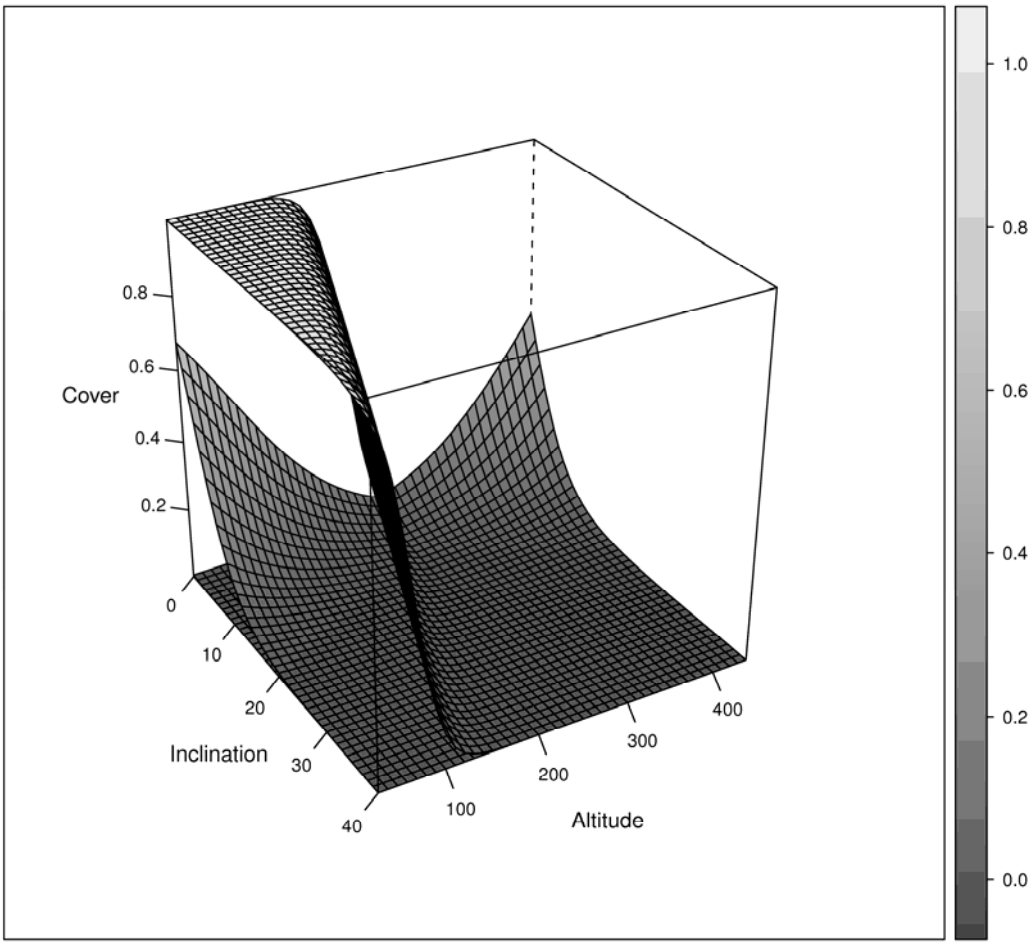
Statistische Auswertung ökologischer Datensätze
-
Kursskript



Michael Rudner & Martin Schaefer
Fakultät für Biologie

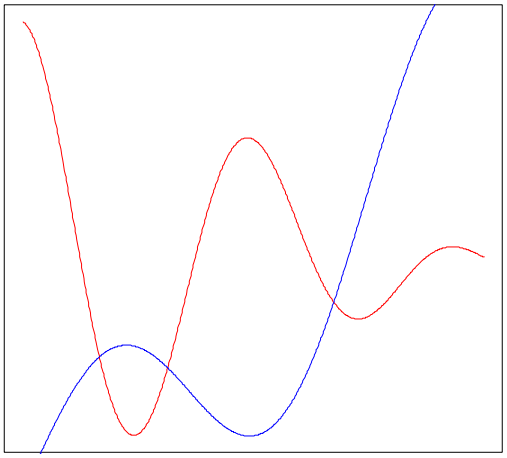
November 2010
Albert-Ludwigs-Universität Freiburg

INHALT

EINLEITUNG

Ziel des Kurses

Kenntnisse zur Anwendung statistischer Verfahren bis hin zu linearer Regression und Varianzanalyse, Fertigkeiten in der Umsetzung von Analysen im Statistikprogramm R



Ziel des Skripts

Kurze Übersicht über wesentliche Themen und Methoden der statistischen Analyse von ökologischen Datensätzen, Hinweise zur Umsetzung in R, komplementär zu den Übersichten zu R-Befehlen zu den einzelnen Kurstagen

Aufbau des Skripts

Das Skript handelt eine Auswahl wesentlicher Aspekte und Verfahren der statistischen Datenanalyse ab. Es stützt sich im Wesentlichen auf die im Literaturverzeichnis genannten Lehrbücher und Skripten. Jedes Thema ist auf einer Seite steckbriefartig zusammengefasst. Der Aufbau folgt soweit möglich einer einheitlichen Struktur:
Gebräuchliche Bezeichnung | Typische Grafik | Was passiert? | Was sagt das Ergebnis aus? | Eignung | Umsetzung in R | Wesentliche Angaben

Weitere Lehrmaterialien

Zur statistischen Datenanalyse in R gibt es die Folien zur Vorlesung sowie Übersichten zu den geeigneten R-Anweisungen. Dazu gibt es eine Reihe von Übungsblättern und die passenden R-Skripten. Zum Einstieg in die Benutzung von R existieren zwei E-Learning-Module, die in WebKit an der Universität Freiburg umgesetzt wurden. Sie werden Studierenden der Biologie über CampusOnline zur Verfügung gestellt.

Eignung des Skripts

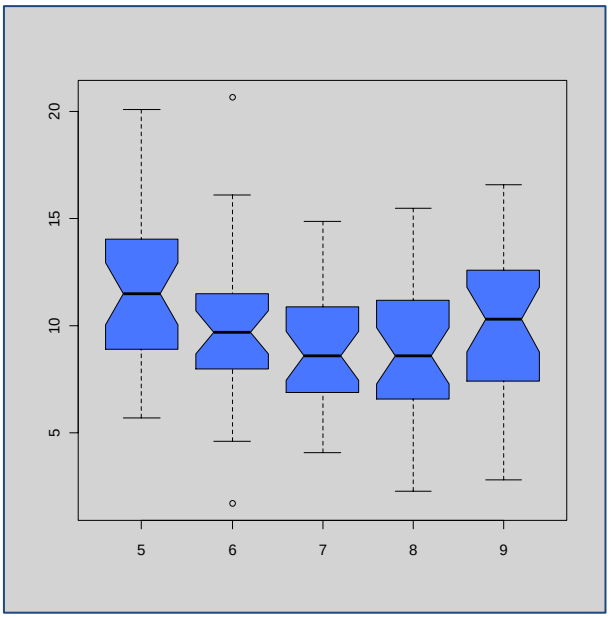
Das Skript ist als unterstützendes Material für den Kurs „Statistische Auswertung ökologischer Datensätze“ gedacht. Auch soll es eine rasche Orientierung bieten, um bei vorhandenen Kenntnissen den Einstieg ins Thema zu erleichtern, wenn Auswertungen eigener Daten anstehen. Für eingehendere Studien verweisen wir auf die Lehrbücher, die in den Literaturhinweisen genannt sind.

