PC-practicum 5: one-way anova

Uitgewerkt voorbeeld ANOVA: de koekoek

Het is bekend dat de koekoek niet zelf een nest bouwt maar zijn eieren legt in de nesten van andere vogels. Sinds 1892 weet men reeds dat het soort koekoekseieren eigen is aan de locatie waar ze gevonden worden. Een studie in 1940 toonde aan dat de koekoeken elk jaar terugkeren naar hetzelfde grondgebied en eieren leggen in de nesten van welbepaalde "pleegouder"-vogels. Bovendien paren koekoeken enkel binnen hun grondgebied. Op die manier zijn geografische subsoorten ontwikkeld, elk met een dominante pleegouder-soort. Hierdoor kon een specialisatie optreden van de koekoek aan de pleegouder-soort via natuurlijke selectie, zodat de koekoekseieren een hogere kans kregen om geadopteerd te worden door de pleegouder-soort.

De dataset koekoek.txt bevat de lengte (variabele `lengte`) van de koekoekseieren (in mm) van willekeurig gekozen geparasiteerde nesten. In totaal bevat de dataset 120 observaties en voor elk ei is aangegeven van welke vogelsoort (variabele `soort`) het nest is. De codering voor soort is als volgt:

- `soort=1`: graspieper
- `soort=2`: boompieper
- `soort=3`: heggemus
- `soort=4`: roodborstje
- `soort=5`: witte kwikstaart
- `soort=6`: winterkoning

In deze analyse zullen we nagaan of de pleegouder-soort een invloed heeft op de gemiddelde lengte van de koekoekseieren.

Libraries laden

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#install.packages("tidyr")
library(tidyr)
#install.packages("multcomp")
library(multcomp)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

Dataset koekoek.txt inlezen

```
koekoek<-read.table("https://raw.githubusercontent.com/statOmics/statistiekBasisCursusData/master/practi
head(koekoek)
```

```
##   lengte soort
## 1  21.87     1
## 2  22.88     1
## 3  24.61     1
## 4  22.95     1
## 5  19.55     1
## 6  22.42     1
```

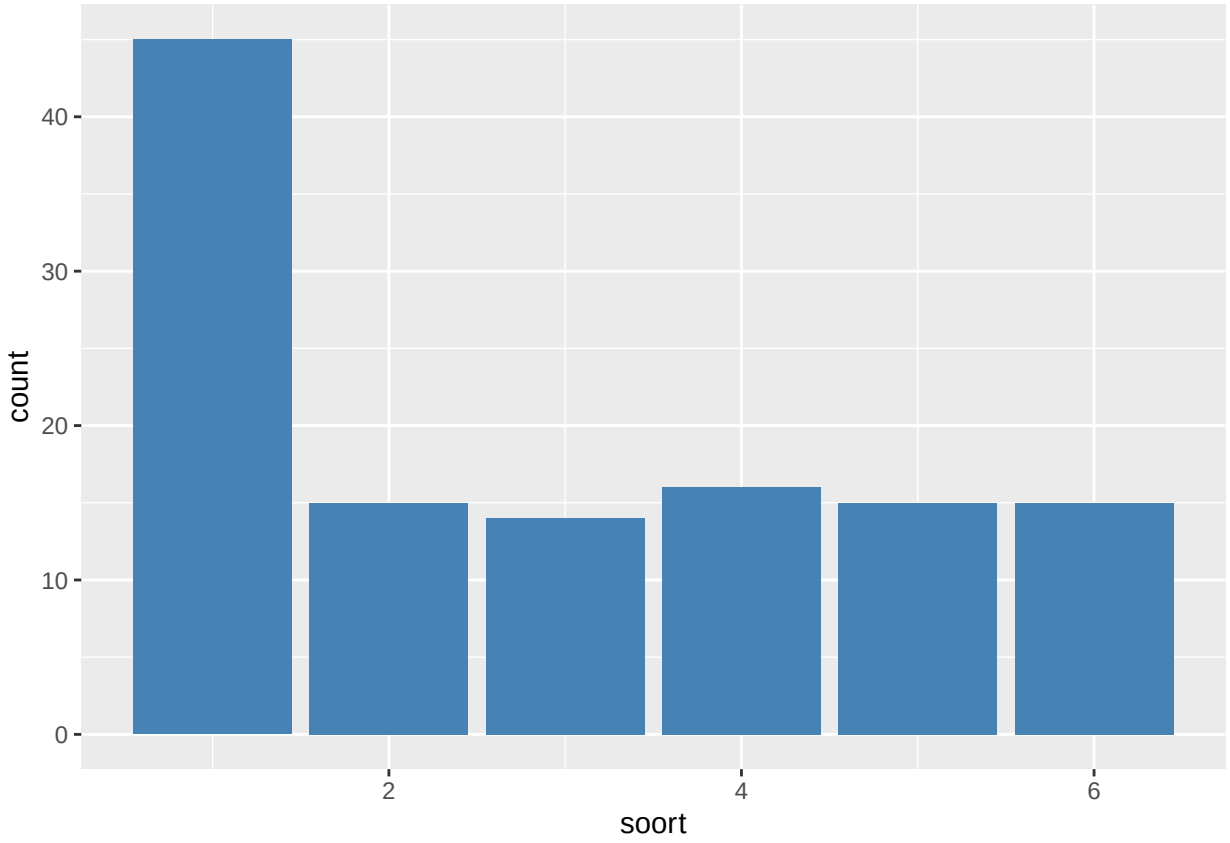
Hoeveel observaties zijn er voor elke soort?

