Varianzanalyse & Versuchsplanung — Serie 1

1. Der Datensatz `stream ("http://stat.ethz.ch/Teaching/Datasets/WBL/stream.dat")` enthält den Zinkgehalt eingeteilt in vier Klassen (Variable `ZINC`) von verschiedenen Flüssen (Variable `STREAM`) und die Variable `DIVERSITY`, welche die Artenvielfalt an der entsprechenden Flussstelle beschreibt. Zusätzlich gibt es die Variable `ZNGROUP`, welche die verschiedenen Zinkgruppen numerisch codiert. Wir wollen untersuchen, ob Zink einen signifikanten Zusammenhang mit der Artenvielfalt aufweist.

a) Wir sollten die Daten zuerst anschauen. Betrachten Sie den Dataframe. Mit dem R-Befehl `str("Dataframe")` sehen Sie, dass die Variable `ZINC` bereits als Faktor identifiziert wurde, jedoch nicht die Variable `ZNGROUP`, die von R als kontinuierliche Variable aufgefasst wird. Korrigieren Sie dies! Zeichnen Sie ein Streudiagramm der Diversität gegen `ZNGROUP`. Gibt es extreme Beobachtungen?
R-Hinweise:
- Um eine kontinuierliche Variable in einen Faktor umzuwandeln, benutzen Sie `d.stream[,"ZNGROUP"] <- as.factor(d.stream[,"ZNGROUP"])`
- Mit `summary("Dataframe")` werden zu jeder Variable einige Informationen ausgegeben.
- Streudiagramme (scattorplot) kann man mit der aus der Regression bekannten Formelschreibweise zeichnen:
  `plot(DIVERSITY ~ ZNGROUP, data = d.stream)`
- R zeichnet hier automatisch einen Boxplot, da auf der x-Achse ein Faktor aufgetragen wird. Wenn man trotzdem ein Streudiagramm sehen möchte, muss man den Faktor als kontinuierliche Variable der Plot-Funktion übergeben:
  `plot(DIVERSITY ~ as.numeric(ZNGROUP), data = d.stream)`

b) Berechnen Sie eine einfache Varianzanalyse. Testen Sie anhand des Outputs, ob sich die Diversitäten bei verschiedenen Zinkgehalten unterscheiden.
R-Hinweise:
- Um das Modell $y_{ij} = \mu_i + e_{ij}$ zu fitten, benützen wir die R-Funktion `aov`:
  `aov("Zielvariable" ~ "erklärende Variable", data = "Dataframe")`.
- Mit `summary("angepasstes Modell")` wird eine ANOVA-Tabelle ausgegeben.

c) Überprüfen Sie die Modellannahmen mit einer Residuenanalyse (Tukey-Anscombe Plot, Normal Plot).
R-Hinweise:
- Wie bei der Regression kann auch hier mit `resid("angepasstes Modell")` auf die Residuen und mit `fitted...` auf die angepassten Werte zugegriffen werden.
- Mit `qqnorm("Residuen")` wird ein qq-Plot der Residuen gezeichnet. Eine passende Gerade kann mit `qqline("Residuen")` eingefügt werden.

d) Geben Sie die Schätzwerte für die Parameter an!
R-Hinweis: Die geschätzten Parameterwerte erhalten Sie mit `summary.lm("angepasstes Modell")`. Dabei wird der auf 0 gesetzte Parameter nicht angezeigt. Alle Parameterwerte erhält man mit `dummy.coef("angepasstes Modell")`.

(Quelle: Quinn, G. and Keough, M., *Experimental Design and Data Analysis for Biologists, Cambridge, 2002, p. 173 f.*)

2. Die folgenden Daten `hafer` (`"../WBL/hafer.dat"`) stammen aus einem Experiment mit behandeltem und unbehandeltem Hafer-Saatgut. Das Saatgut wurde in drei Gruppen aufgeteilt. Die Gruppen 1 und 2 wurden separat mit demselben Wirkstoff gebeizt. Ein Teil blieb unbehandelt (Gruppe Check). Anschliessend wurden die Samen in je 7 Töpfen pro Gruppe zum Keimen gebracht. Am Ende des Versuchs wurde der Ertrag pro Topf (in Gramm) gemessen. Wie die Töpfe im Gewächshaus angeordnet waren, lässt sich leider nicht mehr eruieren.