Umsetzung in R

Es wird der Standardumfang von R genutzt. Nur an wenigen Stellen sind gesonderte Paktete erforderlich.

STATISTIK

Ziel



Bei wissenschaftlichen Arbeiten werden statistische Verfahren meist angewendet, um Hypothesen zu überprüfen oder Zusammenhänge zwischen steuernden Variablen und einer Ergebnisvariablen quantitativ abzubilden. Dies kann bis zu Modellen führen, die Vorhersagen erlauben. Diese Verfahren zählen zur *analytischen* oder *schließenden Statistik*. Die *deskriptive Statistik* beschäftigt sich mit der Zusammenfassung und Darstellung der Daten.

Was passiert?

In der Regel wird die Verteilung der Daten untersucht und dargestellt. Es wird überprüft, ob zwischen verschiedenen Variablen Zusammenhänge bestehen. Sofern Abhängigkeiten vermutet werden, können die Wirkgrößen beziffert werden. Dem Versuchsdesign muss dabei immer Rechnung getragen werden. In der Regel wird mit statistischen Tests geprüft, ob sich aus dem Datensatz Rückschlüsse ziehen lassen, die übertragen werden können. So wird z.B. überprüft, ob zwei Mittelwerte, die sich vom Betrag her unterscheiden, auch echte Unterschiede anzeigen und nicht nur zufällig differieren.

Wann sind welche Verfahren anzuwenden?

Einfachere Anwendungen sind Darstellungen von Verteilungen und Mittelwertvergleiche. Bei linearen Abhängigkeiten kann die Regressionsanalyse angewendet werden. Versuche mit mehreren Varianten können mit Hilfe der Varianzanalyse ausgewertet werden.
Es ist generell zu unterscheiden zwischen parametrischen Verfahren, die eine bestimmte Verteilung der Daten voraussetzen, und den nicht-parametrischen Verfahren, die sich auf ein breiteres Spektrum von Daten anwenden lassen, meist aber weniger trennscharf sind.

Wesentliche Richtungen

Grafische Analyse	Prüfen des Datensatzes auf Mess- und Eingabefehler
Test zu Verteilungen	Ermitteln der Verteilung der Daten
Korrelationsanalyse	Zusammenhänge zwischen Datenreihen oder Variablen ohne dass ein direkter Wirkungszusammenhang vermutet wird
Regressionsanalyse	Quantitative Analyse von Wirkungszusammenhängen zwischen Variablen
Varianzanalyse	Quantitative Analyse von Experimenten mit Varianten

ABLAUF EINER ANALYSE

Ziel

- Prüfen der Verteilung von Daten
- Quantifizieren der Fehler
- Testen von Hypothesen zur statistischen Absicherung von Untersuchungsergebnissen
- Quantifizieren von Wirkgrößen

Typen

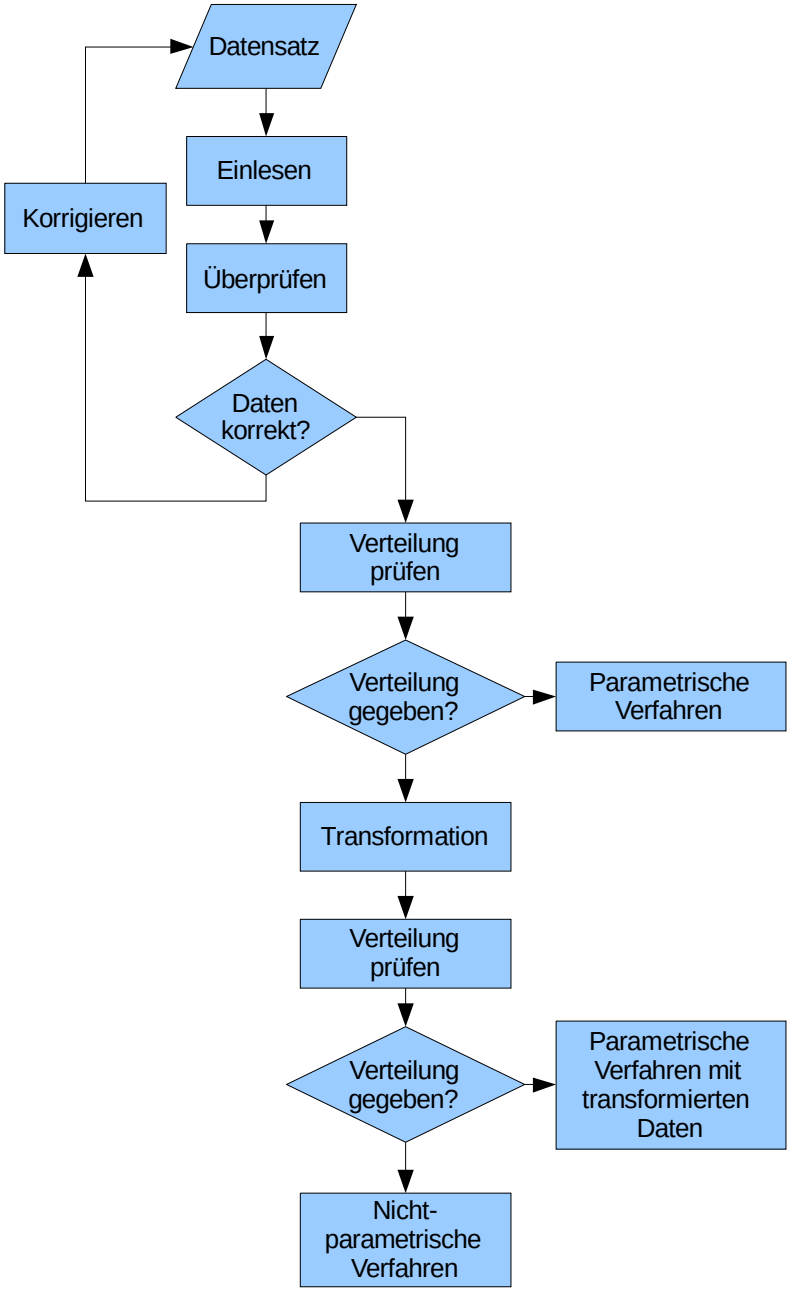
- Grafische Überprüfung
- Prüfen von Korrelationen
- Vergleichende Tests
- Statistische Modelle

Was passiert?

- Prüfen von Voraussetzungen
- Sofern möglich, Anwenden parametrischer Verfahren, sonst nicht-parametrischer Verfahren
- Ökologische Interpretation des Ergebnisses

Eignung / Anforderung an die Daten

- Ausreichender Stichprobenumfang
- Daten müssen repräsentativ erhoben worden sein.
- Daten müssen voneinander unabhängig sein (Autokorrelation muss explizit berücksichtigt werden)



Dingend zu beachten!

Das beste Testergebnis taugt nur so viel, wie seine sinnvolle ökologische Interpretation!

EINLESEN VON DATEN

Was passiert?