Tel het aantal observaties per soort en sla het resultaat op in `count`. Maak een barplot voor de variabele `soort`.

```
count <- koekoek %>% count(soort)
count
```

```
##   soort  n
## 1     1 45
## 2     2 15
## 3     3 14
## 4     4 16
## 5     5 15
## 6     6 15
```

```
koekoek %>% ggplot(aes(x = soort)) + geom_bar(fill = "steelblue")
```

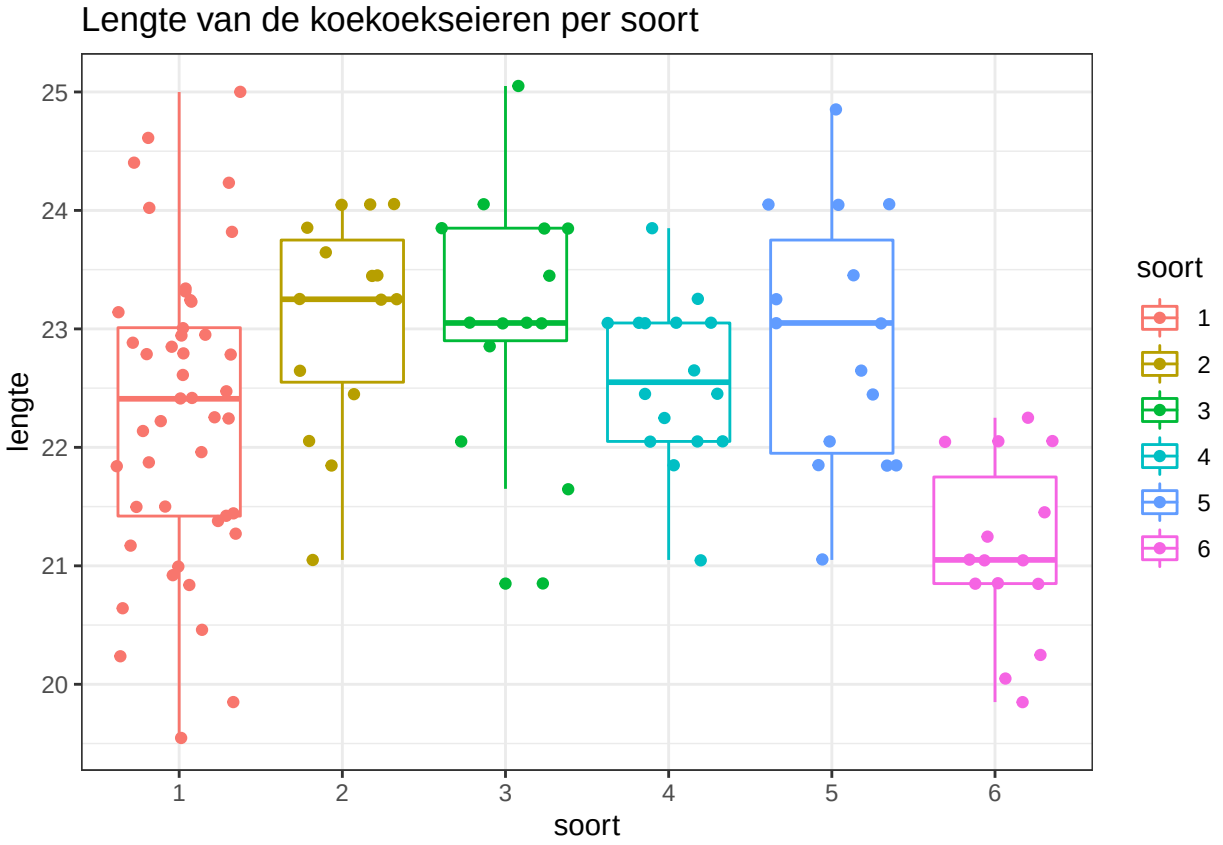


Data-exploratie

Genereer een boxplot van de lengte van de koekoekseieren voor elk van de 6 vogelsoorten. Plot ook de individuele observaties.

```
koekoek$soort <- as.factor(koekoek$soort)
boxplot <- ggplot(data=koekoek,aes(x=soort, y=lengte, col=soort)) +
  geom_boxplot() +
  geom_jitter() +
  theme_bw() +
```

```
  ggtitle("Lengte van de koekoekseieren per soort")
boxplot
```



Welke test kan men uitvoeren om de gemiddelde lengte simultaan te vergelijken tussen alle soorten?

In vorige lessen zagen we enkel de two-sample t-test om twee gemiddelden met elkaar vergelijken. We hebben echter ook reeds gezien dat de two-sample t-test een specifieke versie is van een lineair model, namelijk van een lineair model waarbij de covariaat een categorische variabele X is met 2 levels, i.e.

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Bijvoorbeeld, indien Y_i de lengte van persoon i voorstelt en X_i het geslacht van die persoon waarbij $X_i = 0$ indien persoon i een vrouw is, en $X_i = 1$ indien niet. In dat geval, stelt β_0 de gemiddelde lengte voor vrouwen voor, en β_1 staat voor het verschil in gemiddelde lengte tussen vrouwen en mannen. De gemiddelde lengte voor een man kan men dan bekomen door $E(Y_{male}) = \beta_0 + \beta_1$. Men kan dit ook schrijven als

$$E(Y|female) = \beta_0$$

$$E(Y|male) = \beta_0 + \beta_1$$

Dit lineair model kan echter ook makkelijk veralgemeend worden naar factoren met meerdere levels. Er bestaat echter ook een manier waarbij we **alle levels simultaan kunnen testen**, men zal namelijk testen of de gehele factor variabele een invloed heeft op de respons. In de context van ons voorbeeld, zal men kunnen testen of de pleegouder-soort uberhaupt een effect heeft op de gemiddelde lengte van koekoekseieren. Zo'n

een test heet een one-way ANOVA. Men noemt de test 'one-way' omdat het enkel 'main effects' test, met andere woorden het model bevat geen interacties (enkel relevant voor studenten biochemie en biomedische wetenschappen).

Geef de nul- en alternatieve hypothese voor de test

Stel dat μ_1 de gemiddelde lengte van koekoekseieren voor graspiepers (`soort=1`) voorstelt, en idem voor $\mu_2, \ldots, \mu_6$. De nul- en alternatieve hypothese voor een ANOVA kan men dan voorstellen als

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

H_A: Voor minstens één $i \neq j$ is $\mu_i \neq \mu_j$

In woorden, zegt de nulhypothese dat de gemiddelde lengte van koekoekseiren onafhankelijk is van de pleegouder-soort: er is geen systematisch verschil in gemiddelde lengte van het ei. De alternatieve hypothese zegt dat de gemiddelde lengte verschilt tussen **minstens twee pleegouder-soorten**. Merk op dat men bij het verwerpen van de nulhypothese **niet weet tussen welke soorten** er een verschil is!

