

CA06 - STAT100

Fábio Rodrigues Pereira

Oppgave 0

Se på løsningsforsalget til modul 8, diskuter tre ting dere ikke fikk til helt eller forstod fra forrige uke, som dere nå skjønner litt bedre. Skriv ned hvilke tema som ble diskutert og gjerne en setning av hva dere kom frem til.

(B) Since the variables are paired, then $D_i = X_i - Y_i$, where $D_i = \mu_D + E_i$ where $E_i \sim N(0, \sigma_D)$ for $i = 1, \dots, 15$. \blacksquare

(B) $\hat{\mu}_D = \bar{D} = \frac{1}{m_d} \sum_{i=1}^{m_d} D_i \approx 0,20$ and $\hat{\sigma}_D = SD[\bar{D}] = \frac{1}{m_d-1} \sum_{i=1}^{m_d} (D_i - \bar{D})^2 = 0,23$ and $SE[\hat{\mu}_D] = \frac{SD[\bar{D}]}{\sqrt{m_d}} = \frac{0,23}{\sqrt{15}} \approx 0,06$. \blacksquare

(F) $H_0: \mu_D = 0$ vs $H_1: \mu_D > 0$; $\bar{D} \sim T(\hat{\mu}_D = 0,2; \hat{\sigma}_D = 0,23)$ for $i = 1, \dots, 15$ and $\alpha = 0,05$; $T = \frac{\bar{D} - \hat{\mu}_D}{SE[\bar{D}]} = \frac{0,2 - 0}{0,23/\sqrt{15}} = \frac{0,2}{0,06} = 3,33$. \blacksquare

↳ Since $T = 3,33 > 1,761$, then reject H_0 . \blacksquare

(J) $X_i = \mu_X + E_i$ where $E_i \sim N(0, \sigma_X)$ for $i = 1, \dots, 15$; $\hat{\mu}_X = \bar{X} = \frac{1}{m_X} \sum_{i=1}^{m_X} X_i \approx 1,75$ and $\hat{\sigma}_X = \sqrt{\frac{1}{m_X-1} \sum_{i=1}^{m_X} (X_i - \bar{X})^2} \approx 0,24$. \blacksquare

$Y_i = \mu_Y + E_i$ where $E_i \sim N(0, \sigma_Y)$ for $i = 1, \dots, 15$; $\hat{\mu}_Y = \bar{Y} = \frac{1}{m_Y} \sum_{i=1}^{m_Y} Y_i \approx 1,56$ and $\hat{\sigma}_Y = \sqrt{\frac{1}{m_Y-1} \sum_{i=1}^{m_Y} (Y_i - \bar{Y})^2} \approx 0,30$. \blacksquare

(K) Spredsel variance := $\frac{(m_X-1)s_X^2 + (m_Y-1)s_Y^2}{m_X+m_Y-2} = \hat{\sigma}_D^2 = 0,074$. \blacksquare

$\hat{\sigma}_D = \sqrt{\hat{\sigma}_D^2} \approx \sqrt{0,074} \approx 0,272$ and $SE[\hat{D}] = \hat{\sigma}_D \cdot \sqrt{\frac{1}{m_X} + \frac{1}{m_Y}} \approx 0,099$. \blacksquare

Oppgave 1

Introduksjon til variansanalyse

→ venstre

Volum (målt i milliliter) av venstre hjertekammer (heretter bare kalt volum) ble målt på et tilfeldig utvalg mannlige toppidrettsutøvere innen svømming, langdistanseløping, langrenn og bryting. Resultatet er lagt i en Rdata-fil kalt idrett og er i tillegg vist på slutten av denne oppgaven.

R-datasettet idrett finner du under Modul data. Last det ned på egen maskin. Last også ned R-script-fila «Modul9_Kollokvieoppgave.R».

Åpne Rstudio og åpne både datafila idrett og R-script-fila fra File-Open-menyen i Rstudio.

La Y_{ij} være volum for idrettsutøver j innen idrettsgren i .

A) Hvilke verdier får du for Y_{32} og for Y_{23} ?

→ $Y_{3,2} := \text{Volume of athlete 2 in report 3 (crosscountry)} = 184$. \blacksquare

→ $Y_{2,3} := \text{Volume of athlete 3 in report 2 (running)} = 182$. \blacksquare

1a)

```

5 ~`{r}
6 # Create Variables:
7 Swimming <- c(177, 178, 177, 162, 158, 184, 194, 171, 176, 171)
8 Running <- c(180, 152, 182, 157, 171, 172, 164, 145, 162, 157)
9 CrossCountry <- c(198, 184, 192, 191, 182, 183, 198, 198, 178, 170)
10 Wrestling <- c(166, 153, 164, 132, 152, 151, 182, 166, 152, 145)
11
12 # Create data frame with the variables:
13 volumes <- data.frame(Swimming, Running, CrossCountry, Wrestling)
14
15 # Show dimension and data frame:
16 cat("Dimension (size, groups): ", dim(volumes))
17 cat("\n\n")
18 volumes
19 ```

Dimension (size, groups): 10 4

Swimming Running CrossCountry Wrestling
1 177 180 198 166
2 178 152 184 153
3 177 182 192 164
4 162 157 191 132
5 158 171 182 152
6 184 172 183 151
7 194 164 198 182
8 171 145 198 166
9 176 162 178 152
10 171 157 170 145