Daten, die erfasst wurden, werden zur weiteren Analyse in das Statistikprogramm geladen.

```
"";"S.Length";"S.Width";"P.Length";"P.Width";"Species"
"1";5,1;3,5;1,4;0,2;"setosa"
"2";4,9;3;1,4;0,2;"setosa"
"3";4,7;3,2;1,3;0,2;"setosa"
"4";4,6;3,1;1,5;0,2;"setosa"
"5";5;3,6;1,4;0,2;"setosa"
"6";5,4;3,9;1,7;0,4;"setosa"
"7";4,6;3,4;1,4;0,3;"setosa"
"8";5;3,4;1,5;0,2;"setosa"
"9";4,4;2,9;1,4;0,2;"setosa"
"10";4,9;3,1;1,5;0,1;"setosa"
"11";5,4;3,7;1,5;0,2;"setosa"
"12";4,8;3,4;1,6;0,2;"setosa"
"13";4,8;3;1,4;0,1;"versicolor"
"14";4,3;3;1,1;0,1;"versicolor"
"15";5,8;4;1,2;0,2;"versicolor"
"16";5,7;4,4;1,5;0,4;"versicolor"
"17";5,4;3,9;1,3;0,4;"versicolor"
"18";5,1;3,5;1,4;0,3;"versicolor"
"19";5,7;3,8;1,7;0,3;"versicolor"
"20";5,1;3,8;1,5;0,3;"versicolor"
"21";5,4;3,4;1,7;0,2;"versicolor"
"22";5,1;3,7;1,5;0,4;"versicolor"
"23";4,6;3,6;1;0,2;"versicolor"
```

Anforderung an die Daten

Statistikprogramme erwarten üblicherweise, dass die Daten in einer Tabelle nach Variablen in Spalten organisiert sind. D.h. es gibt zu jeder Messgröße eine Spalte und eine andere Spalte zu den Varianten des Experiments, in der festgehalten wird, ob die Messwerte in der Zeile zur Kontrolle oder einer anderen Variante gehören.
Die Daten sollten in einer Trennzeichen-getrennten Textdatei vorliegen (csv-Format), die sich aus jedem Tabellenkalkulationsprogramm exportieren lässt. Sie sollten wissen, welches Zeichen zur Trennung der Datensätze verwendet wird und welches Zeichen als Dezimaltrennzeichen verwendet wird. Die Einlesefunktion muss dazu passen oder entsprechend spezifiziert werden.

Typen

Die Funktion `read.csv("Datei.txt")` hat mehrere Varianten, die eine unterschiedliche Voreinstellung bezüglich der Trennzeichen für die Spalten und der Dezimaltrennzeichen haben.
Diese Eigenschaften können über die Angabe der Parameter `sep` und `dec` näher bestimmt werden.

Funktion	Dezimal-trenn-zeichen	Trennzeichen für Spalten
`read.table`	`dec="."`	`sep=""`
`read.csv`	`dec="."`	`sep=","`
`read.csv2`	`dec=","`	`sep=";"`
`read.delim`	`dec="."`	`sep="\t"`
`read.delim2`	`dec=","`	`sep="\t"`

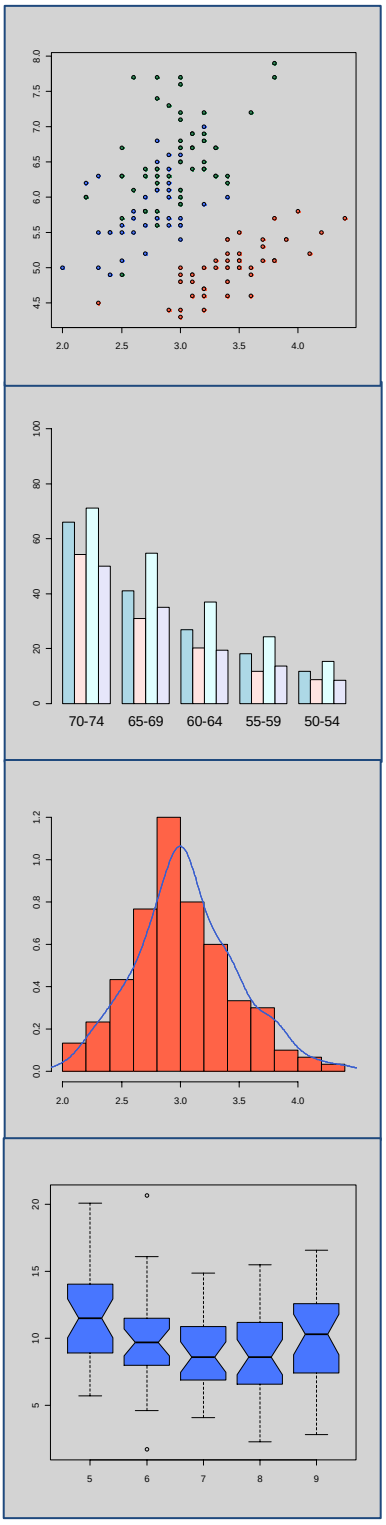
Dringend zu beachten!

R bezieht sich auf das aktuelle Arbeitsverzeichnis. Wenn Ihre Datei dort zu finden ist, müssen Sie den Verzeichnis-Pfad in der Einlesefunktion nicht angeben. Sie müssen in jedem Fall die Dateiendung mit angeben (= die 3 Zeichen nach dem Punkt, also .txt oder .csv). Links geneigte Schrägstriche, wie sie in Windows-Pfaden verwendet werden, müssen immer doppelt angegeben werden.
Das Ergebnis der Einlesefunktion muss einem Objekt zugewiesen werden, da es sonst nur in die Konsole (=Ausgabefenster in R) geschrieben wird und nicht weiter zur Verfügung steht.

Befehle in R

Arbeitsverzeichnis setzen: `setwd("Z:\\Kurs1\\Daten");` Prüfen mit `getwd()`.
Einlesen und Zuweisen mit der passenden Funktion,
z.B.: `dat <- read.csv2("Datei.csv")`
Prüfen mit `str(dat)` oder `summary(dat)` oder `fix(dat)` bzw. `edit(dat)`

GRAFIK



Ziel

Veranschaulichen von Datensätzen und von Zusammenhängen

Typen

Streudiagramm (Streudiagramm-Matrix), Liniendiagramm, Säulendiagramm, Histogramm, Boxplot,

Was passiert?

R wählt abhängig von den an die Funktion `plot()` übergebenen Argumenten, die passende Darstellung aus. Falls eine andere Darstellung gewünscht wird, muss diese eigens spezifiziert werden. Es wird ein Grafikfenster geöffnet, falls nötig. Die Datenpunkte werden aufgetragen. Entsprechend der übergebenen Daten wird der Wertebereich der beiden Achsen festgelegt. Beim späteren Einfügen von Grafikelementen geben die Achsenwerte den Bezugsrahmen, d.h. es kann nur dargestellt werden, was innerhalb der Grenzen des Koordinatensystems der Achsen liegt.

