Eksamen Matematisk Statistik F2019

vejledende besvarelse

6/16/2019

Disclaimer

Det her er en vejledende og meget kort besvarelse. Ved nogle af opgaverne findes flere rigtige svarmuligheder.

Opgave 1

Der findes forskellige svarmuligheder for testet. Man kan bruge enten G-testet, eller Pearsons χ^2 test. Man kan også beslutte sig for en simulationstest. I det følger bruges Pearsons χ^2 test.

(1.a)

Jeg bruger Pearson's χ^2-test for homogenitet. Under nulhypotesen antages teststørrelsen at være χ^2-fordelt med $df = (3-1)\cdot(2-1) = 2$ frihedsgrader.

(1.b, Pearson's test)

Kan regnes i hånden eller med R.

Med R:

```
prop.test(c(5, 11, 7), c(13, 20, 13))
```

```
##
##  3-sample test for equality of proportions without continuity
##  correction
##
## data:  c(5, 11, 7) out of c(13, 20, 13)
## X-squared = 0.96923, df = 2, p-value = 0.6159
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3
## 0.3846154 0.5500000 0.5384615
```

eller:

```
obs <- cbind(c(5, 11, 7), c(8, 9, 6))
chisq.test(obs)
```

```
##
##  Pearson's Chi-squared test
##
## data:  obs
## X-squared = 0.96923, df = 2, p-value = 0.6159
```

eller:

```
obs <- cbind(c(5, 11, 7), c(8, 9, 6))
chisq.test(obs, simulate = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  obs
## X-squared = 0.96923, df = NA, p-value = 0.7486
```

eller

```
obs <- cbind(c(5, 11, 7), c(8, 9, 6))
N <- sum(obs)
forv <- outer(rowSums(obs), colSums(obs))/N

C <- sum((obs - forv)^2/forv)
pval <- 1- pchisq(C, 2)

C
```

```
## [1] 0.9692308
```

```
pval
```

```
## [1] 0.615934
```

Teststørrelsen er $C = 0.969$, og p–værdien er $p_{\text{obs}} = 0.616$.

(1.c)

Data strider ikke imod antagelsen, at kvinder har de samme chancer at blive medlem i EP i alle tre lande.

Opgave 2

(2.a)

Lad Y_{ijk}, $i = 1, 2, 3$, $j = 1, 2$, $k = 1, \ldots, 8$ være trykstyrken i k-te forsøg målt med j-te måler for i-te blandemaskin.

Oprindeligt model:

$$M_0 : Y_{ijk} \sim N(\mu_{ij}, \sigma_{ij}^2),$$

Y_{ijk} uafhængige.

Vi vil teste, om modellen kan reduceres til model med ens varianser:

$$M_1 : Y_{ijk} \sim N(\mu_{ij}, \sigma^2).$$

(Model M_1 svarer tuk nulhypotese: $H_0 : \sigma_{ij}^2 = \sigma^2$ for alle i, j i model M_0)

Der bruges et Bartlett test:

```
beton <- read.csv("betonstyrke.csv")
bartlett.test(beton$styrke, beton$gruppe)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  beton$styrke and beton$gruppe
## Bartlett's K-squared = 6.8635, df = 5, p-value = 0.231
```

p-værdien $p_{\text{obs}} = 0.231$ ligger tydeligt over 0.05. Vi konkluderer at data ikke strider imod model M_1.

(2.b)

Model med additiv virkning:

$$M_2 : Y_{ijk} \sim N(\mu_{ij}, \sigma^2), \quad \mu_{ij} = \mu + \alpha_i + \beta_j, \quad \alpha_1 = 0,\ \beta_1 = 0.$$

Test, om model M_1 må reduceres til M_2:

```
M1 <- lm(styrke ~ gruppe, data = beton)
M2 <- lm(styrke ~ blander + maaler, data = beton)
anova(M2, M1)
```

```
## Analysis of Variance Table
##
## Model 1: styrke ~ blander + maaler
## Model 2: styrke ~ gruppe
##   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
```

Testet resulterer i en stor p-værdi på 0.8534. Derfor forkastes ikke hypotesen, at data kan beskrives ved model M_2 frem for M_1.

(2.c)

I denne delopgave skal vises, at blandemaskinen ikke påvirker trykstyrken, dvs., at data kan beskrives ved modellen

$$M_3 : \mu_{ij} = \mu + \beta_j, \quad \beta_1 = 0.$$

```
M3 <- lm(styrke ~ maaler, data = beton)
anova(M3, M2)
```

```
## Analysis of Variance Table
##
## Model 1: styrke ~ maaler
## Model 2: styrke ~ blander + maaler
##   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
```

Testet resulterer i en stor p-værdi på 0.5284. Derfor forkastes ikke hypotesen, at data kan beskrives ved model M_3 frem for M_2.

