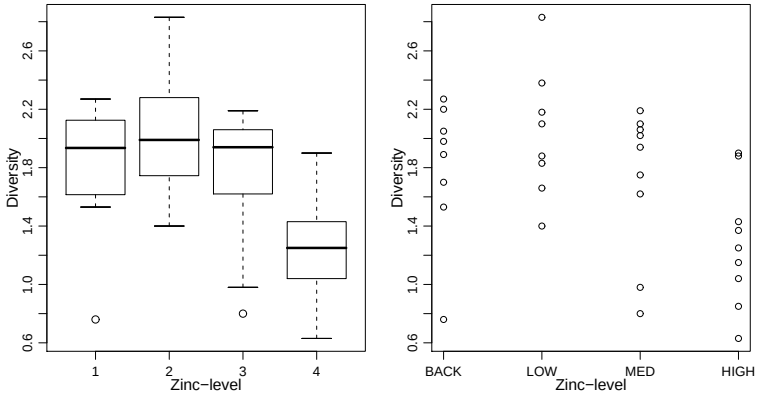
Varianzanalyse & Versuchsplanung — Musterlösungen zur Serie 1

1. **a)** Erste Informationen über die Daten erhalten wir mit

```
> summary(d.stream)
   STREAM     ZINC      DIVERSITY      ZNGROUP
 Arkan:7   BACK:8   Min.   :0.630   1:8
 Blue :7   HIGH:9   1st Qu.:1.377   2:8
 Chalk:5   LOW :8   Median :1.855   3:9
 Eagle:4   MED :9   Mean   :1.694   4:9
 Snake:5            3rd Qu.:2.058
 Splat:6            Max.   :2.830
```

Die Anzahl Daten pro Gruppe (zwei mal 8 und zwei mal 9) ist an der unteren Grenze dafür, dass ein Boxplot hilfreich ist. Man sieht die Verteilung besser im Streudiagramm.



Im Boxplot werden zwei Ausreisser dargestellt. Im Streudiagramm scheinen diese zwei Punkte nicht klare Ausreisser zu sein.

```
> plot(DIVERSITY~ZNGROUP, data=d.stream)
> d.stream[,"ZNGROUP"] <- as.factor(d.stream[,"ZNGROUP"])
> plot(DIVERSITY~ZNGROUP, data=d.stream)
```

Bemerkung: Man kann auch mit der Variablen `ZINC` arbeiten, dann wird jedoch die Reihenfolge verändert (nach Alphabet). Sie werden noch lernen, wie man spezifische Reihenfolgen definieren kann.

b) R-Befehle:

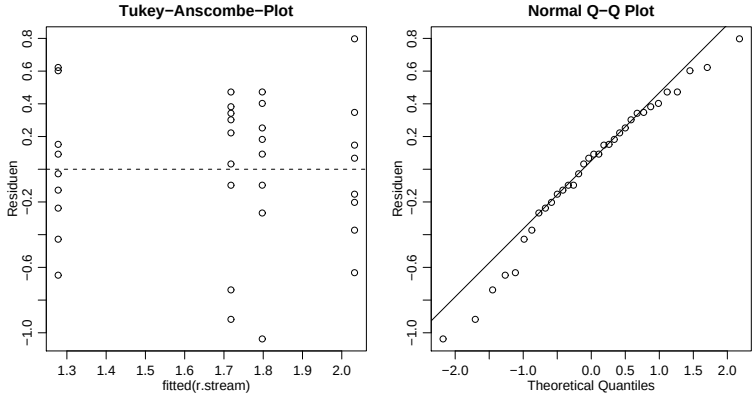

```
> r.stream <- aov(DIVERSITY ~ ZNGROUP, data = d.stream)
> summary(r.stream)
            Df Sum Sq Mean Sq F value  Pr(>F)
ZNGROUP      3 2.5666  0.8555  3.9387 0.01756 *
Residuals   30 6.5164  0.2172
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Der F-Test ergibt ein signifikantes Resultat mit einem P-Wert von 0.018.

c) R-Befehle:

```
> op <- par(mfrow = c(1,2))
> plot(resid(r.stream) ~ fitted(r.stream),
+      main = "Tukey-Anscombe-Plot", ylab = "Residuen")
> abline(h = 0, lty = 2)
> qqnorm(resid(r.stream), ylab = "Residuen")
> qqline(resid(r.stream))
> par(op)
```

Im ersten Befehl richtet man das Grafikfenster so ein, dass zwei Plots nebeneinander erscheinen und speichert gleichzeitig die vorherigen Einstellungen im Objekt `op` (für old par) ab. Im letzten Befehl werden wieder die ursprünglichen Einstellungen aktiviert.



Weder im Tukey-Anscombe-Plot noch im Normal Plot sind Abweichungen von den Modellannahmen erkennbar.

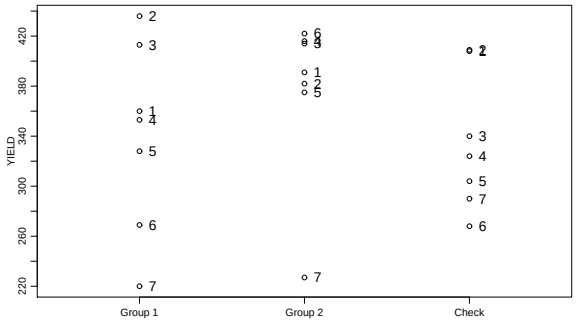
d) R-Befehl:

```
> dummy.coef(r.stream)
Full coefficients are

(Intercept):         1.7975
ZNGROUP:                1          2          3          4
                0.00000000 0.23500000 -0.07972222 -0.51972222
```

Hier sehen wir, dass $\hat{\alpha_1}$ auf 0 gesetzt wird, und somit ist $\hat{\mu_1} = 1.7975$, $\hat{\alpha_2} = 0.235 \Rightarrow \hat{\mu_2} = 1.7975 + 0.235 = 2.0325$, usw.

2. a) Wir betrachten zuerst den Dotplot von `YIELD` gegen `GROUP` mit der Replikatsnummer als Label.



In Gruppe 1 haben die Beobachtungen eine leicht grössere Streuung als bei den anderen Gruppen. Ausser dem 6. und 4. Replikat in Gruppe 2 haben Beobachtungen mit kleiner Replikatsnummer einen grösseren Ertrag als Beobachtungen mit grosser Replikatsnummer. In Gruppe 2 fällt eine Beobachtung mit einem sehr kleinen Ertrag auf. Es ist Replikat 7 mit einem Wert von 227.

Bemerkung: Mit so wenigen Beobachtungen macht es keinen Sinn Boxplots zu zeichnen. Die Verteilung der Punkte ist mit einem Dotplot besser zu sehen.
R-Befehl:

```
# Die Variable GROUP im Dataframe d.hafer ist hier numerisch:
> plot(d.hafer$GROUP,d.hafer$YIELD,xlim=c(0.5,3.5),
     xaxt="n",xlab="",ylab="YIELD")
> axis(1,at=c(1,2,3),labels=c("Group 1","Group 2","Check"))
> text(d.hafer$GROUP+0.08,d.hafer$YIELD,labels=d.hafer$REP,cex=1.3)
```

b) R-Befehl:

```
> d.hafer$GROUP <- factor(d.hafer$GROUP)
> r.hafer <- aov(YIELD ~ GROUP, data = d.hafer)
> summary(r.hafer)
            Df Sum Sq Mean Sq F value Pr(>F)
GROUP        2   6831    3416  0.7614 0.4815
Residuals   18  80744    4486
```

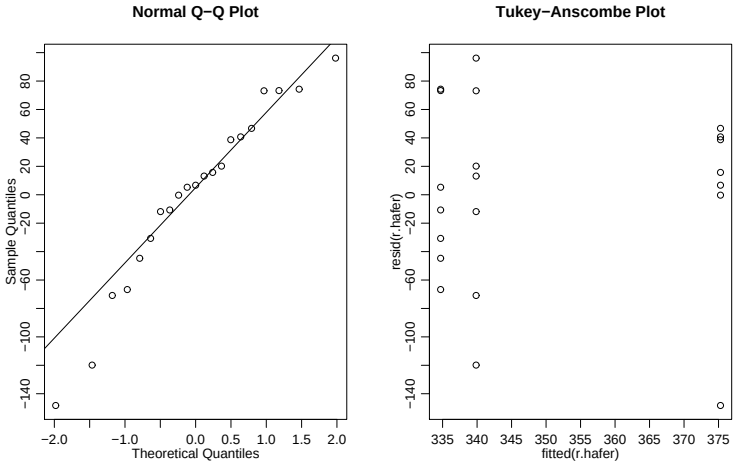
Der P-Wert zum F-Test beträgt 0.481. Es sind also **keine signifikanten** Unterschiede zwischen den 3 Gruppen vorhanden. Die Nullhypothese $H_0 : \alpha_1 = \alpha_2 = \alpha_3$ im Modell $y_{ij} = \mu+\alpha_i+e_{ij}$ kann somit auf dem 5%-Niveau nicht verworfen werden.

Bemerkung: Auch ohne die auffällige Beobachtung (Replikat 7 in Gruppe 2) ist der F-Test nicht signifikant (der P-Wert beträgt 0.106).

Der Aufwand des Beizens scheint sich nicht zu lohnen.

c) Um die Fragen zu beantworten, betrachten wir den Q-Q Plot und den Tukey-Anscombe Plot:
R-Befehl:

```
> op <- par(mfrow=c(1,2))
> plot(resid(r.hafer) ~ fitted(r.hafer),
+      main="Tukey-Ansombe-Plot",ylab="Residuen")
> abline(h=0,lty=2)
> qqnorm(resid(r.hafer), ylab="Residuen")
> qqline(resid(r.hafer))
> par(op)
```



Wie wir schon in a) bemerkt haben, fällt ein Ausreisser (Replikat 7 der Gruppe 2) auf. Die Verteilung der Residuen scheint leicht schief zu sein, allerdings nicht dramatisch. Wenn man sich den Ausreisser wegdenkt ist aber die Varianz der Fehler nicht (mehr) konstant!

d) Das Modell $y_{ij} = \mu+\alpha_i+e_{ij}$ vermag nicht die gesamte Information aus den Daten zu schöpfen. Die Abbildung in a) zeigt, dass die Replikate nicht unabhängig sind. Beobachtungen mit kleiner Replikatsnummer haben tendenziell einen grösseren Ertrag als Beobachtungen mit grosser Replikatsnummer.

Wir können eine Einweg-Varianzanalyse mit dem Faktor `REP` rechnen:

R-Befehl:

```
> d.hafer$REP <- factor(d.hafer$REP)
> r.hafer2 <- aov(YIELD ~ REP, data = d.hafer)
> summary(r.hafer2)
            Df Sum Sq Mean Sq F value  Pr(>F)
REP          6  55616    9269  4.0605 0.01443 *
Residuals   14  31959    2283
```

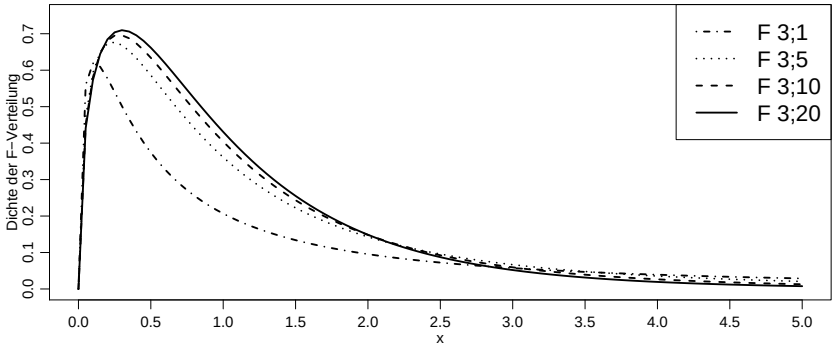
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In diesem Modell gibt es einen signifikanten Einfluss des Faktors `REP`. Vermutlich wurden die Töpfe mit derselben Replikatsnummer gruppiert und an den gleichen Ort des Gewächshauses gestellt. Deshalb könnten die Wachstumsbedingungen für die sieben Replikatengruppen unterschiedlich sein.
Unser Dozent vermutet, dass der 6. Wert in der Gruppe 2 statt 422 eher 242 heissen sollte.

3. **a)** F-Verteilungen bei wachsendem Nenner-Freiheitsgrad:

```
f.plotF <- function(df1, df2, xlim=c(0,5), ylim=c(0,.75), add=TRUE, lty=1,
                    lwd=2, col=1, main=NULL) {
  curve(df(x, df1=df1, df2=df2), xlim=xlim, ylim=ylim, add=add, lty=lty,
        lwd=lwd, col=col, main=main, ylab="Dichte der F-Verteilung")
}

f.plotF(df1=3, df2=1, lty=4, add=F)
f.plotF(df1=3, df2=5, lty=3)
f.plotF(df1=3, df2=10, lty=2)
f.plotF(df1=3, df2=20, lty=1)
legend("topright", paste("F", paste(rep(3,4), c(1,5,10,20), sep=";")),
       lty=4:1, lwd=2, col=rep(1,4), cex=1.5)
```

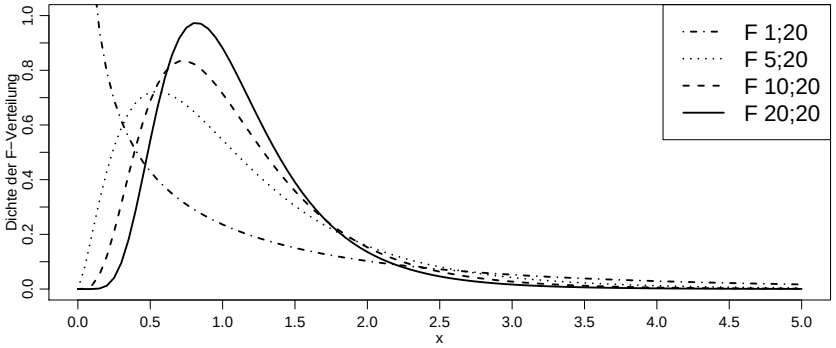


b) 95%-Quantile:

```
> dflist <- list(c(3,1), c(3,5), c(3,10), c(3,20))
> sapply(dflist, function(x) qf(0.95, df1=x[[1]], df2=x[[2]]))
[1] 215.707345   5.409451   3.708265   3.098391
## oder
> qf(0.95,3,1)
[1] 215.7073
> qf(0.95,3,5) # etc.
[1] 5.409451
```

c) F-Verteilungen bei wachsendem Zähler-Freiheitsgrad:


```
f.plotF(df1=1, df2=20, lty=4, ylim=c(0,1), add=F)
f.plotF(df1=5, df2=20, lty=3)
f.plotF(df1=10, df2=20, lty=2)
f.plotF(df1=20, df2=20, lty=1)
legend("topright", paste("F",paste(c(1,5,10,20), rep(20,4), sep=";")),
       lty=4:1, lwd=2, col=rep(1,4), cex=1.5)
```



d) 95%-Quantile:

```
> dflist <- list(c(1,20), c(5,20), c(10,20), c(20,20))
> sapply(dflist, function(x) qf(0.95, df1=x[[1]], df2=x[[2]]))
[1] 4.351244 2.710890 2.347878 2.124155
## oder
> qf(0.95,1,20)
[1] 4.351244
> qf(0.95,5,20) # etc.
[1] 2.71089
```

e) p-Werte für F=2.37:

```
> dflist <- list(c(1,20), c(5,20), c(10,20), c(20,20))
> sapply(dflist, function(x) pf(2.37, df1=x[[1]], df2=x[[2]], lower.tail=F))
[1] 0.13935977 0.07643008 0.04819802 0.03022462
## oder
> pf(2.37, 1, 20, lower.tail=F)
[1] 0.1393598
> pf(2.37, 5, 20, lower.tail=F) ## etc.
[1] 0.07643008
```

f) Falls in einem ersten ANOVA-Modell eine Interaktion nicht signifikant ist und diese aus dem Modell entfernt wird, so wächst für das neue Modell der Freiheitsgrad des Fehler um eben diese Freiheitsgrade der Interaktion an. Das mittlere Quadrat ändert sich dadurch nicht stark, da der Beitrag der Interaktion “klein” ist (sonst wäre die Interaktion ja signifikant!). Die Macht des F-Tests wächst aber an im Vergleich zum vollen Modell mit Interaktion.

g) Die Anzahl Freiheitsgrade des Zählers wird durch die Anzahl Stufen jedes Faktors bestimmt.