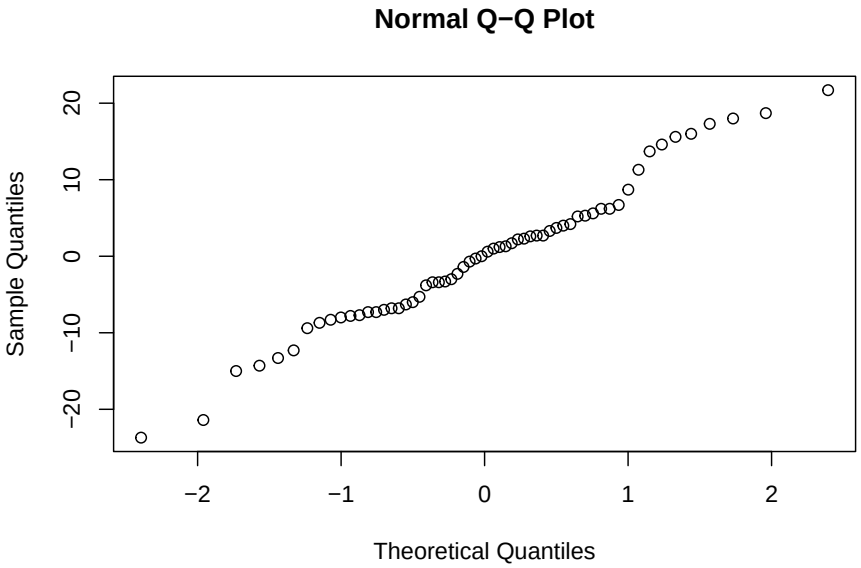
(4.a)

Modellen er

$$Y_{tsi} \sim N(\mu + \delta_t + \delta_s + \delta_{ts}, \sigma^2_{ts}), \;\; \sigma^2_{ts} \geq 0$$

hvor $\delta^{(RS)}_{ts}$ er forskellig fra 0 for mindst en kombination af t og s. t refererer her naturligvis til type, s til størrelse og Y til skumvolumen. My og deltaerne er naturligvis reelle tal. Jeg har valgt ikke at medtage $^{(R)}$, $^{(S)}$ og $^{(RS)}$ i opskrivningen, som Ute gør i sine noter, da vi her ikke har opdelt de to typer af kategorier i rækker og søjler.

```
data <- read.csv('aeg.csv')
M0 <- lm(skumvol ~ type * stor, data = data)
qqnorm(residuals(M0))
```



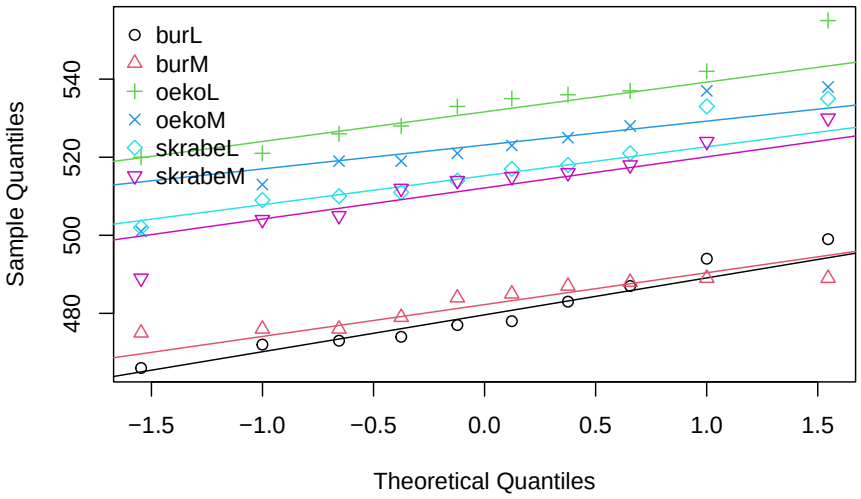
Vi ser ved QQ-plot, at data tilnærmelsesvis er normalfordelte.

```
data$gruppe <- paste(data$type, data$stor, sep="")

qqnorm_grouped <- function(y, groups, col = NULL, pch = NULL, pt.lwd = NULL,
   main = "", xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
                              ..., legend = TRUE, qqlines = TRUE){
  # lidt fejlbehandling
  if (missing(y)) stop ("no data given")
  if (missing(groups)) stop ("no groups given")
  if (length(y) != length(groups)) stop ("y and groups need to be vectors of same length")

  # lav en faktor variabel, hvis ikke groups allerede er det
  groups = factor(groups)
  # valg farver og symboler hvis de mangler
  ngroups <- nlevels(groups)
  if (is.null(col)) col <- 1 : ngroups
  col <- rep(col, length.out = ngroups)
  if (is.null(pch)) pch <- 1 : ngroups
  pch <- rep(pch, length.out = ngroups)
  # beregn x og y værdier til fraktildiagram: en slags "stille" qqnorm
  allqq <- tapply(y, groups, qqnorm, plot.it = FALSE)
  # nu har vi en liste med x og y til alle grupper.
  xrange <- range(sapply(allqq, function(qqgruppe) range(qqgruppe$x)))
  yrange <- range(y)
  # tom plot
  plot(xrange, yrange, type="n", xlab = xlab, ylab = ylab, main = main)
  # plot punkterne for hver gruppe i den passende farve
  if (is.null(pt.lwd)) pt.lwd = par("lwd")
```

```
  for (i in seq_along(allqq)) {
    points(allqq[[i]], col = col[i], pch = pch[i], lwd = pt.lwd,...)
    if (qqlines) qqline(allqq[[i]]$y, col = col[i], ...)
  }
  # tilføj legende
  if(legend) legend("topleft", legend = names(allqq), col = col, pch = pch,
                    bty = "n", pt.lwd = pt.lwd, ...)
}
qqnorm_grouped(data$skumvol, data$gruppe)
```



Og vi ser, at når vi betragter hver gruppe isoleret set, er data pænt normalfordelte.