(2.d)

Konfidensintervallet kan regnes "i hånden", eller ved R. Beregning "i hånden", fra summary:

```
summary(M3)
```

```
##
## Call:
## lm(formula = styrke ~ maaler, data = beton)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0458  -3.0458  -0.8625   3.4208   9.7208
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.0458     0.9825  55.011  < 2e-16 ***
## maalerb      -5.8667     1.3894  -4.222 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.813 on 46 degrees of freedom
## Multiple R-squared:  0.2793, Adjusted R-squared:  0.2637
## F-statistic: 17.83 on 1 and 46 DF,  p-value: 0.0001128
```

Forskellen mellem de to målere estimeres til $\hat{\beta}_2 = -5.8667$, med en standardfejl på $\widehat{\text{SE}}[\hat{\beta}_2] = 1.3894$, som er estimeret med 46 frihedsgrader. Derved beregnes konfidensintervallet som følger:

```
-5.8667 + c(-1, 1)* qt(0.975, 46) * 1.3894
```

```
## [1] -8.663417 -3.069983
```

Det fås også vha R-funktionen `confint`:

```
confint(M3, parm = 2)
```

```
##             2.5 %    97.5 %
## maalerb -8.663384 -3.069949
```

Da konfidensintervallet ikke indeholder 0, kan vi ikke antage, at der ikke er forskel på de to trykstyrkemålere.

(2.e)

Residual-standardafvigelsen estimeredes til $\sqrt{s^2} = 4.813$. Estimatoren til residualvariansen har en skaleret χ^2 fordeling med $df = 46$ frihedsgrader, idet $df \cdot S^2/\sigma^2 \sim \chi^2(df)$. Derved gælder

$$P(q_1 \leq df \cdot S^2/\sigma^2 \leq q_2) = 0.95, \quad q_1 = F^{-1}_{\chi^2(df)}(0.025),\ q_2 = F^{-1}_{\chi^2(df)}(0.975)$$
$$\Longrightarrow P(df \cdot S^2/q_2 \leq \sigma^2 \leq df \cdot S^2/q_1) = 0.95$$

Konfidensintervallet kan beregnes i hånden eller direkte fra den fittede model som følger:

```
s2 <- summary(M3)$sigma^2 # eller 4.813^2
chiquant <- qchisq(c(0.975, 0.025), 46)
s2 / chiquant * 46
```

```
## [1] 15.99602 36.54311
```

Et 95%-konfidensinterval for σ^2 er altså givet ved $[16.0, 36.5]$.

Opgave 3

(3.a)

Det må være **boxplot C**, der beskriver de oprindelige data. Medianen i de logaritmerede data aflæses ved theoretical quantile $\Phi^{-1}(0.5) = 0$, til at være ca. 1.9, og det tredje kvartil aflæses ved theoretical quantile $\Phi^{-1}(0.75) = 0.67$ til at ligge mellem 2.3 og 2.4. Derved må medianen af de oprindelige data ligge ved ca. $e^{1.9} = 6.7$, og det tredje kvartil mellem $e^{2.3} = 10$ og $e^{2.4} = 11$. Dette opfyldes kun af boxplot C.

(3.b)

En enkelt måling har en spredning på $\sigma = 0.34$. For gennemsnittet $\bar{d}$ af n målinger gælder, at $\text{SE}[\bar{d}] = \sigma/\sqrt{n}$. Så beregnes det nødvendige antal af målinger som

$$n \geq (\sigma/\text{SE}[\bar{d}])^2 = (0.34/0.05)^2 = 46.24.$$

Der skal altså mindst laves 47 målinger.

(3.c)

Vi har at gøre med tælledata: antallet af storme i et givet tidsrum. En passende model til den slags data er Poissonmodellen, her

$$X_1 \sim \text{Pois}(\lambda_1), \quad X_2 \sim \text{Pois}(\lambda_2).$$

I denne model formuleres hypoteser, der passer til det faglige spørgsmål, som:

$$\text{nulhypotese}: \ H_0 : \lambda_1 = \lambda_2 \quad \text{vs.} \quad \text{alternative } H_A : \lambda_1 < \lambda_2.$$

Når man får en p-værdi på 0.97 vil man ikke forkaste nulhypotesen, dvs, data strider ikke imod antagelsen, at den forventede antal storme per år er den samme i perioden 2009-2018 som i 1999-2008. Data støtter altså ikke den omtalte frygt.