Fit het model voor de analyse

We fitten een lineair model met als afhankelijke variabele de lengte van de eieren en als onafhankelijke variabele de soort. Merk op dat het belangrijk is om soort op te nemen als factor, wat al in orde werd gebracht bij het genereren van de boxplots.

```
m <- lm(lengte~soort, data = koekoek)
summary(m)
```

```
##
## Call:
## lm(formula = lengte ~ soort, data = koekoek)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7376 -0.7406  0.0975  0.6869  2.7124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.2876     0.1571 141.862  < 2e-16 ***
## soort2        0.8024     0.3142   2.554 0.011975 *
## soort3        0.8339     0.3225   2.585 0.010985 *
## soort4        0.2874     0.3068   0.937 0.350725
## soort5        0.6158     0.3142   1.960 0.052467 .
## soort6       -1.1576     0.3142  -3.684 0.000353 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.054 on 114 degrees of freedom
## Multiple R-squared:  0.254,  Adjusted R-squared:  0.2213
## F-statistic: 7.762 on 5 and 114 DF,  p-value: 2.576e-06
```

De output van het model suggereert dat er inderdaad verschillen lijken te zijn in gemiddelde lengte tussen de pleegoudersoorten. Merk op dat in de standaard output op basis van dit model de p-waarden echter niet aangepast worden voor meervoudig toetsen.

Ga de assumpties voor een ANOVA na.

Merk op dat de assumpties checken niet steeds eenduidig zijn, zeker niet bij een ANOVA test. Zoals beschreven in de cursus, veronderstelt ANOVA een locatie-shift model. Dit wil zeggen dat elke groep een gelijke distributie heeft en er enkel shifts in gemiddelde kunnen optreden. In het bijzonder nemen we de aanname dat elke groep een normale verdeling zou volgen. Dit impliceert dat

- elke groep normaal verdeeld moet zijn.
- elke groep een gelijke variantie moet hebben.

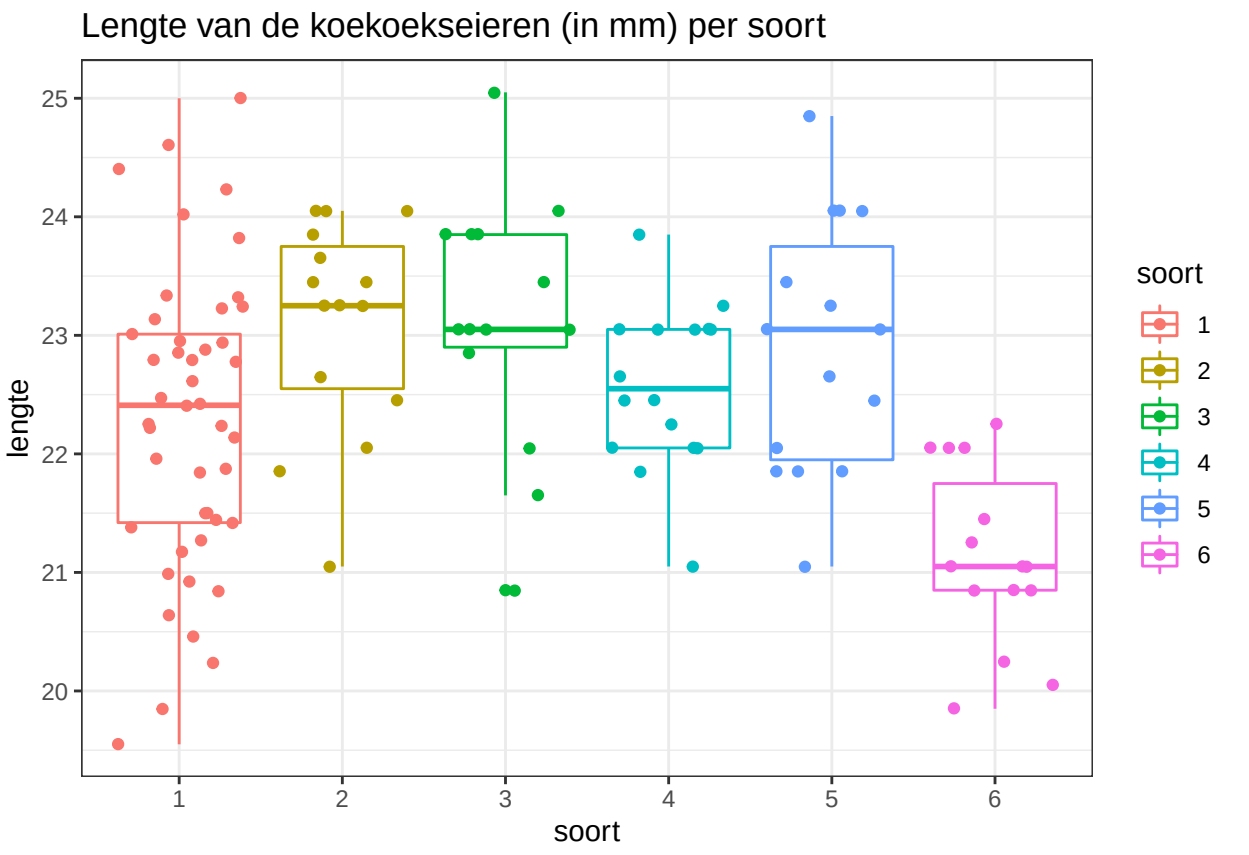
Bovendien neemt de test nog aan dat

- de groepen onafhankelijk zijn van elkaar.
- de gegevens binnen een groep onafhankelijk zijn van elkaar.

Deze laatste twee assumpties zijn voldaan; de nesten werden willekeurig gekozen. De eerste twee assumpties kunnen we nagaan indien er niet te veel groepen zijn. Hier hebben we zes groepen en is het checken van assumpties binnen elke groep nog doenbaar.