(4.b)

```
bartlett.test(data$skumvol,data$gruppe)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data$skumvol and data$gruppe
## Bartlett's K-squared = 4.3553, df = 5, p-value = 0.4995
```

Ved Bartlett-test fås en p-værdi på 0.4995. Vi kan således godt antage, at varianserne for grupperne er ens.

(4.c)

```
M1 <- lm(skumvol ~ type * stor, data = data)

q <- c(0.975, 0.025)
df <- summary(M1)$df[2]
s_squared <- summary(M1)$sigma ^ 2
df * s_squared / qchisq(q, df = df)
```

```
## [1]  71.71746 153.55049
```

Et 95% CI er altså [71.71746,153.55049].

(4.d)

Vi vil teste:

$$H_0: \ Y_{tsi} \sim N(\mu + \delta_t^{(R)} + \delta_s^{(S)}, \sigma^2)$$

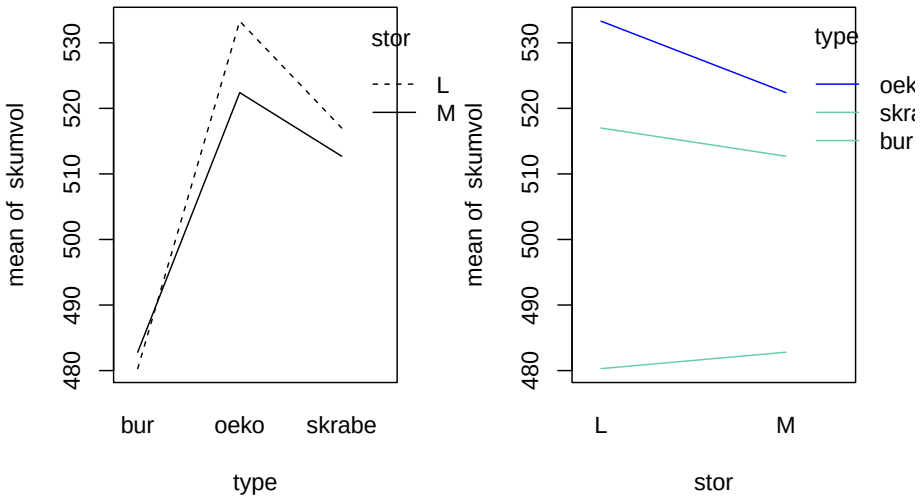
mod

$$H_A: \ Y_{tsi} \sim N(\mu + \delta_t^{(R)} + \delta_s^{(S)} + \delta_{ts}^{(RS)}, \sigma^2)$$

hvor $\delta_{ts}^{(RS)}$ er forskellig fra 0 for mindst en kombination af t og s.

Vi vil nu undersøge grafisk om data kan stamme fra en additiv model. Vi laver en interaktionsplot:

```
par(mfrow = c(1, 2))
with(data, interaction.plot(type, stor, skumvol,
    lty = c("dashed","solid")))
with(data, interaction.plot(stor, type, skumvol,
    col = c("aquamarine3", "blue"), lty = "solid"))
```



Vi ser, at linjerne ikke helt har samme hældning. Datasættets meget lille størrelse taget i betragtning er en additiv model dog stadig en mulighed, da forskellen ikke er enorm.

Vi laver nu en test for at se, om vi kan reducere:

```
M2 <- lm(skumvol ~ type + stor, data = data)
anova(M2,M1)
```

```
## Analysis of Variance Table
##
## Model 1: skumvol ~ type + stor
## Model 2: skumvol ~ type * stor
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     56 5913.2
## 2     54 5464.3  2    448.93 2.2183 0.1186
```

Vi får en p-værdi på 0.1186 ved anova-test, så vi kan godt reducere til en additiv model.

(4.e)

For at kunne reducere yderligere, skal vi kunne fjerne en af de to variable fra modellen. Vi prøver:

```
M3a <- lm(skumvol ~ stor, data = data)
anova(M3a,M2)
```

```
## Analysis of Variance Table
##
## Model 1: skumvol ~ stor
## Model 2: skumvol ~ type + stor
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     58 28723.8
## 2     56  5913.2  2     22811 108.01 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Når vi fjerner type, fås en ekstremt lille p-værdi, så det kan vi ikke. Nu prøver vi at fjerne størrelse:

```
M3b <- lm(skumvol ~ type, data = data)
anova(M3b, M2)
```

```
## Analysis of Variance Table
##
## Model 1: skumvol ~ type
## Model 2: skumvol ~ type + stor
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     57 6182.0
## 2     56 5913.2  1    268.82 2.5458 0.1162
```

Her fås en p-værdi på 0.1162, så vi kan godt reducere til en model, hvor skumvolumen kun afhænger af type. Herfra kan vi ikke meningsfyldt reducere modellen yderligere. Skumvolumen kan således meningsfyldt modelleres ud fra en lineærmodel, hvor skumvolumenet i et æg er normalfordelt og hvor normalfordelingens middelværdi kun afhænger af æggets type (bur, øko, skrabe). Således har ægstørrelsen ikke nogen nævneværdig indflydelse på skumvolumen.