# Opgave 3

2024-10-18

# Data

```
data1 <- read.table("data/nov2022opg1.txt", header = 1)
head(data1, n = 3)
```

```
##   variety afstand udbytte
## 1       G       8     5.8
## 2       V       7     6.3
## 3      R1       6     4.9
```

```
unique(data1$variety)
```

```
## [1] "G"  "V"  "R1" "F"  "Re" "M"  "E"  "P"
```

# Delopgave 1

Det er den ensidet variansanalyse. Med en kategorisk forklarende variabel variety og en kvantitativ kontinuert responsvariabel udbytte

```
mod1 <- lm(udbytte ~ variety, data = data1)
round(summary(mod1)$coefficients, 4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.550     0.6763  9.6843   0.0000
## varietyF      -0.925     0.9565 -0.9671   0.3432
## varietyG       0.000     0.9565  0.0000   1.0000
## varietyM      -0.950     0.9565 -0.9932   0.3305
## varietyP      -0.050     0.9565 -0.0523   0.9587
## varietyR1     -3.050     0.9565 -3.1887   0.0039
## varietyRe     -0.975     0.9565 -1.0193   0.3182
## varietyV      -0.250     0.9565 -0.2614   0.7960
```

Vores Intercept er E

```
varietyE <- 6.550
varietyR1 <- varietyE + (-3.050)

varietyE
```

```
## [1] 6.55
```

```
varietyR1
```

```
## [1] 3.5
```

Vi for et estimat for middelværdien i hver gruppe

$$\hat{\alpha_{\texttt{E}}} = 6.55, \;\; \hat{\alpha_{\texttt{R1}}} = 3.5$$

# Opgave 2

```
confint(mod1)
```

```
##                 2.5 %     97.5 %
## (Intercept)  5.154084  7.9459156
## varietyF    -2.899123  1.0491228
## varietyG    -1.974123  1.9741228
## varietyM    -2.924123  1.0241228
## varietyP    -2.024123  1.9241228
## varietyR1   -5.024123 -1.0758772
## varietyRe   -2.949123  0.9991228
## varietyV    -2.224123  1.7241228
```

```
E_low <- 5.154084
E_high <- 7.9459156

diff_low <- -5.024123
diff_high <- -1.0758772

R1_low <- E_low + diff_low
R1_high <- E_high + diff_high

E_low
```

```
## [1] 5.154084
```

```
E_high
```

```
## [1] 7.945916
```

```
diff_low
```

```
## [1] -5.024123
```

```
diff_high
```

```
## [1] -1.075877
```

```
R1_low
```

```
## [1] 0.129961
```

```
R1_high
```

```
## [1] 6.870038
```

95 % - konfidensinterval for E (5.15, 7.94)

95 % - konfidensinterval for R1 (0.12, 6.87)

95 % - konfidensinterval for forskellen (-5.02, -1.07)

konfidensintervalet skal forstås således at den sande ukendte populationsmiddelværdi for gruppen E ligger i intervallet (5.15, 7.94). Samme logik følger for de næste to intervaller.

# Opgave 3

Vi laver en anova test for at se om delmodelen kan forklare data bedre. Dette kunne også være gjordt med drop1, i begge tilfælde får vi samme p-værdi * FullModel : En ensidet variansanalyse med variety som forklarende variabel * Nullmodel : End del model af FullModel, som betragtes som en enkelt stikprøve uden en forklarende variabel

```
Fullmodel <- lm(udbytte ~ variety - 1, data = data1)
Nullmodel <- lm(udbytte ~ 1, data = data1)

anova(Nullmodel, Fullmodel)
```

```
## Analysis of Variance Table
##
## Model 1: udbytte ~ 1
## Model 2: udbytte ~ variety - 1
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     31 73.000
## 2     24 43.915  7    29.085 2.2708 0.06342 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi kan ud fra data ikke konkludere at der er forskel. Vi forkaster ikke nulhypotesen $\alpha_{g(1)} = \ldots = \alpha_{g(k)}$ for k grupper, med et signifikansnivuea på 95%, da vi for en p-værdi på over 5%

# Opgave 4

**Data:**
Par $(x_1, y_1), \ldots, (x_n, y_n)$ bestående af kvantitative data, både for vores responsvariabel udbytte og den forklarende variabel afstand. Vi bemærker at udbytte er koninutert imens afstand er diskret.

**Statistisk model:**

$$y_i = \mu_i + \epsilon$$

hvor $\epsilon_i \sim N(0, \sigma^2)$

Vi antager $y_1, \ldots, y_n$ er uafhængihed, og at alle $y_i$ er normaltfordelt med middelværdi $\mu_i = \alpha + \beta x_i$ (ret linje) med spredning $\sigma$.

**De ukendte populationsparametre**
Skæringen $\alpha$, hældningen $\beta$ og spredningen $\sigma$ er alle ukendte parametre, som vi kan estimere: $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$