Voor het nagaan van homoscedasticiteit werken we met boxplots:

```
boxplot <- ggplot(data=koekoek,aes(x=soort, y=lengte, col=soort)) +
  geom_boxplot() +
  geom_jitter() +
  theme_bw() +
  ggtitle("Lengte van de koekoekseieren (in mm) per soort")
boxplot
```



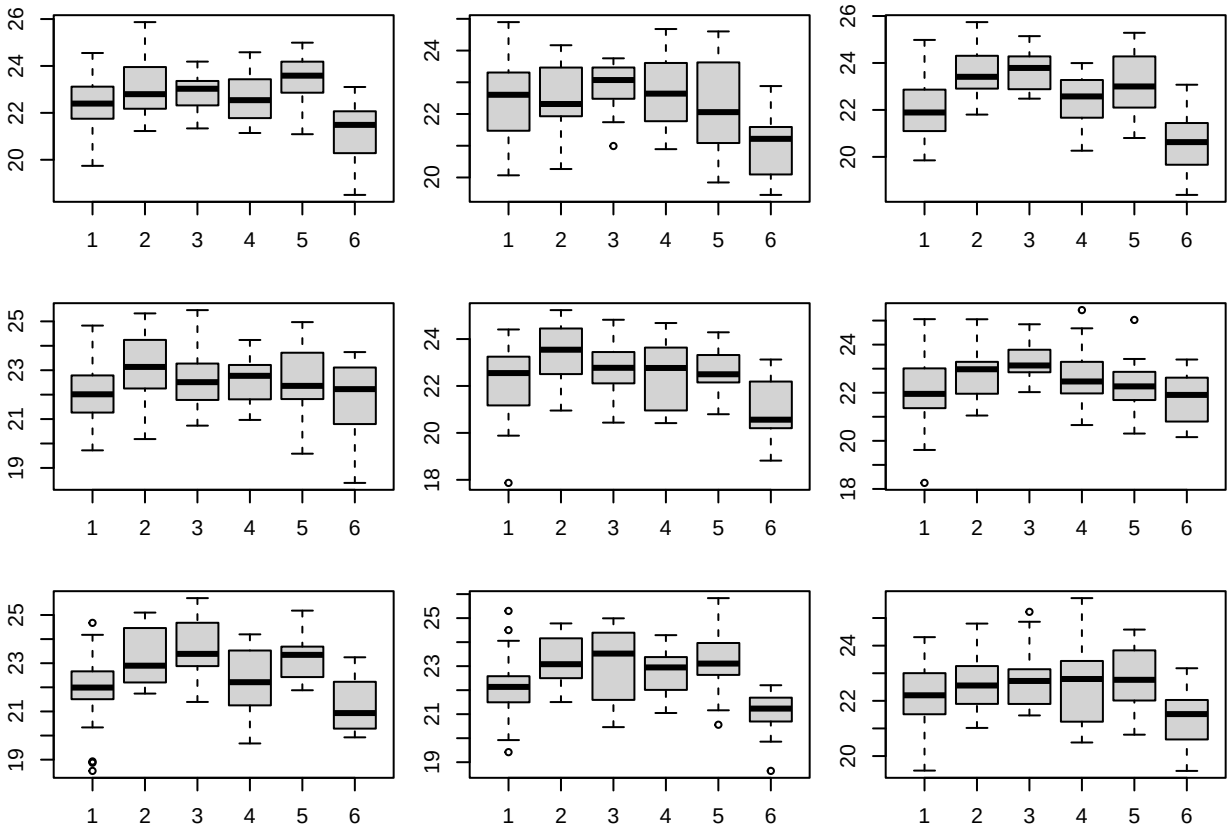
De data lijken gelijke varianties te hebben, en deze assumptie lijkt alvast niet geschonden.

Het is echter niet altijd eenvoudig om te beoordelen of de varianties sterk van mekaar verschillen of niet. Om een beter idee te krijgen, kunnen we eens een aantal boxplots simuleren met dezelfde steekproefgrootte als in de dataset en in de veronderstelling dat de varianties gelijk zijn.

```
set.seed(52)
par(mfrow=c(3,3), mar=c(3,2,1,1))
sd1<-koekoek %>% pull(lengte) %>%sd()
means<- koekoek %>% group_by(soort) %>% summarise(m=mean(lengte))
means$m
```

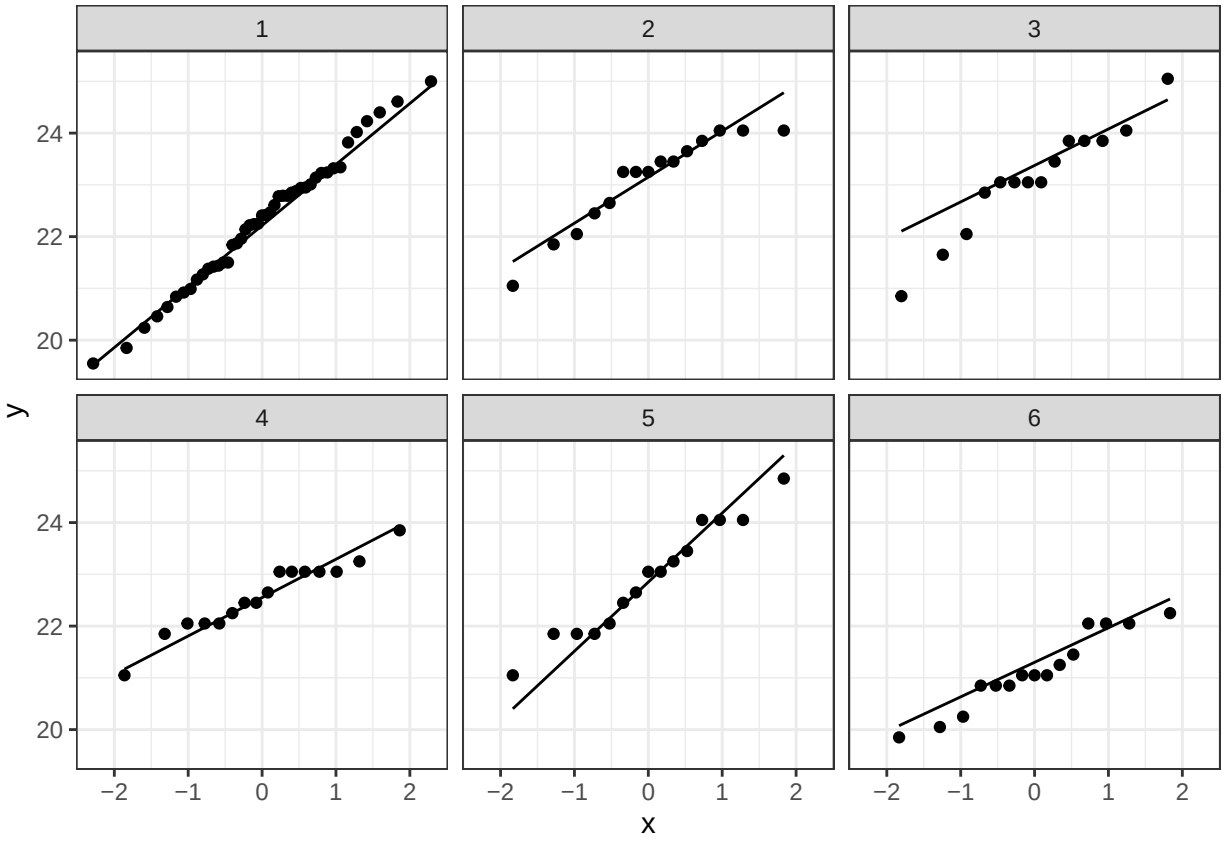
```
## [1] 22.28756 23.09000 23.12143 22.57500 22.90333 21.13000
```

```
for(i in 1:9){
  a<-rnorm(45, mean=means$m[1], sd=sd1)
  b<-rnorm(15, mean=means$m[2], sd=sd1)
  c<-rnorm(14, mean=means$m[3], sd=sd1)
  d<-rnorm(16, mean=means$m[4], sd=sd1)
  e<-rnorm(15, mean=means$m[5], sd=sd1)
  f<-rnorm(15, mean=means$m[6], sd=sd1)
  boxplot(a,b,c,d,e,f)
}
```