Eignung / Anforderung an die Daten

Falls mehrere Variablen übergeben werden, um z.B. ein Streudiagramm zu zeichnen, müssen beide Variablen gleich viele Einträge aufweisen. (Die Vektoren müssen gleich lang sein.) Die Daten müssen numerisch sein, allenfalls für die Unterscheidung von Faktoren können nicht-numerische Daten verwendet werden.

Befehle in R

Der Befehl `plot()` ist sehr vielfältig und abhängig vom übergebenen Datentyp sehr leistungsfähig. Im Plotbefehl können Sie einige Parameter für die Darstellung festlegen. Eine weitere Funktion zur Festlegung von Grafikparametern ist `par()`. Die damit getroffenen Festlegungen gelten bis zur nächsten Änderung derselben.

In eine bestehende Grafik können mit den Funktionen `points()`, `lines()`und `text()` weitere Punkte, Linien und Zeichenfolgen eingefügt werden.

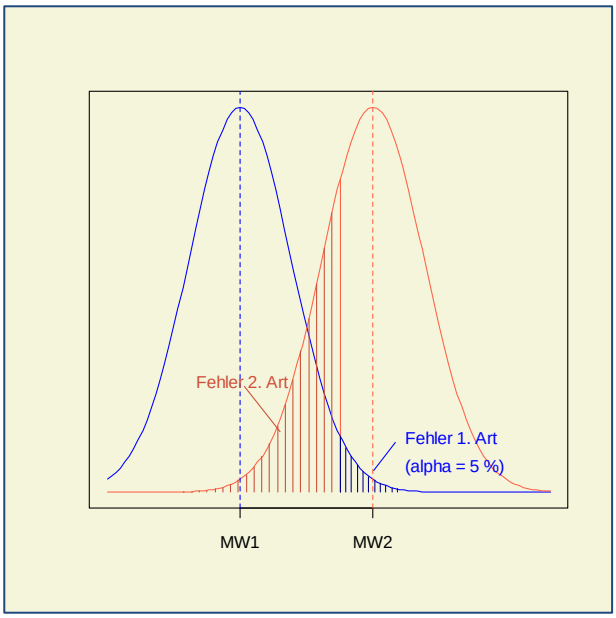
Beispiel:
```
x <- seq(0,3.14,0.1); y <- sin(x) ;
plot(x,y,type="l") ; points(x,y,pch=20)
```

Parameter `type` : `p`: Punkte, `l`: Linien, `b`: beides, `h`: vertikale Linien
`o`: überlagerte Punkte und Linien, `s`: Stufendarstellung
`n`: nichts (Zeichnet eine leere Grafik, legt aber den Wertebereich und die Achsenbeschriftung fest)

STATISTISCHE TESTS

Ziel

Statistische Tests werden eingesetzt, um zu prüfen, wie groß die Wahrscheinlichkeit ist, dass ein bestimmtes Ergebnis zufällig entstanden ist – unter der Prämisse der Nullhypothese. Dies wird verwendet, um einen sog. Signifikanten Zusammenhang oder Unterschied nachzuweisen.



Typen

Parametrische Tests – diese beruhen auf bestimmten Annahmen, z.B. der Normalverteilung der Daten
Nicht-parametrische Tests sind nicht an bestimmte Verteilungen gebunden.

Was passiert?

Es wird eine Kenngröße errechnet, die sogenannte Teststatistik - z.B. die t-Statistik. Dann das Quantil der Verteilung der Teststatistik an der Signifikanzschwelle ermittelt - z.B. der Wert der t-Statistik, die noch mit 2,5% Wahrscheinlichkeit auftritt. Ist die Teststatistik größer als der Vergleichswert an der Signifikanzschwelle, wird die Nullhypothese verworfen – man spricht von einem signifikanten Ergebnis (z.B. signifikant auf dem 5%-Niveau).
Meist wird direkt die Wahrscheinlichkeit für die Teststatistik ermittelt.

Eignung / Anforderung an die Daten

Parametrische Tests setzen oft voraus, dass die Daten einer bestimmten Verteilung folgen.
Bei Test auf Verteilungen ist zu bedenken, dass die abhängigen Variablen getrennt nach Gruppen (Behandlungen, Arten o.ä.) getestet werden sollen. Unter dem Einfluss eines zweistufigen Faktors wäre ja eine zweigipflige Verteilung zu erwarten (≠ Normalverteilung).

Dingend zu beachten!

Bei beidseitigen Tests ist das Vergleichsniveau nur halb so hoch (im Beispiel daher 2,5%)
Der Fehler, den man begeht, wenn man die Nullhypothese ablehnt obwohl sie tatsächlich stimmt, ist der Fehler 1. Art. Dieser wird über das Signifikanzniveau α bzw. den p-Wert quantifiziert. Der Fehler 2. Art, die Nullhypothese nicht abzulehnen obwohl sie falsch ist, kann nicht über αoder den p-Wert beziffert werden.
Wenn man die Nullhyopthese nicht verwirft ist damit noch nicht bewiesen, dass sie zutrifft.

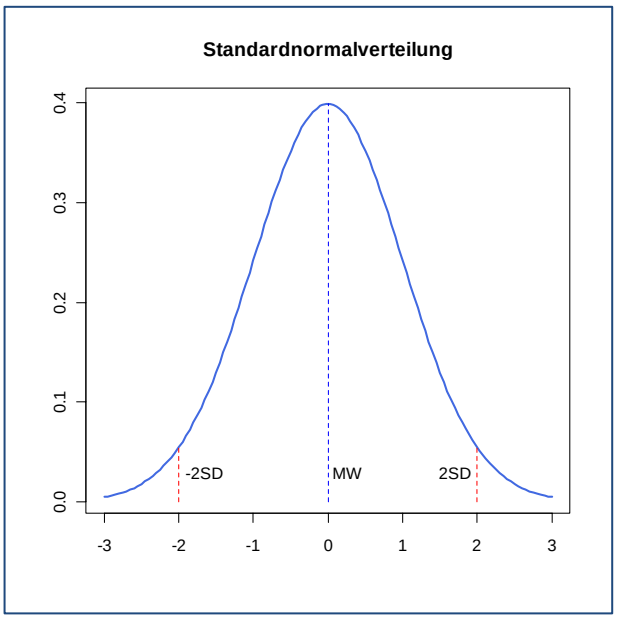
Befehle in R

Es gibt verschiedene Testfunktionen für besondere Zwecke, z.B. `shapiro.test()`, `ks,test()`, `t.test()`, `var.test()`.
Zusätzlich gibt es Funktionen, um die Quantile, die Wahrscheinlichkeitsdichte der Verteilungen, und entsprechende Zufallswerte zu erzeugen, z.B. `qnorm()`, `dnorm()`, `rnorm()`, `qt()`, `dt()`, `rt()` oder `qf()`, `df()`, `rf()`.

NORMALVERTEILUNG

Ziel

Die Normalverteilung beschreibt eine häufige Verteilung von Daten, die dann gegeben ist, wenn Daten aufgrund von zufälligen Einflüssen um einen Mittelwert schwanken, z.B. die Mittelwerte bei wiederholten Probenahmen in einer Population. Da viele statistische Verfahren normalverteilte Daten voraussetzen, nimmt sie einen zentralen Platz in der statistischen Analyse ein. Gegebenenfalls wird auch versucht, über Transformationen die Daten einer Normalverteilung anzunähern.