```

```

20 ~`{r}
21 # Exercise 1a):
22
23 cat("Y_3_2 := Athlete 2, Sport 3 (CrossCountry):", volumes[2, 3])
24 cat("\nY_2_3 := Athlete 3, Sport 2 (Running):", volumes[3, 2])
25 ```

Y_3_2 := Athlete 2, Sport 3 (CrossCountry): 184
Y_2_3 := Athlete 3, Sport 2 (Running): 182

```

B)

Hva er Y_{ij} i denne undersøkelsen? → Volume of the ventricle of athlete j and sport i . //

Hva er k og hva er n i denne undersøkelsen? → k := size of sports; n := size of athletes. //

i) μ_i og σ er de ukjente parameterne i denne undersøkelsen, tolk disse.

ii) Hvordan vil du tolke ε_{ij} ?

iii) Hvorfor er ε_{ij} en ukjent størrelse?

i) μ_i := average of the volume of the ventricle in sport i ;

σ := standard deviation / dispersion from the mean of the volumes of the ventricles. //

ii) $Y_{ij} = \mu_i + \varepsilon_{ij} \rightarrow \varepsilon_{ij} = Y_{ij} - \mu_i \rightarrow$ for $\hat{\mu}_i = \bar{Y}_i \rightarrow \varepsilon_{ij} = Y_{ij} - \bar{Y}_i$:= residual where $\varepsilon_{ij} \sim N(0, \sigma)$ ↗

iii) ε_{ij} is unknown because μ_i is unknown ↗

C)

Summetegnet spiller en vesentlig rolle i variansanalysen.

Gi en tolkning av følgende størrelser,

$$\frac{\sum_{i=1}^4 \sum_{j=1}^{10} Y_{ij}}{40} \quad := \text{This is the average/mean of all groups and samples} = \bar{Y}_j //$$

$$\frac{\sum_{j=1}^{10} Y_{ij}}{10} \quad (\text{kall denne/disse } \bar{Y}_i) \quad := \text{This is the mean/average of a group } i = \bar{Y}_i //$$

$$\frac{\sum_{i=1}^4 \sum_{j=1}^{10} (Y_{ij} - \bar{Y}_j)^2}{39} \quad := \text{This is the total variance between all groups and samples.} //$$

$$\frac{\sum_{j=1}^{10} (Y_{ij} - \bar{Y}_i)^2}{9} \quad \text{kall denne/disse } S_i^2 \quad := \text{This is a group } i \text{ variance.} // \blacksquare$$

D)

i) Hvordan vil du estimere μ_i , og hvordan vil du estimere σ^2 ?

Finn estimatene vha RStudio.

ii) Finn også de empiriske standardavvikene (og variansene) til hver gruppe. Kommenter tallene: Ser det ut til å være forskjeller på forventet hjertevolum for ulike idrettsutøvere? Ser det ut som om standardavviket er likt i alle gruppene? Kva kan du da si om antakelsen $\varepsilon_{ij} \sim N(0, \sigma^2)$?

$$\hat{\mu}_i = \bar{Y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{ij}; \hat{\mu}_1 \approx 174,8; \hat{\mu}_2 \approx 164,2; \hat{\mu}_3 \approx 187,4; \hat{\mu}_4 \approx 156,3$$

$$\hat{\sigma}^2 = \frac{(m_1-1)S_1^2 + (m_2-1)S_2^2 + (m_3-1)S_3^2 + (m_4-1)S_4^2}{m_1+m_2+m_3+m_4-k} = \frac{\left(\frac{m_1-1}{m_1-1}\right) \cdot \sum_{j=1}^{m_1} (Y_{1j} - \bar{Y}_1)^2 + \dots + \left(\frac{m_4-1}{m_4-1}\right) \cdot \sum_{j=1}^{m_4} (Y_{4j} - \bar{Y}_4)^2}{m_1+\dots+m_4-k} = \frac{\sum_{i=1}^4 \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2}{m-k} = \frac{SSE}{m-k} = MSE$$

$$\therefore \hat{\sigma} = \sqrt{MSE}$$

$$\hat{\sigma}_1^2 = \frac{1}{m_1-1} \sum_{j=1}^{m_1} (Y_{1j} - \bar{Y}_1)^2 \approx 105,51 \text{ and } \hat{\sigma}_1 = \sqrt{\hat{\sigma}_1^2} \approx 10,27; \hat{\sigma}_2^2 \approx 144,4 \text{ and } \hat{\sigma}_2 \approx 12,01; \hat{\sigma}_3^2 \approx 91,37 \text{ and } \hat{\sigma}_3 \approx 9,56$$