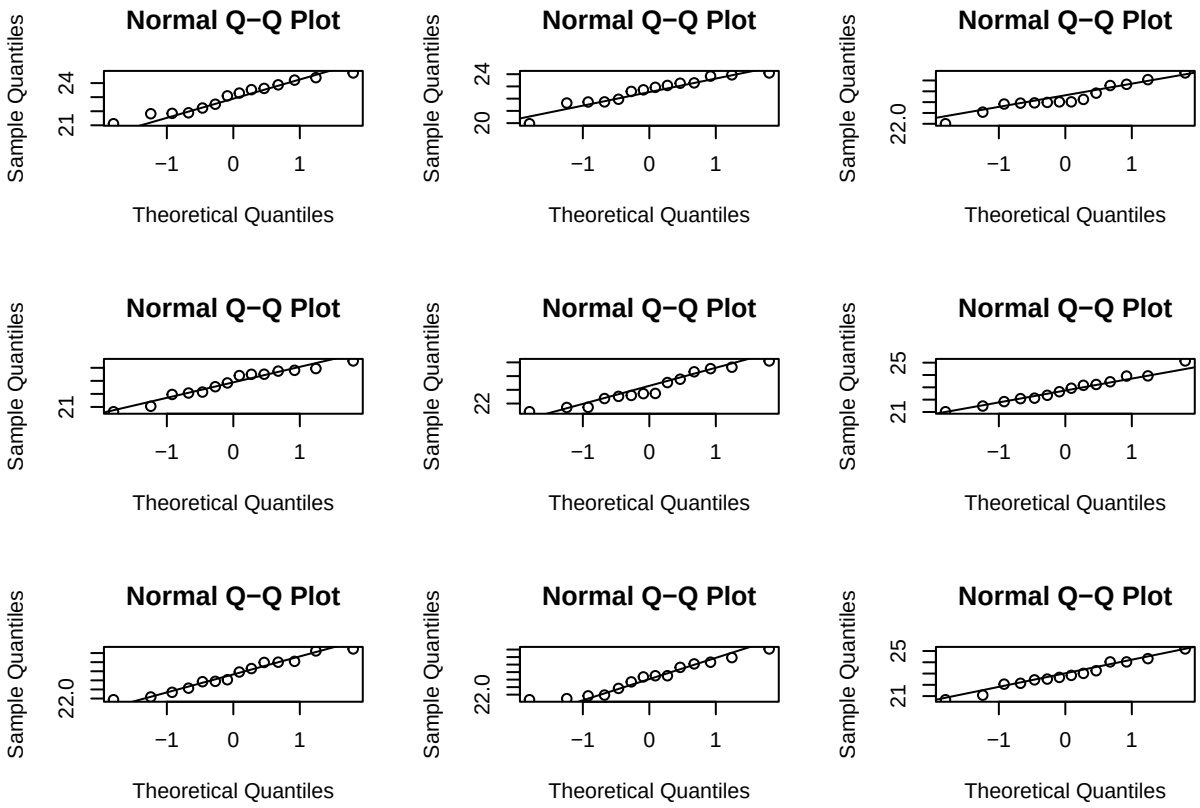
We zullen nu de assumptie van normale verdeling binnen elke groep nagaan:

```
# Maak QQ-plot voor de lengte van de koekoekseieren per soort
plot_qq <- koekoek %>% ggplot(aes(sample = lengte)) +
                       geom_qq() + # qq-punten
                       geom_qq_line() + # qq-lijn
                       theme_bw() +
                       facet_wrap(~soort)
plot_qq
```



Bij de derde soort lijkt op het eerste zicht niet volledig voldaan aan normaliteit. We stelden eerder al vast dat soort drie slechts 14 observaties bevat. We kunnen opnieuw eens vergelijken met gesimuleerde data onder de nulhypothese. De afwijkingen die we in onze qqplot zien lijken niet zeer uitzonderlijk te zijn. Ook in de gesimuleerde data lijken sommige punten in de staart af te wijken van de rechte op de plot.