Typen

Normalverteilung
Standardnormalverteilung (SNV)
Schiefe Verteilungen (linkschief, rechtsschief)

Was passiert?

Eine Normalverteilung ist durch Mittelwert (MW) und Standardabweichung (SD) gekennzeichnet. (SNV: MW = 0, SD = 1). Fast alle Werte liegen im Intervall [MW-3xSD;MW+3xSD].

Eignung / Anforderung an die Daten

Voraussetzung für viele parametrische Verfahren, wie den t-Test (Mittelwertvergleich)
Bei bekannter Verteilung kann die Wahrscheinlichkeit für das Auftreten bestimmter Werte ermittelt werden.

Dringend zu beachten!

Wird die SNV herangezogen, um Wahrscheinlichkeiten für das Auftreten bestimmter Werte oder Wertebereiche zu ermitteln, muss anschließend wieder rücktransformiert werden.

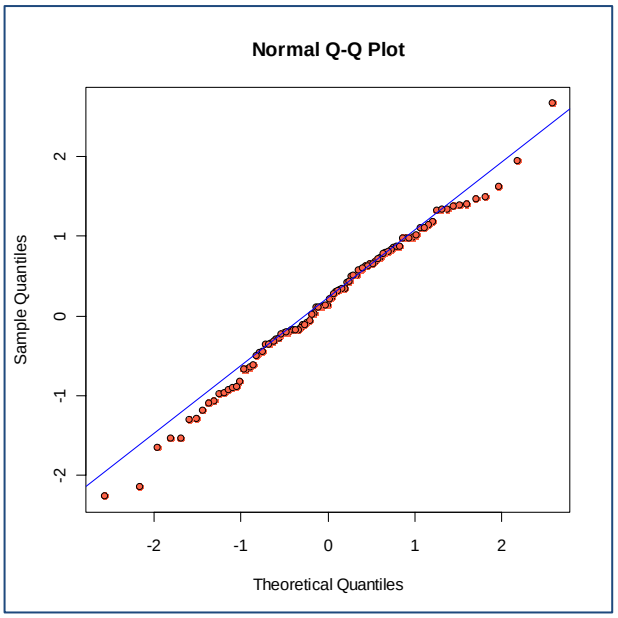
Befehle in R

Die folgenden Funktionen Arbeiten mit der Standardnormalverteilung, außer wenn explizit `mu`=Mittelwert und `sd`=Standardabweichung angegeben werden.
`rnorm()` erzeugt Zufallswerte zu einer Normalverteilung.
`dnorm()` berechnet die Dichte der Verteilung (Glockenkurve).
`pnorm()` ermittelt die Wahrscheinlichkeit für das Auftreten von Werten <x.
`qnorm()` ermittelt das Quantil zur angegebenen Wahrscheinlichkeit.

TEST AUF NORMALVERTEILUNG

Ziel

Es soll geprüft werden, ob die Daten die Anforderungen verschiedener *parametrischer* Verfahren erfüllen, die eine Normalverteilung der Daten voraussetzen.



Was passiert?

Die Daten werden mit einem idealen normal verteilten Datensatz verglichen. Die Abweichungen werden in einer Teststatistik z.B. dem W-Wert des Shapiro-Wilk-Tests zusammengefasst. Abhängig von den Freiheitsgraden wird ermittelt, wie hoch die Wahrscheinlichkeit ist, diesen Wert für die Teststatistik zu erhalten, wenn tatsächlich eine Normalverteilung zu Grunde liegt. Der p-Wert gibt somit an, wie groß die Irrtumswahrscheinlichkeit ist, wenn man die Nullhypothese ablehnt. Die Nullhypothese der Tests ist, dass kein Unterschied zu einer Normalverteilung vorliegt.
Bei kleinem p-Wert (z.B. $p < 0.05$ bei einer 5% -Signifikanzschwelle) würde man die Nullhypothese verwerfen. Die Daten sind damit nicht normal verteilt.
Mit der grafischen Überprüfung mittel eines Quantil-Quantil-Plots (Q-Q-Plot) kann festgestellt werden, inwieweit die Daten normalverteilt sind und an welcher Stelle große Abweichungen bestehen. Dabei werden die Quantile der Variablen gegen die theoretischen Quantile einer Normalverteilung aufgetragen. Eine Linie, die den idealen Verlauf der Werte anzeigt, ist hilfreich für die Interpretation.

Eignung / Anforderung an die Daten

Immer dann durchzuführen, wenn von einer Normalverteilung aufgrund der Natur der Daten ausgegangen werden kann. Bei Zähldaten ist das eher unsinnig - da wird üblicherweise eine Poisson-Verteilung erwartet.

Dringend zu beachten!

Jede Variable muss einzeln getestet werden. Es kann vorkommen, dass eine Variable insgesamt nicht normalverteilt ist, bei Betrachtung für verschiedene Varianten (z.B. Pflanzenarten) aber durchaus.

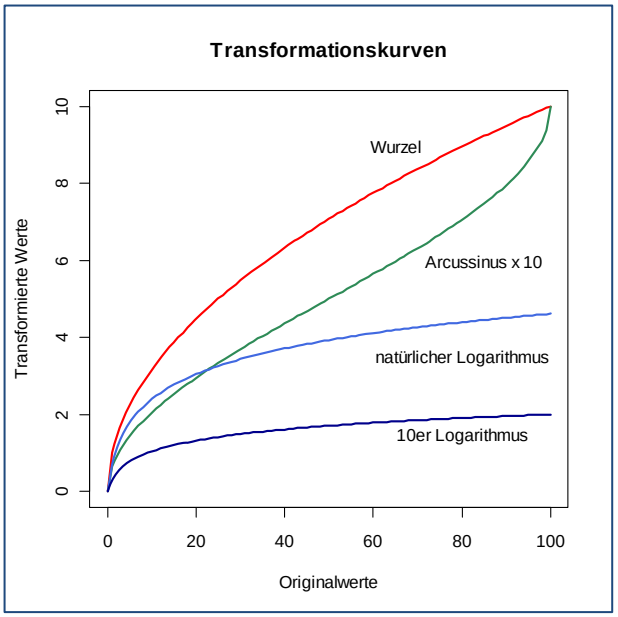
Befehle in R

`shapiro.test()`: Shapiro-Wilk-Test auf Normalverteilung
`ks.test()`: Kolmogorov-Smirnov-Test auf Normalverteilung (kann auch für andere Verteilungen verwendet werden.)
`qqnorm()`: Zeichnet Q-Q-Plot.
`qqline()`: Zeichnet ideale Linie in den Q-Q-Plot

TRANSFORMATION

Ziel

- Sind Daten nicht normalverteilt, kann es in vielen Fällen mit Hilfe einer Transformation gelingen, normalverteilte Daten zu erhalten. Der transformierte Datensatz kann dann mit parametrischen Verfahren analysiert werden.
- Angleichen der Wertebereiche bzw. des Einflusses unterschiedlich skalierter Variablen (z.B. Höhenlage, pH-Wert) (Standardisierung)
- Verändern der Fehlerverteilung bei Prozentskala, um lineare Verfahren anwenden zu können (Arcussinus-Transformation)



Typen

Monotone Transformation wird auf alle Werte unabhängig von den anderen Werten angewendet und hat damit keine Änderung des Rangs zur Folge. **Relativierung** hängt von den anderen Werten ab (z.B. Mittel- oder Maximalwert).