$\hat{\sigma}_4^2 \approx 189,12$ and $\hat{\sigma}_4 \approx 13,75$ → The CrossCountry group has the smallest dispersion from its mean and the groups of wrestling has the largest dispersion of its data from its mean. ↴ E_{ij} is the residual given by $Y_{ij} - \bar{Y}_i$ from $Y_i = \mu_i + \varepsilon_{ij} \rightarrow Y_i - \bar{Y}_i = (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i)$

```

26 - ````{r}
27 # Exercise 1d):
28
29 cat("Estimate mean for group 1 (Swimming): ", mean(volumes$Swimming),
30     "\nEstimate variance for group 1 (Swimming): ", var(volumes$Swimming),
31     "\nEstimate standard deviation for group 1 (Swimming): ", sd(volumes$Swimming),
32     "\n",
33     "\nEstimate mean for group 2 (Running): ", mean(volumes$Running),
34     "\nEstimate variance for group 2 (Running): ", var(volumes$Running),
35     "\nEstimate standard deviation for group 2 (Running): ", sd(volumes$Running),
36     "\n",
37     "\nEstimate mean for group 3 (CrossCountry): ", mean(volumes$CrossCountry),
38     "\nEstimate variance for group 3 (CrossCountry): ", var(volumes$CrossCountry),
39     "\nEstimate standard deviation for group 3 (CrossCountry): ", sd(volumes$CrossCountry),
40     "\n",
41     "\nEstimate mean for group 4 (Wrestling): ", mean(volumes$Wrestling),
42     "\nEstimate variance for group 4 (Wrestling): ", var(volumes$Wrestling),
43     "\nEstimate standard deviation for group 4 (Wrestling): ", sd(volumes$Wrestling))
44 ...
```

```

Estimate mean for group 1 (Swimming): 174.8
Estimate variance for group 1 (Swimming): 105.5111
Estimate standard deviation for group 1 (Swimming): 10.27186

Estimate mean for group 2 (Running): 164.2
Estimate variance for group 2 (Running): 144.4
Estimate standard deviation for group 2 (Running): 12.01666

Estimate mean for group 3 (CrossCountry): 187.4
Estimate variance for group 3 (CrossCountry): 91.37778
Estimate standard deviation for group 3 (CrossCountry): 9.559172

Estimate mean for group 4 (Wrestling): 156.3
Estimate variance for group 4 (Wrestling): 189.1222
Estimate standard deviation for group 4 (Wrestling): 13.75217
```

E)

For å gå videre må dere stacke data, se kommentarer i R-script-fila om hvorfor og hvordan.

↳ stack() in R

Bruk de nye variablene til å finne

total gjennomsnitt (uavhengig av idrettsgren), $\mu_{values} \approx 170,675$ //
total standardavvik (uavhengig av idrettsgren) $\hat{\sigma}_{values} \approx 16,19336$ //

```
45 ````{r}
46 # Exercise 1e):
47
48 sports <- stack(volumes) # or unstack()
49 cat("Mean of values: ", mean(sports$values),
50     "\nVariance of values: ", var(sports$values),
51     "\nStandard Deviation of values : ", sd(sports$values))
52 ````
```

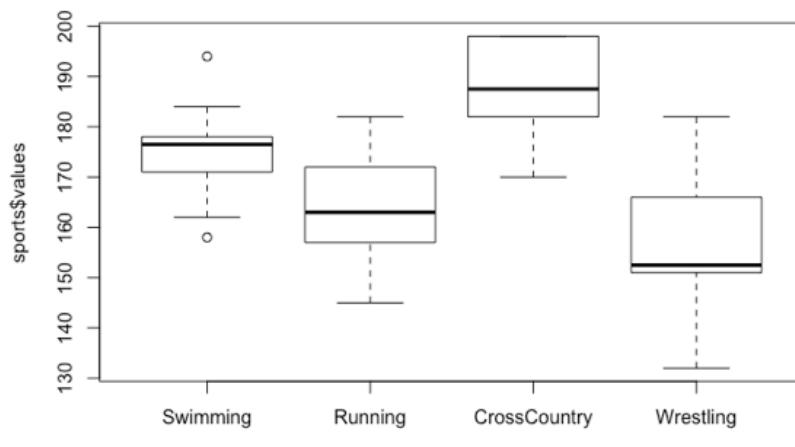
```
Mean of values: 170.675
Variance of values: 262.225
Standard Deviation of values : 16.19336
```

F)

Lag et boksplott der du deler inn idrettsgreiner.

Ved å studere dette samt verdiene du fant i D og E, tror du at det er virkelig forskjell mellom gruppene eller tror du at de forskjellene vi ser er innenfor det vi kan forvente av normal variasjon for fire i utgangspunktet like grupper?

```
53 ````{r}
54 # Exercise 1f):
55
56 boxplot(sports$values~sports$ind)
57 ````
```



We do see differences between groups, for example, the mean volume of the crosscountry group is above of the 4th quantile of the others groups. Other example is that the wrestling's dispersion is bigger than the others dispersions, and wrestling's mean below others 1st quantiles. Then the varionle between groups are clearly different. //

G)

Siden ε_{ij} og dens standardavvik σ er ukjent, må vi nøye oss med det vi kaller residualer, nemlig

$$\hat{\varepsilon}_{ij} = e_{ij} = Y_{ij} - \bar{Y}_i$$

der \bar{Y}_i er det vi kaller tilpasset verdi, og den skrives av og til som \hat{Y}_i .