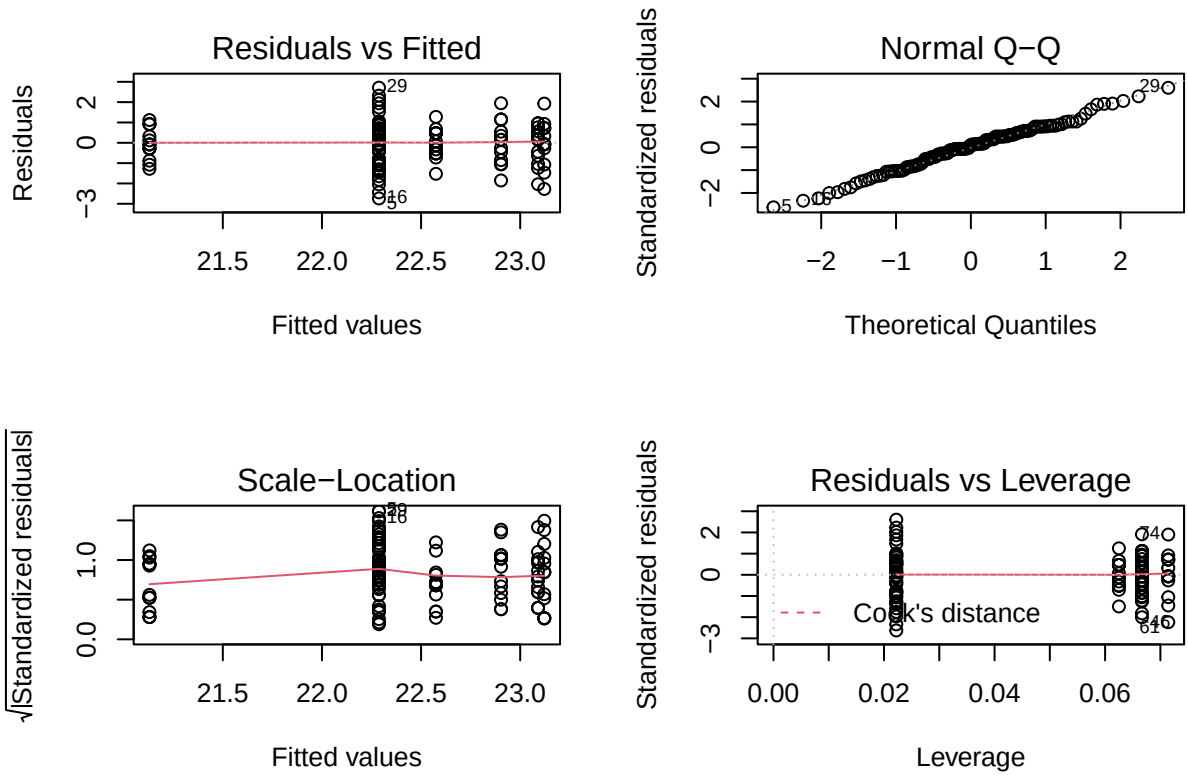
```
par(mfrow=c(3,3))
for(i in 1:9){
  x <- rnorm(n=14, mean=koekoek %>% filter(soort==3) %>%pull(lengte) %>%mean(),
             sd=koekoek %>% filter(soort==3) %>%pull(lengte) %>%sd())
  qqnorm(x)
  qqline(x)
}
```



Elke groep lijkt een normale verdeling te volgen en deze assumptie is ook voldaan.

Indien men veel groepen moet vergelijken, kan het efficiënter zijn om slechts één plot te moeten beoordelen. In dat geval kan men ervoor kiezen om niet voor elke groep apart een QQ-plot te maken, maar kan men de residuen van het lineair model checken. Merk op dat men dan checkt voor een normale distributie van alle residuen van de respons variabele ten opzichte van hun groepsgemiddelde, en dus niet voor een normale distributie binnen elke groep.

```
par(mfrow=c(2,2))
plot(m) # Enkel figuur rechts boven is relevant
```




```
par(mfrow=c(1,1))
```

De QQ-plot vertoont geen systematische afwijkingen van een normale distributie. Het nagaan van de assumpties op deze manier vormt zeker geen sluitend bewijs (dat hebben we namelijk zelden) dat de data normaal verdeeld is binnen elke groep, maar het is een benadering die we kunnen gebruiken in het geval dat

- er te veel groepen zijn om de assumpties te checken binnen elke groep;
- er te weinig observaties zijn per groep om binnen elke groep de assumpties na te gaan.

Merk op dat je in principe de assumptie van homoscedasticiteit ook op basis van de plot linksboven zou kunnen checken: elke 'kolom' van punten stelt een soort voor (1 soort heeft 1 geschat gemiddelde) en de punten stellen de residuen voor ten opzichte van hun groepsgemiddelde. Men kan deze plot dus ook gebruiken om te kijken of er groepen (soorten) zijn die een verschillende variantie hebben ten opzichte van andere groepen.

Voer de ANOVA uit

We voeren de ANOVA test uit aan de hand van het lineair regressiemodel. In principe testen we dan volgende nulhypothese

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

met de alternatieve hypothese dat minstens twee regressieparameters verschillen van elkaar.

Merk op dat deze nulhypothese evenwaardig is aan de nulhypothese die we eerder formuleerden. Als alle regressieparameters $\beta_1, \ldots, \beta_5 = 0$, betekent dit dat er geen verschil is tussen de 6 groepsgemiddelde lengtes.

```
anova(m)
```

```
## Analysis of Variance Table
##
## Response: lengte
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## soort       5  43.108  8.6215  7.7621 2.576e-06 ***
## Residuals 114 126.622  1.1107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De p-waarde van deze ANOVA test is bijzonder klein. We besluiten dat we de nulhypothese kunnen verwerpen ($p << 0.001$) en dat de gemiddelde lengte van koekoekseieren verschilt tussen minstens twee van de bestudeerde pleegoudersoorten op het 5% significantieniveau.

Aan de hand van dit resultaat weten we echter niet tussen welke soorten er een verschil optreedt, en hiervoor zal men een **post-hoc analyse** moeten uitvoeren. Een post-hoc analyse voert men enkel uit indien de ANOVA test significant was, en bestaat erin om paarsgewijze vergelijkingen uit te voeren tussen de groepen.

Post-hoc analyse

De post-hoc analyse bestaat eruit om paarsgewijze testen uit te voeren. Indien men over k groepen beschikt is het totaal aantal paarsgewijze vergelijkingen gelijk aan $k(k-1)/2$. Bij ons is $k = 6$ waardoor we 15 paarsgewijze vergelijkingen zullen uitvoeren. We kunnen echter niet elke test op het 5% significantieniveau uitvoeren vanwege het meervoudig toetsen probleem. Inderdaad, indien men 15 vergelijkingen zou doen, elk op het 5% significantieniveau, dan is de kans dat we minstens één nulhypothese zouden verwerpen terwijl die eigenlijk waar is niet langer gelijk aan ons significantieniveau (5%). In ons geval, zou deze kans gelijk zijn aan

```
alpha <- 0.05
nComparisons <- 15
1-(1-alpha)^nComparisons
```

```
## [1] 0.5367088
```

Dus indien we elke test op het 5% significantieniveau zouden uitvoeren hebben we, als alle nulhypotheses waar zouden zijn, een kans van $\sim 54\%$ dat we minstens één nulhypothese verkeerd zouden verwerpen! Om deze kans globaal gezien (dit is, over alle paarsgewijze vergelijkingen) op 5% te houden, kunnen we bijvoorbeeld de Bonferroni correctie uitvoeren.

In `R` kunnen we de post-hoc analyse uitvoeren met behulp van het `multcomp` package aan de hand van de `glht` functie. We specifiëren hier in het `linfct` argument dat we *multiple comparisons* (`mcp`) willen uitvoeren waarbij we alle paarsgewijze vergelijkingen voor de `soort` variabele willen testen aan de hand van de `"Tukey"` methode. Het resultaat van deze test slaan we vervolgens op in het object `mcp`, waarop we een `summary` opvragen van dat object. Het `multcomp` package zorgt ervoor dat deze p-waarden automatisch gecorrigeerd worden voor meervoudig toetsen aan de hand van een efficiënte manier (efficiënter dan de Bonferroni methode).