```
mod2 <- lm(udbytte ~ afstand, data = data1)
summary(mod2)
```

```
##
## Call:
## lm(formula = udbytte ~ afstand, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7750 -0.6214  0.1179  0.7357  1.8679
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.71786    0.43785   8.491 1.78e-09 ***
## afstand      0.45714    0.08671   5.272 1.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.124 on 30 degrees of freedom
## Multiple R-squared:  0.4809, Adjusted R-squared:  0.4636
## F-statistic:  27.8 on 1 and 30 DF,  p-value: 1.08e-05
```

Ud fra summery kan vi se de estimeret parametre

$$\hat{\alpha} = 3.72, \hat{\beta} = 0.46, \hat{\sigma} = 1.12$$

# Opgave 5

Vi opstiller en nulhypotese $H_0 : \beta = 0$

Vi kan forkaste nulhypotesen om at der ikke er nogen linær sammenhæng på et signifikantnivue på 95% med en p-værdi på 1.08e-05 (Meget lidt :D )

Derfor ved vi at nul ikke ligger i 95%-konfindenintervallet for den ukendte sande hældning

```
confint(mod2)
```

```
##                2.5 %    97.5 %
## (Intercept) 2.8236482 4.6120661
## afstand     0.2800631 0.6342226
```

Nul ligger ikke i intervallet (0.28, 0.63)

# Opgave 6

```
data1$afstand_kat <- factor(data1$afstand)
data1$afstand_kat
```

```
##  [1] 8 7 6 5 4 3 2 1 4 3 1 2 5 6 8 7 5 8 6 7 2 1 3 4 4 2 1 3 7 6 8 5
## Levels: 1 2 3 4 5 6 7 8
```

```
mod3 <- lm(udbytte ~ variety + afstand_kat, data = data1)
summary(mod3)
```

```
##
## Call:
## lm(formula = udbytte ~ variety + afstand_kat, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06370 -0.27758 -0.03018  0.39009  1.12169
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.8723     0.8245   4.697 0.000208 ***
## varietyF      -1.0194     0.6182  -1.649 0.117488
## varietyG      -0.4933     0.6583  -0.749 0.463877
## varietyM      -0.4373     0.8006  -0.546 0.591971
## varietyP       0.1362     0.6894   0.198 0.845738
## varietyR1     -2.0440     0.8942  -2.286 0.035376 *
## varietyRe     -1.3319     0.6515  -2.044 0.056744 .
## varietyV      -0.3264     0.5692  -0.573 0.573926
## afstand_kat2   2.1721     0.8085   2.687 0.015606 *
## afstand_kat3   1.9735     0.7239   2.726 0.014370 *
## afstand_kat4   3.0233     0.8006   3.776 0.001506 **
## afstand_kat5   3.3548     0.7750   4.329 0.000456 ***
## afstand_kat6   3.3434     0.5755   5.810 2.09e-05 ***
## afstand_kat7   3.5881     0.7693   4.664 0.000223 ***
## afstand_kat8   3.2826     0.8328   3.942 0.001052 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7725 on 17 degrees of freedom
## Multiple R-squared:  0.861,  Adjusted R-squared:  0.7466
## F-statistic: 7.523 on 14 and 17 DF,  p-value: 9.15e-05
```

Vores intercept er i dette tilfeælde varietyE og afstad_kat1, så estimatet kan aflæses direkte til at være 3.87

```
3.0233 - 2.1721
```

```
## [1] 0.8512
```

Den forventede forskel på udbytter for to områder som ligger henholdvis 2 m og 4 m fra hækken er 0.85

# Opgave 7

```
data1$x <- 1/data1$afstand
mod4 <- lm(udbytte ~ x + variety, data = data1)
summary(mod4)
```

```
##
## Call:
## lm(formula = udbytte ~ x + variety, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16747 -0.45419  0.00321  0.42138  1.30419
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6288     0.3874  19.695 6.76e-16 ***
## x            -3.9832     0.5106  -7.801 6.59e-08 ***
## varietyF     -0.8325     0.5118  -1.627  0.11746
## varietyG     -0.2573     0.5128  -0.502  0.62065
## varietyM     -0.3691     0.5171  -0.714  0.48252
## varietyP      0.4574     0.5158   0.887  0.38441
## varietyR1    -1.8052     0.5360  -3.368  0.00266 **
## varietyRe    -0.9655     0.5117  -1.887  0.07187 .
## varietyV     -0.2322     0.5117  -0.454  0.65422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7237 on 23 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.7776
## F-statistic: 14.55 on 8 and 23 DF,  p-value: 2.496e-07
```

Den forventede værdi er 7.13

```
(7.6288 + 0.4574) + (-3.9832 * (1/4.2))
```

```
## [1] 7.137819
```

```
predict(mod4, newdata = data.frame(x = (1/4.2), variety = "P"), interval = "p")
```

```
##        fit      lwr      upr
## 1 7.137794 5.455567 8.820021
```

Vi for et 95 % - prædiktionsinterval (5.45 8.82)

Hvilket vil sige at 95% af nye obersevationen kan forventes at falde indenfor det interval, givet udbyttet på et område som ligger i afstanden 4.2 m fra hækken og som beplantes med jordbærsorten P.