- Ⓐ Vanligvis er det datamaskin som gir oss residualene, men regn ut disse for hånd for svømmer nummer 7 og for bryter nummer 4.
- Ⓑ Hva kan du si om en idrettsutøver som har et negativt residual?

$$① \quad \varepsilon_{ij} = Y_{ij} - \bar{Y}_i \rightarrow \begin{cases} \varepsilon_{1,7} = Y_{1,7} - \bar{Y}_1 = 194 - 174,8 = 19,2 \\ \varepsilon_{4,4} = Y_{4,4} - \bar{Y}_4 = 132 - 156,3 = -24,3 \end{cases}$$

```

58 - ````{r}
59 # Exercise 1g):
60
61 cat("Swimmer (j=7, i=1) has volume: ", volumes[7, 1],
62   "\nSwimming mean: ", mean(volumes$Swimming),
63   "\nResidual of Swimmer 7:", volumes[7, 1] - mean(volumes$Swimming),
64   "\n",
65   "\nWrestler (j=4, i=4) has volume: ", volumes[4, 4],
66   "\nWrestling mean: ", mean(volumes$Wrestling),
67   "\nResidual of Wrestler 4:", volumes[4, 4] - mean(volumes$Wrestling))
68 ...

```

Swimmer (j=7, i=1) has volume: 194
 Swimming mean: 174.8
 Residual of Swimmer 7: 19.2

Wrestler (j=4, i=4) has volume: 132
 Wrestling mean: 156.3
 Residual of Wrestler 4: -24.3

* The wrestler who has negative residual expresses that his volume is below the average of his group. ☐

H)

Vi splitter den totale variasjon opp i to deler (derav navnet variansanalyse)), nemlig variasjon innen grupper og variasjon mellom grupper.

Vi har $SS_{total} = SS_{gruppe} + SS_{error}$ (eventuelt $SS_{error} = SS_{residual}$)

Når blir SS_{gruppe} stor (eller liten) og når blir SS_{error} stor (eller liten)?

$$Y_{ij} = \mu_i + \varepsilon_{ij} \rightarrow Y_{ij} = Y_i + \varepsilon_{ij} \rightarrow Y_{ij} - \bar{Y} = (Y_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i)$$

$$\rightarrow \sum_{i=1}^k \sum_{j=1}^m (Y_{ij} - \bar{Y})^2 = \underbrace{\sum_{i=1}^k \sum_{j=1}^m (\bar{Y}_i - \bar{Y})^2}_{SS_{total}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2}_{SS_{residual}}$$

$\sum_{i=1}^k m_i (\bar{Y}_i - \bar{Y})^2$

$SSG = \sum_{i=1}^k \sum_{j=1}^m (Y_i - \bar{Y})^2$ turns big when there are large differences between groups;

$SSR = \sum_{i=1}^k \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2$ turns big when there are large differences within a group. ☐

Før du tester er det et begrep som er svært vanskelig på få tak i er; frihestgrader (degrees of freedom, df). Hver kavdratsum har koblet til seg frihetsgrader, dette er hvor mange lineært uavhengige ledd som er knyttet til denne kvadratsummen. Dette kan vi diskutere i detalj ved en senere anledning.

I)

Vi skal nå teste om det er signifikant gruppeeffekt.

Ⓐ Sett opp nullhypotese og alternativ hyposete både med ord og ved parametere.

Ⓑ Utfør testen (vha Rstudio) både ved å se på F-verdien og ved å bruke p-verdien.

Ⓒ Hvorfor trenger du kun å enten se på F-verdien eller p-verdien for å kunne trekke konklusjonen?

Ⓓ Gi en tolkning av denne p-verdien.

$H_0 := \mu_1 = \mu_2 = \mu_3 = \mu_4 \rightarrow \text{all group averages are the same.}$

VS

$H_1 := \text{at least one pair of group averages are different.}$

model := $Y_{ij} = \mu_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$ for $i=1, \dots, k$ and $j=1, \dots, m$

$$F\text{-statistic} := F = \frac{\text{variance between groups}}{\text{variance inside groups}} = \frac{MSB}{MSE} = \frac{\frac{\text{Explained SS}}{(k-1)}}{\frac{\text{Residual SS}}{(m-k)}} = 13,708$$

```
69 - ````{r}
70 # Exercise 1i):
71
72 ##### Calculating Explained Sum of Square ESS:
73 ESS <- 0
74 k <- 4
75
76 for (i in 1:k){
77   ESS <- ESS + (nrow(volumes[i]) * (mean(volumes[, i]) - mean(sports$values)) ^ 2)
78 }
79
80 ESS_df <- k-1
81 MSB <- ESS/ESS_df
82
83 cat("The Explained Sum of Square is ", ESS, "\nThe ESS's degree of freedom is ", ESS_df,
84   "\nThe Mean Square Factor MSB is ", MSB)
85
86 ##### Calculating Residual Sum of Square RSS:
87 RSS <- 0
88 n <- 10
89
90 for (i in 1:k){
91   for (j in 1:n){
92     RSS <- RSS + (volumes[j, i] - mean(volumes[, i])) ^ 2
93   }
94 }
95
96 RSS_df = k*n - k
97 MSE <- RSS/RSS_df
98
99 cat("\n\nThe Residual Sum of Square is ", RSS, "\nThe RSS's degree of freedom is ", RSS_df,
100   "\nThe Mean Square Error MSE is ", MSE)
101
102 ##### Calculating F-statistic:
103 cat("\n\nThe F-Statistic is ", MSB/MSE)
104
```