```
library(multcomp)
mcp <- glht(m,linfct=mcp(soort="Tukey"))
summary(mcp)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = lengte ~ soort, data = koekoek)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0  0.80244    0.31421   2.554  0.11433
## 3 - 1 == 0  0.83387    0.32252   2.585  0.10661
## 4 - 1 == 0  0.28744    0.30676   0.937  0.93405
## 5 - 1 == 0  0.61578    0.31421   1.960  0.36551
## 6 - 1 == 0 -1.15756    0.31421  -3.684  0.00456 **
## 3 - 2 == 0  0.03143    0.39164   0.080  1.00000
## 4 - 2 == 0 -0.51500    0.37877  -1.360  0.74536
## 5 - 2 == 0 -0.18667    0.38483  -0.485  0.99648
## 6 - 2 == 0 -1.96000    0.38483  -5.093  < 0.001 ***
## 4 - 3 == 0 -0.54643    0.38569  -1.417  0.71096
## 5 - 3 == 0 -0.21810    0.39164  -0.557  0.99327
## 6 - 3 == 0 -1.99143    0.39164  -5.085  < 0.001 ***
## 5 - 4 == 0  0.32833    0.37877   0.867  0.95210
## 6 - 4 == 0 -1.44500    0.37877  -3.815  0.00288 **
## 6 - 5 == 0 -1.77333    0.38483  -4.608  < 0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

In de output hiervan zien we de verschillende paarsgewijze vergelijkingen die werden uitvoerd, bijvoorbeeld `2 - 1 == 0` duidt erop dat voor dit contrast wordt getest of het gemiddelde voor soort `2` min het gemiddelde voor soort `1` gelijk is aan nul of niet. In de tweede kolom wordt het verschil in gemiddelden weergegeven, met hun standaard error en teststatistiek in de respectievelijk derde en vierde kolom. De laatste kolom geeft aangepaste p-waarden weer op een globaal significantieniveau van 5%. Aan de hand van de aangepaste p-waarden zien we dat de gemiddelde lengte van soort 6 (winterkoning) verschilt van alle andere soorten. De effectgrootte is voor alle soorten negatief, hetgeen erop duidt dat de gemiddelde lengte van koekoekseieren lager is in nesten van winterkoning in vergelijking met andere soorten.

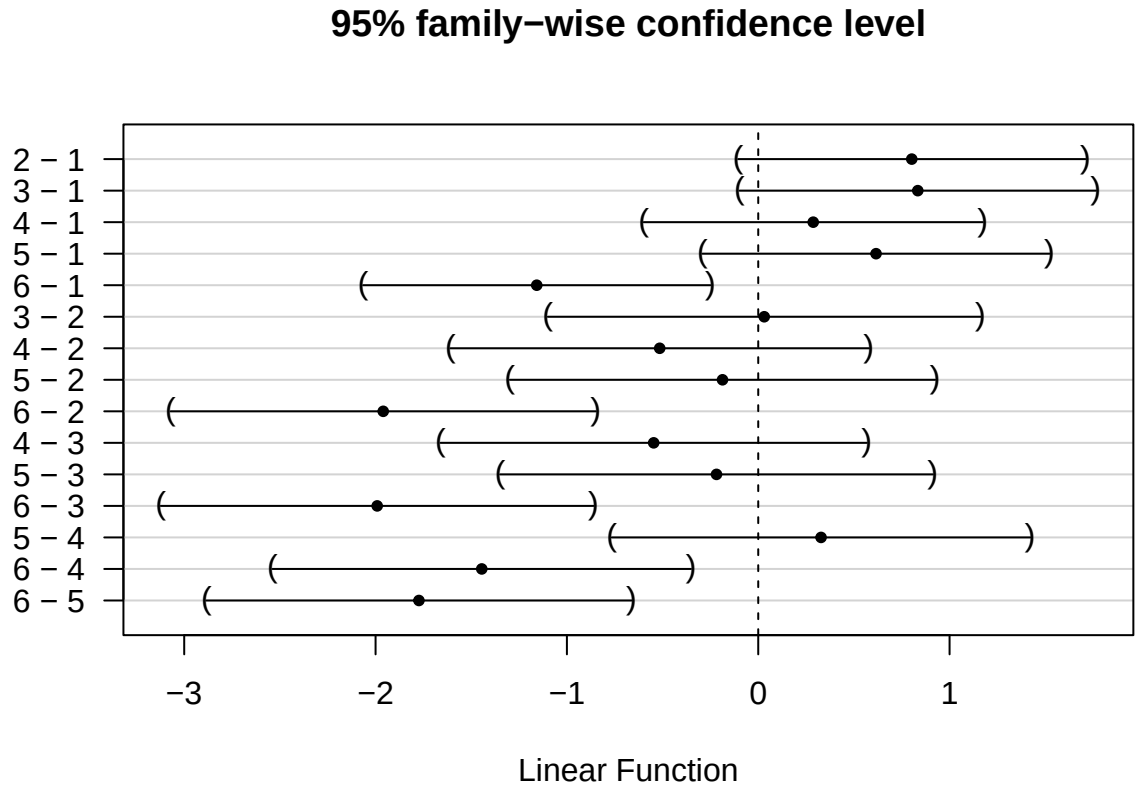
Voor de rapportering zullen we ook betrouwbaarheidsintervallen voor elke paarsgewijze vergelijking opvragen. We kunnen deze ook makkelijk grafisch voorstellen aan de hand van de nuttige `plot` functie die zo op een `glht` object kan toegepast worden.

```
confint(mcp)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
```