Was passiert?

Häufig verwendet werden Wurzel- und logarithmische Transformation sowie die Box-Cox-Transformation (der allgemeine Fall). Hierbei werden in der Regel die Intervalle im höheren Wertebereich verkürzt und im unteren Wertebereich relativ dazu gestreckt.
Bei einer Standardisierung wird der Mittelwert auf 0 verschoben und durch Division eine Standardabweichung von 1 erreicht. Die Arcussinus-Transformation führt zur Streckung der Intervalle im unteren und oberen Wertebereich.

Eignung / Anforderung an die Daten

Da der Logarithmus von 0 nicht definiert ist, muss vor einer logarithmischen Transformation zu allen Werten eine Konstante addiert werden, die höchstens so groß sein sollte wie der kleinste positive Wert im Datensatz ($x' = \log(x+c)$). Standardisierungen werden vor allem bei Analysen mit mehreren unabhängigen Variablen notwendig.

Dringend zu beachten!

Eine Rücktransformation wird erforderlich, wenn man Aussagen auf den Wertebereich der Originaldaten treffen will, z.B. Prognosen. Bei Relativierungen ist es wichtig, die zur Transformationen verwendeten Parameter zu merken, da sonst keine Rücktransformation möglich wäre. Die Umkehrfunktion zum Logarithmus ist die Exponentialfunktion.

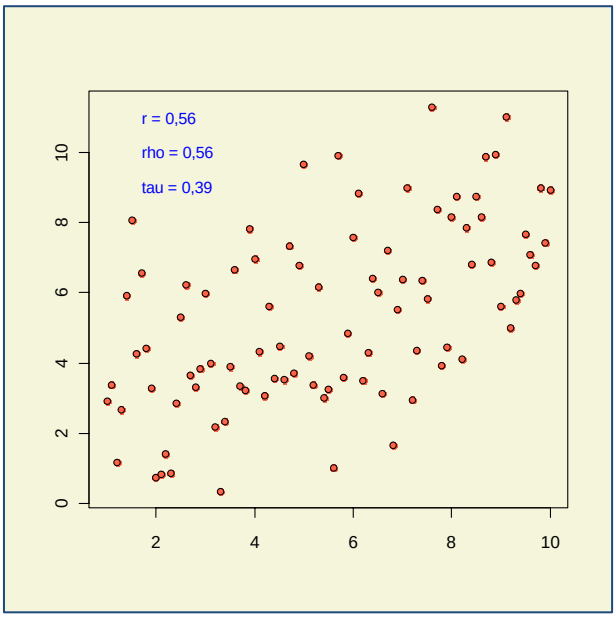
Befehle in R

`sqrt()` Wurzelfunktion, `log()` natürlicher Logarithmus, `exp()` Exponentialfunktion
`D.arcsin <- 2/pi*asin(sqrt(Deckung/100))` Arcussinus-Transformation

KORRELATION

Ziel

Grundfrage: „Gibt es einen (statistischen) Zusammenhang zwischen zwei Variablen?“ Ermitteln wie stark zwei Größen miteinander zusammenhängen, die gemeinsam ansteigen oder sich gegenläufig verhalten. Bei der Analyse der Korrelation wird kein kausaler Zusammenhang impliziert. Dies geschieht oft bei der Interpretation von Korrelationsmaßen.



Typen

Pearsons Korrelationskoeffizient r
Spearmans Rangkorrelation ρ
KendallsKorrelation τ
Jeweils mit Werten von -1 bis +1;
Werte um 0 ➔ keine Korrelation

Was passiert?

Pearsons Korrelationskoeffizient
r(x,y) = Kovarianz(x,y)/Wurzel(Varianz(x)* Varianz(y))

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \cdot \sum (y-\bar{y})^2}}$$

Spearmans Rangkorrelationskoeffizient
wird aus Unterschieden der Ränge ermittelt

$$R = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)}$$

Eignung / Anforderung an die Daten

Korrelation	Koeffizient	Lineare Beziehung	Datenlevel
Pearson Produkt-Moment-K.	r	vorausgesetzt	intervallskaliert
Spearman Rangkorrelation	ρ [rho]	nicht nötig	ordinal, äquidistant
Kendall Korreltaion	τ [tau]	nicht nötig	ordinal

Dingend zu beachten!

Sie müssen überlegen, welche ökologischen Prozesse hinter der Korrelation stehen, diese sozusagen auf Plausibilität prüfen. Damit können Sie dem Problem der Scheinkorrelationen aus dem Weg gehen.
Beachten Sie auch den Bezugsraum, ob das nun Sippen sind oder räumliche Skalen!

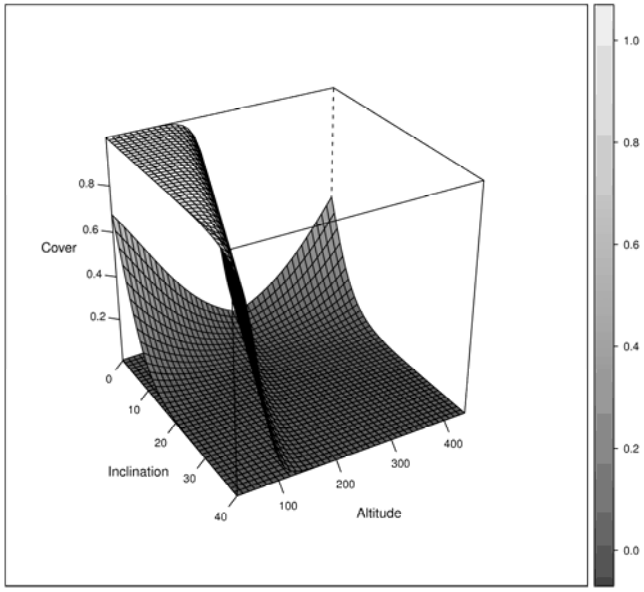
Befehle in R

cor(x,y,method=c(“pearson“,“spearman“,“kendall“)) Korellationskoeffizent
cor.test(x,y,method=c(“pearson“,“spearman“,“kendall“)) K.koeffizient und p-Wert
Die Funktion kann auch auf eine Matrix angewendet werden, nicht jedoch auf eine Tabelle.

STATISTISCHE MODELLE

Ziel

Statistische Modelle beschreiben quantitativ den Zusammenhang zwischen unabhängigen (UV) und abhängigen Variablen (AV). Ziel ist ein minimal adäquates Modell.



Typen

Regressionsmodelle z.B. lineare Regression aber auch komplexere wie die verallgemeinerten linearen Modelle (GLM). Varianzanalysen (kategoriale UV)

Was passiert?

Maximum-Likelihood Schätzung von Regressionskoeffizienten, um ein Modell mit den kleinstmöglichen Residuen zu erreichen (Minimieren der Varianz).