The Explained Sum of Square is 5453.075
 The ESS's degree of freedom is 3
 The Mean Square Factor MSB is 1817.692
 The Residual Sum of Square is 4773.7
 The RSS's degree of freedom is 36
 The Mean Square Error MSE is 132.6028
 The F-Statistic is 13.70779

```
105 - ````{r}
106 # Exercise 1i):
107
108 model <- lm(sports$values ~ sports$ind)
109 anova(model)
110
111 Analysis of Variance Table
112 Response: sports$values
113   Df Sum Sq Mean Sq F value Pr(>F)
114 sports$ind 3 5453.1 1817.7 13.708 4.038e-06 ***
115 Residuals 36 4773.7 132.6
116
117 Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
118
119
120 summary(model)
121
122
123 Call:
124 lm(formula = sports$values ~ sports$ind)
125
126 Residuals:
127   Min   1Q Median   3Q   Max
128 -24.30 -5.85 -1.20  8.15 25.70
129
130 Coefficients:
131
132 (Intercept) 174.800 3.641 48.003 < 2e-16 ***
133 sports$indRunning -10.600 5.150 -2.058 0.046853 *
134 sports$indCrossCountry 12.600 5.150 2.447 0.019431 *
135 sports$indWrestling -18.500 5.150 -3.592 0.000972 ***
136
137 Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
138
139 Residual standard error: 11.52 on 36 degrees of freedom
140 Multiple R-squared:  0.5332, Adjusted R-squared:  0.4943
141 F-statistic: 13.71 on 3 and 36 DF, p-value: 4.038e-06
```

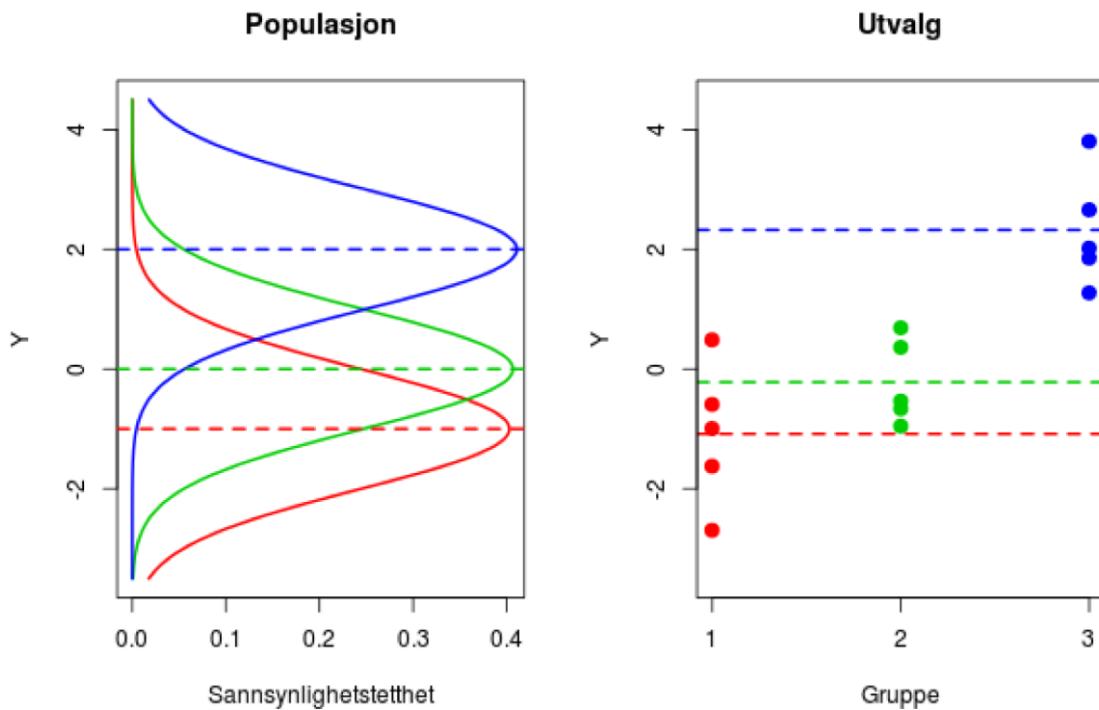
Note: The p-value has significance level $\alpha = 0,001$

☰ Because: If $F > F_\alpha$ or $p\text{-value} < \alpha$, then reject H_0

$Y_{ij} = \mu_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$ for $i=1, \dots, k=4$ and $j=1, \dots, m=10$

Conclusion: Since $F = 13,70 > F_{3, 36, \frac{\alpha}{2}} = 3,5$ for $\alpha = 0,05$, then reject H_0 . The averages are different. ☐

Oppgave 2



Én på hver gruppe går inn på <https://solve.shinyapps.io/Variansanalyse>.

Vi skal bruke denne applet'en til å se hvordan variansanalyseresultatene avhenger av balansen mellom antall observasjoner, forskjellen mellom forventningene og verdien på støystandardavviket (σ).

Til venstre i app'en kan vi bestemme de *sanne* verdiene av forventningene til tre grupper, dvs verdiene av μ_1 , μ_2 og μ_3 og samt den *sanne* verdien av støystandardavviket (σ).

Vi er altså i den uvanlige (og urealistiske) situasjonen av vi kjenner populasjonsparametrene og kan sette verdien av disse. Vi vil så benytte *simulerte tilfeldige* utvalg fra disse populasjonene og se hvordan våre statistiske metoder takler utfordringene med estimering og hypotesetesting. Slike simuleringer brukes mye i statistisk forskning for å studere hvor gode våre metoder er.