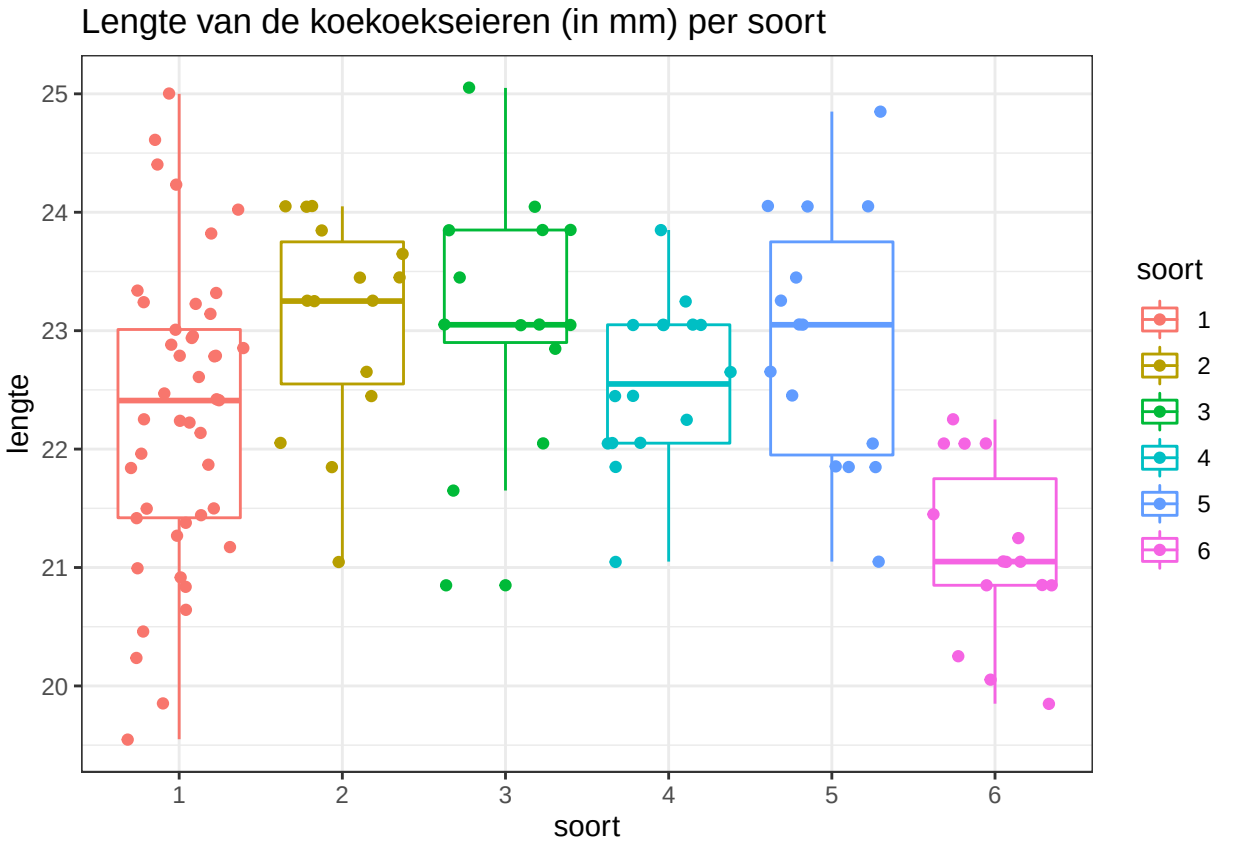
```
##
## Fit: lm(formula = lengte ~ soort, data = koekoek)
##
## Quantile = 2.8888
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##             Estimate lwr      upr
## 2 - 1 == 0  0.80244 -0.10525  1.71014
## 3 - 1 == 0  0.83387 -0.09781  1.76556
## 4 - 1 == 0  0.28744 -0.59872  1.17361
## 5 - 1 == 0  0.61578 -0.29191  1.52347
## 6 - 1 == 0 -1.15756 -2.06525 -0.24986
## 3 - 2 == 0  0.03143 -1.09994  1.16280
## 4 - 2 == 0 -0.51500 -1.60918  0.57918
## 5 - 2 == 0 -0.18667 -1.29836  0.92502
## 6 - 2 == 0 -1.96000 -3.07169 -0.84831
## 4 - 3 == 0 -0.54643 -1.66060  0.56774
## 5 - 3 == 0 -0.21810 -1.34946  0.91327
## 6 - 3 == 0 -1.99143 -3.12280 -0.86006
## 5 - 4 == 0  0.32833 -0.76585  1.42252
## 6 - 4 == 0 -1.44500 -2.53918 -0.35082
## 6 - 5 == 0 -1.77333 -2.88502 -0.66164
```

```
plot(mcp)
```



Men kan de resultaten zelf ook nog eens bekijken op basis van de ruwe data:

```
boxplot <- ggplot(data=koekoek,aes(x=soort, y=lengte, col=soort)) +
  geom_boxplot() +
  geom_jitter() +
  theme_bw() +
  ggtitle("Lengte van de koekoekseieren (in mm) per soort")
boxplot
```



Besluit

We vinden een extreem significante afhankelijkheid tussen de gemiddelde lengte van koekoekseieren en de pleegoudersoort (one-way ANOVA test, $p << 0.001$). Op een globaal 5% significantieniveau vinden we verschillen in gemiddelde lengte van de koekoekseieren tussen verschillende soorten. De gemiddelde lengte van koekoekseieren in nesten van winterkoning is kleiner in vergelijking met deze in nesten van alle andere bestudeerde soorten: graspieper (Tukey test, verschil=-1.16, aangepaste p-waarde = 0.005, 95% BI: [-2.07 ; -0.25]), boompieper (Tukey test, verschil=-1.96, aangepaste p-waarde < 0.001 , 95% BI: [-3.07 ; -0.85]), heggenmus (Tukey test, verschil=-1.99, aangepaste p-waarde < 0.001 , 95% BI: [-3.12 ; -0.86]), roodborstje (Tukey test, verschil=-1.45, aangepaste p-waarde = 0.003 , 95% BI: [-2.53 ; -0.35]) en witte kwikstaart (Tukey test, verschil=-1.77, aangepaste p-waarde < 0.001 , 95% BI: [-2.88 ; -0.66]).

We vinden onvoldoende bewijs voor een verschil in gemiddelde lengte van de koekoekseieren tussen de overige soorten.