Eignung / Anforderung an die Daten

Die Werte sollen möglichst gleichmäßig über den Wertebereich der UV verteilt sein. Pro geschätztem Modellparameter sollen mehrere am besten 10 Beobachtungen vorliegen.
Residuen sollen homogen über den Wertebereich verteilt sein (Varianzhomogenität)
Residuen sollen normalverteilt sein.

Dingend zu beachten!

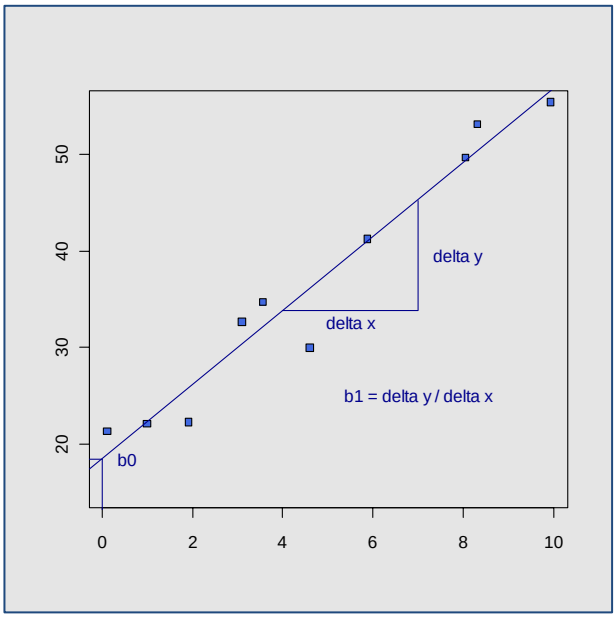
Zunächst klären, welches die AV ist und welches die UV sind. Arbeitshypothesen!!
Hypothesengeleitete Modellbildung, d.h. nur die UV ins Modell aufnehmen, von welchen man sich aufgrund der ökologischen Prozesse einen Effekt erwartet.
UVs die stark positiv oder negativ untereinander korrelieren (Pearson $r > 0.7$) dürfen nicht gemeinsam in ein Modell aufgenommen werden (Problem der Multikollinearität).
Sparsamkeitsprinzip (*parsimony*): So wenig Parameter wie möglich, eher Modelle mit wenigen Annahmen, Vereinfachen bis zum kleinsten adäquaten Modell, einfache Erklärungen komplexen vorziehen.
Ein Modell ohne schlüssige ökologische bzw. biologische Erklärung ist nichts wert!

Befehle in R

Lineare Regressionsmodelle werden mit `lm()` ermittelt. GLMs mit `glm()`. Für Varianzanalysen wird die Funktion `aov()` genutzt. Um im Zuge der Modellbildung zu prüfen, ob die Hinzunahme eines Modellparameters (UV) das Modell verbessert, kann die Funktion `anova(modell1,modell2)` verwendet werden.

REGRESSION

Ziel



Statistische Methode zur Abbildung der Ausprägung einer Variablen in Abhängigkeit von einer oder mehreren anderen Variablen.
Ursachenanalysen & Wirkungsprognosen
Welche Faktoren beeinflussen abhängige Variable (AV)?
Wenn sich UV (unabhängige Variable) ändert, welchen Einfluß hat dies auf AV?

Typen

Eine unabhängige (erklärende) Variable: univariate Regression (Y= f(X))
Mehrere unabhängige Variablen: multiple Regression ($Y= f(X_1, X_2\ldots, X_j,\ldots X_J)$)

Was passiert?

Es wird die Linie gesucht, die die Daten am besten repräsentiert bzw. am wahrscheinlichsten werden lässt. Bei linearer Regression: Minimieren der Summe der Fehlerquadrate.

Eignung / Anforderung an die Daten

Alle Variablen sollten kontinuierlich sein. Ggf. Umformung von kategorialen Variablen in Dummy-Variablen (0,1). Wenn AV nominal (z.B. Anwesenheit: ja, nein), dann logistische Regression verwenden. Lineare Regression setzt einen linearen Zusammenhang zwischen AV & UV voraus ($y = b_0 + b_1 * x$).

Dingend zu beachten!

F-Statistik: Besitzt **Modell** Vorhersagequalität für Grundgesamtheit?
r^2 : Abschätzung des Effekts, wie viel der Streuung durch UV erklärt wird. r^2 wird von Anzahl der UVs beeinflusst, daher wird bei vielen UVs häufig korrigiertes r^2 benutzt. Dies nimmt mit Anzahl der UVs ab, da Erklärungsanteil jeder UV zufällig bedingt sein kann.

$$R^2 = \frac{\sum_{k=1}^{K}(\hat{y}_k - \bar{y})^2}{\sum_{k=1}^{K}(y_k - \bar{y})^2} = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}}$$

r^2 zwischen 0 und 1. In Ökologie in signifikanten Modellen 0.1-0.8
Verteilung der Residuen prüfen!

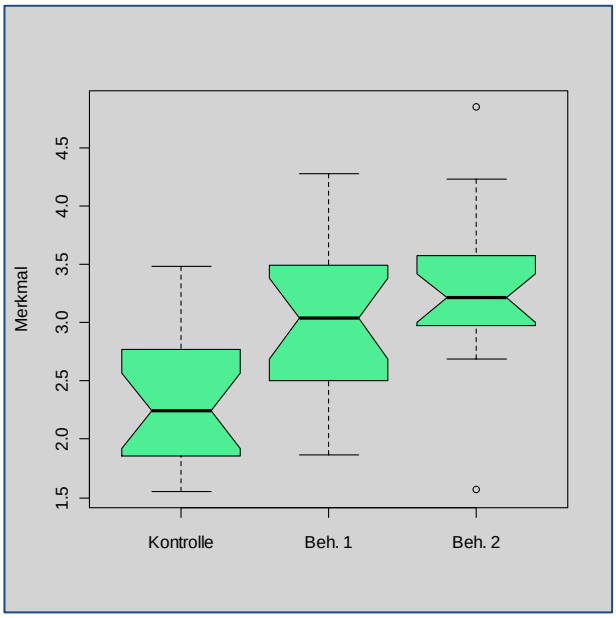
Befehle in R

Für univariate lineare Regression: `lm(y~x, data=Datensatz)` bzw. `lm(y~x-1,…)` wenn die Regressionsgerade durch den Ursprung gehen soll, also kein Achsenabschnitt berechnet wird. Für multivariate lineare Regression: `lm(y~var1+var2+…, data=Datensatz)`. Bei der Berücksichtigung von Interaktionstermen wird das `+` durch ein `*` ersetzt.
Logistische Regression kann mit `glm()`, der Funktion für verallgemeinerte lineare Modelle oder mit der Funktion `lr()` aus dem Paket Design berechnet werden.

VARIANZANALYSE

Ziel

Über den Vergleich der Varianzen der abhängigen Variablen unter dem Einfluss eines Faktors (bei verschiedenen Faktorstufen, z.B. Behandlungen) wird ermittelt, ob es einen Unterschied der Mittelwerte gibt.
Es wird quantifiziert, wie stark sich ein Faktor (eine Behandlung) auf die abhängige Variable auswirkt.



Typen

One-way-ANOVA – ein Faktor
Two-way-ANOVA – zwei Faktoren
Multi-way-ANOVA – > 2 Faktoren

Was passiert?