A) Start med å justere forventningene slik at

$$\mu_1 = -1, \mu_2 = 0, \text{ og } \mu_3 = 1, \text{ samt at } \sigma = 3.$$

Velg at det skal trekkes $n=5$ observasjoner fra hver populasjon.

Trykk på knappen «Oppdater figurer». Det vil da bli trukket 5 observasjoner fra hver populasjon og du vil se disse i figuren til høyre som prikker med tre ulike farger, én farge for hvert utvalg (gruppe). De stiplede linjene angir gjennomsnittet i hvert utvalg. I figuren til venstre ser vi de tre normalfordelingene (satt på høykant) som er bestemt av de parameterverdiene vi satte for forventningene og standardavviket.

④ På bakgrunn av figurene: Synes du gjennomsnittene gode estimerer på de sanne forventningene?

⑤ Trykk på fanen «Anovatabell». Hva ble verdien av testobservatoren F ? Vi kan bruke F eller p-verdien i siste kolonne i tabellen til å teste hypoteser.

Hvilke hypoteser tester vi i så fall?

Får du forkastning av nullhypotesen med testnivå 5%?

⑥ Du fikk her enten forkastning eller ikke forkastning av nullhypotesen.

⑦ Ut fra det du vet om de sanne parameterverdiene, gjorde du en korrekt beslutning i testen? **Yes!**
Siden H_0 er gal burde den vært forkastet.

⑧ Yes, because we know that the population is normal distributed, then symmetrical //

⑨ $F = 3,2513$; Yes, F or p -value can be used to hold or reject the H_0 claim //

$H_0: \mu_1 = \mu_2 = \mu_3 \quad \text{vs} \quad H_1: \text{at least one pair is not equal}$

$Y_{ij} = \mu_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, 3)$ for $i=1,2,3=k$ and $j=1, \dots, 5=m$

$$F = \frac{\frac{ESS}{k-1}}{\frac{RSS}{m-k}} = \frac{MSF}{MSE} = 3,2513 \rightarrow F = 3,2513 < F_{\alpha=0,05} = 5,4095 //$$

and $P(F > F_{\alpha} | H_0 \text{ is true}) = 0,074 > \alpha = 0,05 //$

Hold H_0 //

B)

Lat som om en ny student skal gjøre samme forsøk som det du gjorde i A. Gjenta hele prosedyren. Merk at du nå får nye parameterestimater, og ny F- og p-verdier.

$H_0: \mu_1 = \mu_2 = \mu_3 \quad \text{vs} \quad H_1: \text{at least one pair is not equal}$

$Y_{ij} = \mu_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, 3)$ for $i=1,2,3=k$ and $j=1, \dots, m=30$

$$F = \frac{\frac{ESS}{k-1}}{\frac{RSS}{m-k}} = \frac{MSF}{MSE} = 4,2348 \rightarrow F = 4,2348 > F_{\alpha=0,05} \approx 2,92 // \rightarrow \text{reject } H_0$$

and $P(F > F_{\alpha} | H_0 \text{ is true}) = 0,017 < \alpha = 0,05 // \rightarrow \text{reject } H_0$

C)

Bruk samme parameterverdier som i oppgave A.

La nå antall observasjoner i hver gruppe øke gradvis over verdiene 5, 10, 20, 30, 40, og 50.

⑧ For hver verdi av n notér deg verdien på testobservatoren F .

⑨ Hva er tendensen for verdien av F og evnen til å forkaste en nullhypotesen?

⑩ Lag gjerne et plott over F-verdiene for å se trenden.

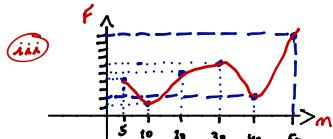
⑪ Forklar hvorfor du ser det du ser, gjerne basert på hvordan F er beregnet.

Moral: Hvis H_0 er gal, blir det lettere og lettere og forkaste denne hvis n vokser.

⑫ for $m=5 \rightarrow F = 4,6215$; for $m=10 \rightarrow F = 1,2558$; for $m=20 \rightarrow F = 5,0174$; for $m=30 \rightarrow F = 6,213$ //

for $m=40 \rightarrow F = 2,0646$; for $m=50 \rightarrow F = 10,134$ //

⑬ as $m \rightarrow \infty$ or $F \rightarrow \infty$, despite some exceptions //



⑭ F is the ratio between variance between groups and variance inside the groups. If the number of samples goes up, due m goes up, then F goes up, in other words

$$F = \frac{\frac{ESS}{k-1}}{\frac{RSS}{m-k}} = \frac{ESS}{RSS} \cdot \frac{(m-k)}{(k-1)} = \uparrow F$$

D)

La det nå være $n=15$ observasjoner i hver gruppe og hold forventningene uforandret på

$$\mu_1 = -1, \mu_2 = 0, \text{ og } \mu_3 = 1,$$

men varier standardavviket (σ) gradvis over verdiene 1, 3, 5, 7 og 9.