	Treatment		
Replicate	Group 1	Group 2	Check
1	360	391	408
2	436	382	409
3	413	414	340
4	353	416	324
5	328	375	304
6	269	422	268
7	220	227	290

Der Datensatz `hafer` enthält die Variablen `REP` (Replicate), `YIELD` (Ertrag), `GROUP` (Code 1 bis 3) und `TREATM` (Gruppenbezeichnung, Treatment).

a) Stellen Sie die Daten zuerst graphisch dar. Kommentar!

b) Vergleichen Sie die drei Gruppen mit einer Einweg-Varianzanalyse (mit `GROUP` als Faktor). Gibt es signifikante Unterschiede auf dem 5%-Niveau?
Lohnt sich das Beizen des Saatgutes?

c) Überprüfen Sie die Modellannahmen mit einer Residuenanalyse (Tukey-Ansombe Plot, Normal Plot). Ist die Varianz der Fehler konstant? Sind die Fehler normalverteilt?

d) Vermag das statistische Modell in c) $y_{ij} = \mu + \alpha_i + e_{ij}$ die gesamte Information aus den Daten zu schöpfen? Betrachten Sie die Replikate als Stufen eines Faktors und führen Sie eine entsprechende Varianzanalyse durch.
Liegt evtl. noch ein Schreibfehler vor?

(Quelle: Ostle, B. and R.W Mensing (1975), *Statistics in Research*, Iowa State University Press, Ames., S. 417)

3. (fakultativ) Für die Signifikanztests in der Varianzanalyse bedienen wir uns der F-Verteilungen (nach Ronald Aylmer **F**isher (1890-1962)). Die F-Verteilung ist die Verteilung eines Quotienten. In dessen Zähler steht das mittlere Quadrat (**M**ean **S**quare) von ν_1 unabhängigen standardnormalverteilten Zufallsvariablen, im Nenner dasjenige von ν_2 standardnormalverteilten unabhängigen Zufallsvariablen. Der Parameter ν_1 heisst Anzahl "Freiheitsgrade des Zählers", ν_2 ist die Anzahl "Freiheitsgrade des Nenners".

Um ein Gefühl für die beiden Parameter der F-Verteilung zu bekommen, werden wir einige (der unendlich vielen!) F-Verteilungen untersuchen.

R-Hinweise finden Sie nach den Aufgaben!

a) Stellen Sie grafisch die Wahrscheinlichkeitsdichten der folgenden F-Verteilungen dar:
$F_{3,\mathbf{1}}$ $F_{3,\mathbf{5}}$ $F_{3,\mathbf{10}}$ $F_{3,\mathbf{20}}$

Wie verändert sich der Verlauf einer F-Verteilung, wenn der Freiheitsgrad des Nenners vergrössert wird?

b) Berechnen Sie für jede der F-Verteilungen aus a) das 95%-Quantil! Was schliessen Sie aus dem Resultat für die Erlangung einer Signifikanz?

c) Untersuchen Sie auch die folgenden F-Verteilungen:
$F_{\mathbf{1},20}$ $F_{\mathbf{5},20}$ $F_{\mathbf{10},20}$ $F_{\mathbf{20},20}$

Wie verändert sich die Form der F-Verteilung, wenn der Freiheitsgrad des Zählers vergrössert wird?

d) Berechnen Sie auch für die F-Verteilungen aus c) jeweils das 95%-Quantil.

e) Geben Sie für jede der Verteilungen in c) die p-Werte für einen F-Wert von 2.37 an.

f) Überlegen Sie sich eine Situation, in der sich die Anzahl Freiheitsgrade im Nenner ändert, während der F-Wert sich nur wenig ändert. (**Tipp:** Skript S. 23)

g) Wodurch kann die Anzahl Freiheitsgrade des Zählers beeinflusst werden?

R-Hinweise:

- Den Wert der Dichte der F-Verteilung mit 3 und 5 Freiheitsgraden an der Stelle x erhalten Sie in R mit `df(x, df1=3, df2=5)`.
  Analog zu `dnorm()`, `pnorm()`, `qnorm()`, etc. erhalten Sie mit `pf()` die kumulative Verteilungsfunktion und mit `qf()` die Quantile der entsprechenden F-Verteilung.
  Mit `pf(`q`, ..., lower.tail=FALSE)` erhält man das Integral unter der Dichtekurve rechts vom Wert q, also den p-Wert.
- Einen Funktionsgraphen zeichnen Sie mit `curve(`*Funktion von* x`, xlim= , ylim= )`. Dabei muss x die "Laufvariable" sein.
- Um verschiedene Funktionen zu vergleichen, benützen Sie `par(mfrow=c( , ))` oder `curve(..., add=TRUE, lty= , lwd=, col=)`.
- `sapply(`*Liste, Funktion*`)` wendet eine Funktion auf jedes Listenelement an.