Die Gesamtvarianz im Datensatz wird mit der Varianz verglichen, die durch die Faktoren erklärt wird. Es wird geprüft, ob die F-Statistik (Quotient der Varianzen) höher ist als ein F-Wert mit 5% Wahrscheinlichkeit bei zutreffender Nullhypothese. Dazu müssen die Freiheitsgrade zu den Varianzen im Zähler und Nenner angegeben werden. Das Ergebnis zeigt also zunächst an, ob sich wenigstens zwei der untersuchten Mittelwerte signifikant unterscheiden. Um herauszufinden, ob es zwischen verschiedenen Behandlungen oder Gruppen von Behandlungen signifikante Unterschiede gibt, muss ein a posteriori Test durchgeführt werden. Hierzu werden Kontraste verwendet, die orthogonal sein müssen, um unzulässige Mehrfachvergleiche zu vermeiden.

Eignung / Anforderung an die Daten

Abhängige Variable (AV) kontinuierlich skaliert, unabhängige Variablen kategorial, nominal oder binär; AV muss bei jeder Faktorstufe normalverteilt sein; Varianzen der AV müssen bei verschiedenen Faktorstufen gleich sein.

Dingend zu beachten!

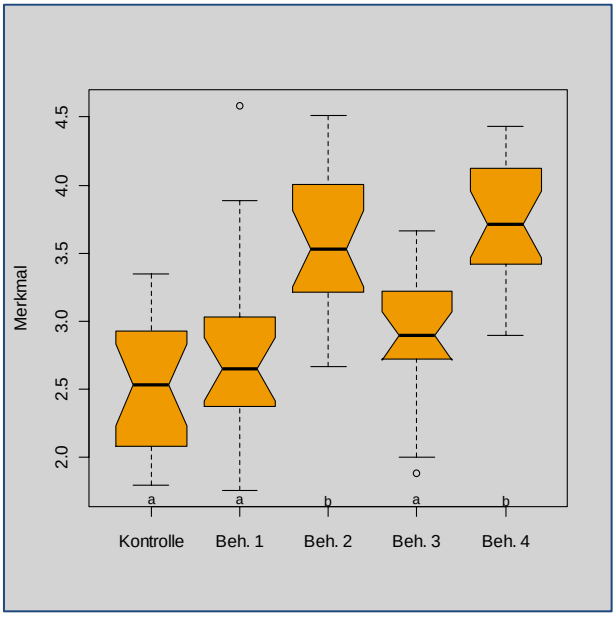
Um Interaktionen untersuchen zu können, muss ein faktorielles Design vorliegen, d.h. es existiert für jede Faktorstufen-Kombination mehr als eine Beobachtung.

Befehle in R

Mit dem Befehl `aov()` wird eine Varianzanalyse berechnet. Mit dem Befehl `summary()` bzw. genauer `summary.aov()` wird die ANOVA-Tabelle aufgerufen. Die Effektgrößen werden mit dem Befehl `summary.lm()` angezeigt.

```
mod <- aov(y ~ var1 + var2, data=dat
```

KONTRASTE



Ziel

Kontraste werden verwendet, um im Rahmen einer Varianzanalyse nachzuprüfen, ob sich die Mittelwerte der abhängigen Variablen zwischen bestimmten Behandlungen (Faktorstufenkombinationen) unterscheiden. Da Mehrfachvergleiche zu vermeiden sind, müssen die Kontraste voneinander unabhängig (orthogonal) sein.

Typen

Treatment-Kontraste (Standard in R)
Helmert-Kontraste (Standard in S-Plus)
A priori-Kontraste: Aufstellung nach Behandlungsvarianten und Hypothesen.
A posteriori Kontraste: Modellvereinfachung mit Gruppieren von Faktorstufen, die sich vermutlich nicht signifikant unterscheiden. Ergebnis zur Signifikanz über Modellvergleich.

Was passiert?

Man muss zunächst die Kontraste festlegen und anschließend die ANOVA berechnen. Dann erhält man für jeden Kontrast ein Ergebnis zur Signifikanz des geprüften Unterschieds.
A priori-Kontraste: Wenn z.B. in einem Experiment zweierlei oberirdische Rückschnittmaßnahmen und zweierlei Rückschnittintensitäten an der Wurzel vorgenommen werden, würde man prüfen wollen, ob die Behandlungen von der Kontrolle verschieden sind. Man würde prüfen, ob die oberirdischen Schnittbehandlungen andere Auswirkungen haben als unterirdische. Und man würde prüfen, ob sich die beiden Intensitäten jeweils eine unterschiedliche Auswirkung haben.

Eignung / Anforderung an die Daten

Der *a priori* Vergleich über Kontraste ist nur sinnvoll, wenn die Varianzanalyse signifikant war. Die Kontrastmatrix wird so aufgestellt, dass die Summe der gewichte der einzelnen Stufen 0 ergibt. Wird die Kontrolle gegen vier Behandlungen getestet, erhält die Kontrolle das Gewicht 4 und die Behandlungen erhalten jeweils das Gewicht -1. Werden zwei Behandlungen gegeneinander getestet, erhält eine das Gewicht 1, die andere das Gewicht -1 und alle anderen Behandlungen das Gewicht 0.

Dingend zu beachten!

Es können maximal k-1 unabhängige (orthogonale) Vergleiche vorgenommen werden (mit k: Anzahl der Behandlungen). Die Summe der Produkte zwischen den Gewichten zweier Behandlungen in der Kontrastmatrix ergibt bei Orthogonalität jeweils 0.

Befehle in R

Mit dem Befehl `contrasts(Datensatz)` werden die Kontraste festgelegt. Zum Beispiel: `contrasts(dat) <- c(c(4,-1,-1,-1,-1),c(0,1,1,-1,-1),c(0,1,-1,0,0), c(0,0,0,1,-1))`; Rückstellen per `contrasts(Datensatz) <- NULL`

LITERATURVERZEICHNIS

Empfehlenswerte Lehrbücher

Bortz, J. (2005): Statistik für Human- und Sozialwissenschaftler. Springer, Heidelberg.
Crawley, M.J. (2005): Statistics, an introduction using R. John Wiley & Sons, Chichester.
Köhler, W., Schachtel, G. & Voleske, P. (2002): Biostatistik. Springer, Berlin, Heidelberg.

Weiterführende Lehrbücher

Bortz, J., Lienert, G.A. & Boehnke, K. (2000): Verteilungsfreie Methoden in der Biostatistik. Springer, Berlin, Heidelberg.
Crawley, M.J. (2005) Statistical Computing. An introduction to data analysis using S-Plus. John Wiley & Sons, Chichester.
Crawley, M.J. (2008): The R-Book. John Wiley & Sons, Chichester.
Sachs ,L. & Hedderich, J. (2009): Angewandte Statistik: Methodensammlung mit R. 13. Auflage. Springer, Berlin, Heidelberg.

Statistikskript

Dormann, C.F., I. Kühn. 2008. Statistische Analyse biologischer Daten (mit dem freien Programmpaket R). 2. Auflage.
URL: http://www.ufz.de/data/deutschstatswork7649.pdf