- i) Hvordan avhenger F og evnen til forkaste nullhypotesen av σ ?*
- ii) Hvorfor er det slik forklar både intuitivt og ved formel for F ?*

iii) Moralen i denne oppgaven er: Dersom du får stor p-verdi og dermed ikke klarer å forkaste kan det skyldes tre årsaker, hvilke kan det være?

i) for $\sigma=1 \rightarrow F=12,996$; for $\sigma=3 \rightarrow F=3,7482$; for $\sigma=5 \rightarrow F=3,2234$;

for $\sigma=7 \rightarrow F=0,8705$; for $\sigma=9 \rightarrow F=0,9533$

** as $\sigma \rightarrow \infty$ as $F \rightarrow 0$*

ii) $F = \frac{ESS}{RSS} \cdot \frac{(n-k)}{(k-1)} = \frac{\sum_{i=1}^k m_i (\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2} \cdot \frac{(n-k)}{(k-1)} = \downarrow F$ and $Y_i = \mu_i + E_{ij}$ where $E_{ij} \sim N(0, \sigma^2)$.

iii) ?

Ekstra:

E) Hvordan tolker vi, og hvordan beregner vi R^2 ? $\rightarrow \frac{ESS}{RSS}$
Prøv med hjelp av simuleringer å se hvordan følgende påvirker R^2 . (Ta gjerne litt hardt i).

Suppose $R^2 = \frac{ESS}{RSS} = 0,30$, then 30% of the dispersion between groups have been accounted for and the remaining 70% is still unexplained.

Øke n , holde alt annet konstant.

Øke σ holde alt annet konstant.

Øke forskjellen mellom μ -ene holde alt annet konstant.

Oppgave 3

Merk dere hvordan eksamsoppgaver høst 2016 Del B oppgave 1 og vår 2017 oppgave 1 er formulert.

Prøv dere på vår 2017. Diskuter dere fram til hvordan denne typen oppgaver løses best mulig (vent med begrepene kontraster og vurdering av modellantagelser, dette vil bli diskutert snart).

Merk at Det foreligger ikke løsningsforslag på disse to oppgavene, og det vil heller ikke bli laget.

Vår 2017:

Oppgave 1

Kvaliteten på frukt er ofte bedømt etter sukkerinnholdet. I et forsøk utført på Bioforsk (nå NIBIO) på Ås i 2011 ble 4 pæresorter sammenlignet, 3 av disse var av Kvede typen. En lagde juice av prøvene og målte sukkerinnholdet i juiceen. Siden sukkerinnholdet kan variere innenfor en pæresort, valgte en å ta 6 gjentak for hver sort. Data finner du i Tabell 1 i Appendix. I Tabell 2 finner du en utskrift fra R-commander. I Figur 1 i Appendix finner du et histogram.

Skriv en **KORT** rapport der du beskriver resultatene fra analysen. Rapporten skal gi modell med modellantagelser, informasjon om og tolkning av parametre og estimatorer, modellvurdering (hvor god modelltilpasningen er og om modellantagelser er oppfylt), hypotesetest(er) og resultattolkninger, inkludert mulige interessante kontraster.

```
FSS
AnovaModel.1 <- aov(Sukkerinnhold ~ Sort, data=Pear2011)
          Df Sum Sq Mean Sq   F value    Pr(>F)
Sort      (k-1)  3  5.602  1.8672  6.228     0.00367
Residuals (m-k) 20  5.997  0.2998
                         RSS
mean      sd   data:n
KvedeA    12.067 →  $\hat{\mu}_1 = \bar{Y}_1$   0.320   6
Kvedeadams 11.467 →  $\hat{\mu}_2 = \bar{Y}_2$   0.520   6
KvedeC    12.550 →  $\hat{\mu}_3 = \bar{Y}_3$   0.804   6
Pyrodwarf 11.350 →  $\hat{\mu}_4 = \bar{Y}_4$   0.423   6
Estimate Std. Error t value Pr(>|t|) DF
Sort c( 0.33 0.33 0.33 -1 ) 0.678 0.258  2.625 0.016 20
```

Tabell 2. Oppgave 1, utskrift fra R-commander

Sukker	Sort
12.1	KvedeA
12.1	KvedeA
12.4	KvedeA
11.8	KvedeA
11.6	KvedeA
12.4	KvedeA
12.7	KvedeC
12.7	KvedeC
12.3	KvedeC
13.8	KvedeC
12.5	KvedeC
11.3	KvedeC
11.3	Pyrodwarf
11.2	Pyrodwarf
12.1	Pyrodwarf
11.3	Pyrodwarf
10.8	Pyrodwarf
11.4	Pyrodwarf
12.4	kvedeadams
11.5	kvedeadams
11.4	kvedeadams
11.4	kvedeadams
10.8	kvedeadams
11.3	kvedeadams

Tabell 1: Data for oppgave 1.

$$\text{model} := Y_{ij} = \mu_i + \epsilon_{ij} \text{ where } \epsilon_{ij} \sim N(0, \sigma^2) \text{ for } i=1, \dots, k=4 \text{ and } j=1, \dots, m=6$$

$$\text{Devide variances} := Y_{ij} - \bar{Y} = (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i) \rightarrow \sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y})^2 = \underbrace{\sum_{i=1}^k \sum_{j=1}^{m_i} (\bar{Y}_i - \bar{Y})^2}_{\text{TSS}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2}_{\text{FSS or ESS}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2}_{\text{RSS}}$$

$$\text{Estimates} := \hat{\mu}_i = \bar{Y}_i \text{ and } \hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\frac{(m_1-1)}{(m_1-1)} \sum_{j=1}^{m_1} (Y_{1j} - \bar{Y}_1)^2 + \dots + \frac{(m_4-1)}{(m_4-1)} \sum_{j=1}^{m_4} (Y_{4j} - \bar{Y}_4)^2}{m_1 + m_2 + m_3 + m_4 - k}} = \sqrt{\frac{\sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2}{(m-k)}} = \sqrt{\frac{\text{Residual SS}}{(m-k)}} = \boxed{\sqrt{\text{MSE}}}$$

$$\text{Hypotesis} := H_0 := \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_1 := \text{not all } \mu_i \text{ are the same}$$

$$\text{Significance level} := \alpha = 0.05$$

$$F\text{-test} := F = \frac{MSF}{MSE} = \frac{FSS}{RSS} \cdot \frac{(m-k)}{(k-1)} \approx 6.228 \rightarrow \text{Since } F \approx 6.228 > F_{4,6,0.05} = 4.533677 \text{ and } p\text{-value} \approx 0.003 < \alpha = 0.05,$$

from tabel

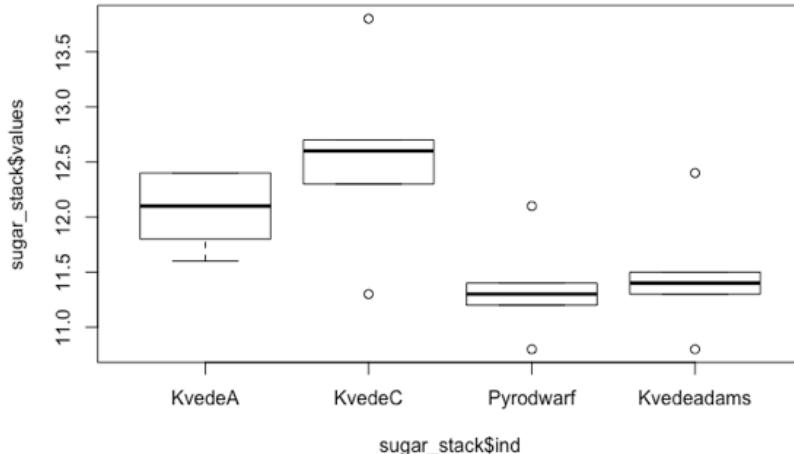
then reject H_0 . * we are 95% secured that not all average groups are the same. \blacksquare

* My calculations in R is in the next page!

```

117 - ````{r}
118 # Exercise 3:
119
120 KvedeA <- c(12.1, 12.1, 12.4, 11.8, 11.6, 12.4)
121 KvedeC <- c(12.7, 12.7, 12.3, 13.8, 12.5, 11.3)
122 Pyrodwarf <- c(11.3, 11.2, 12.1, 11.3, 10.8, 11.4)
123 Kvedeadams <- c(12.4, 11.5, 11.4, 11.4, 10.8, 11.3)
124
125 sugar <- data.frame(KvedeA, KvedeC, Pyrodwarf, Kvedeadams)
126 sugar_stack <- stack(sugar)
127
128 boxplot(sugar_stack$values~sugar_stack$ind)
129 ``

```



```

130 - ````{r}
131 # Exercise 3i):
132
133 ##### Calculating Explained Sum of Square ESS:
134 FSS <- 0
135 k <- 4
136 n <- 6
137
138 for (i in 1:k){
139   FSS <- FSS + (nrow(sugar[i]) * (mean(sugar[, i]) - mean(sugar_stack$values))^2 )
140 }
141
142 FSS_df = k-1
143 MSF <- FSS/FSS_df
144
145 cat("The Explained Sum of Square is ", FSS, "\nThe ESS's degree of freedom is ", FSS_df,
146     "\nThe Mean Square Factor MSF is ", MSF)
147
148 ##### Calculating Residual Sum of Square RSS:
149 RSS <- 0
150
151 for (i in 1:k){
152   for (j in 1:n){
153     RSS <- RSS + (sugar[j, i] - mean(sugar[, i]))^2
154   }
155 }
156
157 RSS_df = k*n - k
158 MSE <- RSS/RSS_df
159
160 cat("\n\nThe Residual Sum of Square is ", RSS, "\nThe RSS's degree of freedom is ", RSS_df,
161     "\nThe Mean Square Error MSE is ", MSE)
162
163 ##### Calculating F-statistic:
164 cat("\n\nThe F-Statistic is ", MSF/MSE)
165 cat("\n\nThe F_0.05_4_6 = ", qf(0.95, 4, 6))
166 ``

```

```

The Explained Sum of Square is  5.601667
The ESS's degree of freedom is 3
The Mean Square Factor MSF is  1.867222

```

```

The Residual Sum of Square is  5.996667
The RSS's degree of freedom is 20
The Mean Square Error MSE is  0.2998333

```

```

The F-Statistic is  6.227534

```

```

The F_0.05_4_6 =  4.533677

```

```

167 - ````{r}
168 model <- lm(sugar_stack$values~sugar_stack$ind)
169 anova(model)
170 ``

```

Analysis of Variance Table

	Response: sugar_stack\$values	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sugar_stack\$ind	3	5.6017	1.86722	6.2275	0.003674 **	
Residuals	20	5.9967	0.29983			

					Signif. codes:	0 **** 0.001 *** 0.01 ** 0.05 .' 0.1 ' ' 1