

# Statistik-Übungen

## mit

Wintersemester 2010/2011

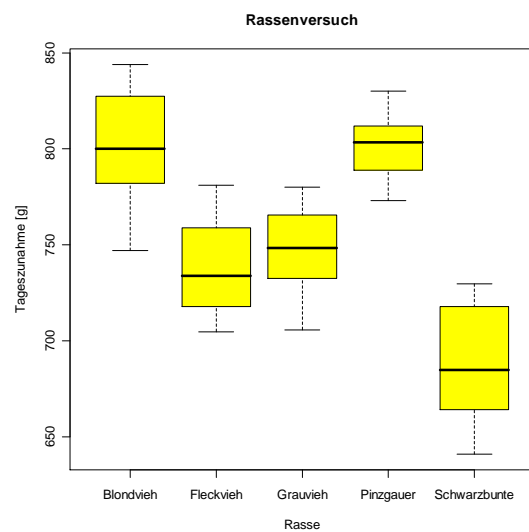
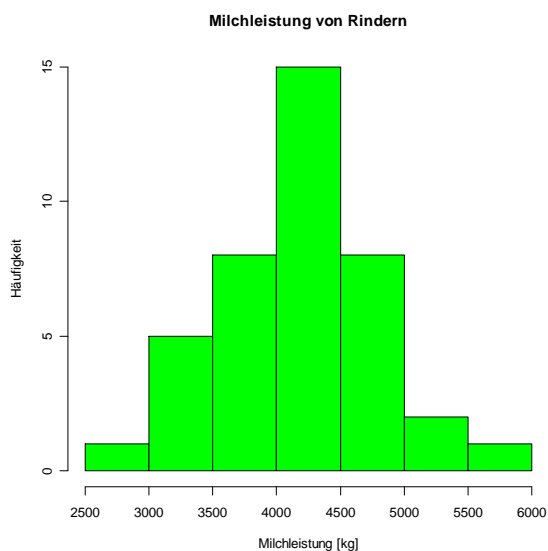
### Inhalt

Beschreibende Statistik .....	3
t-Tests .....	7
Varianzanalyse .....	11
Regression .....	15
Kreuztabellen .....	18
Nichtparametrische Verfahren .....	19

Autoren: Lydia Matiasch, Bernhard Spangl, Robert Wiedermann

Institut für Angewandte Statistik und EDV  
Department für Raum, Landschaft und Infrastruktur  
Universität für Bodenkultur, Wien

<http://www.rali.boku.ac.at/statedv.html>

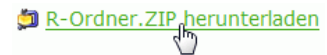


# The R Project for Statistical Computing (<http://www.r-project.org/>)



## Herunterladen / Entzippen

Klicken Sie in Moodle auf den Link 'R-Ordner.ZIP herunterladen' (oder <http://www.boku.ac.at/statedv/EinfDV/SW-Download/R-Ordner.zip>), und speichern Sie die Datei auf Ihrem PC.



Mac- und Linux-User müssen hier andere Wege gehen: Informationen finden Sie in Moodle sowie im Skriptenteil "Statistik-Sprache R" [http://statedv.boku.ac.at/roberts\\_it-kurs-unterlagen/?i=Stat-R](http://statedv.boku.ac.at/roberts_it-kurs-unterlagen/?i=Stat-R)

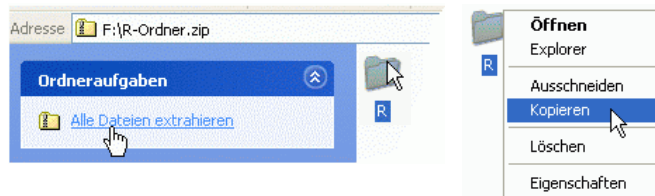
Bevor Sie R verwenden können, müssen Sie die in der .ZIP-Datei enthaltenen Dateien „auspacken“. Dafür gibt es mehrere Möglichkeiten:



Klicken Sie mit der rechten Maustaste auf die Datei 'R-Ordner.ZIP', und wählen im Kontextmenü den Menüpunkt 'Alle extrahieren ...' und folgen dem Extrahier-Assistenten.



Alternativ können Sie die .ZIP-Datei mittels Doppelklick öffnen, dann sehen Sie den „eingepackten“ Ordner "R". Kopieren Sie den Ordner "R" wie gewohnt an eine andere Stelle Ihrer Festplatte oder Ihres USB-Sticks.



Sie können nun nach dem Entpacken die heruntergeladene Datei 'R-Ordner.ZIP' wieder löschen.

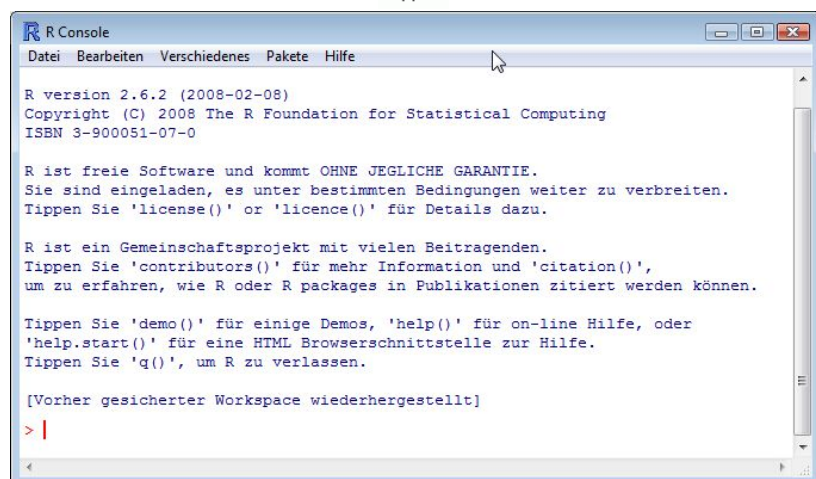
## aufrufen

Doppelklicken Sie auf den Ordner "R" und doppelklicken Sie auf das Batchprogramm '\_R\_aufrufen\_(Rgui.exe).bat'.



Damit starten Sie die R-Console, und können an der roten Eingabeaufforderung '>' Befehle eintippen:

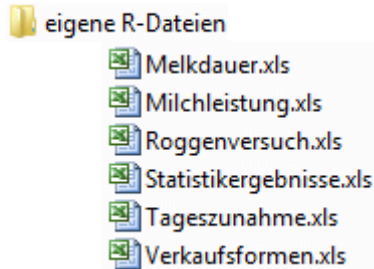
Mit den Cursortasten (Pfeil hinauf) können Sie zu den letzten Befehlen zurückblättern:



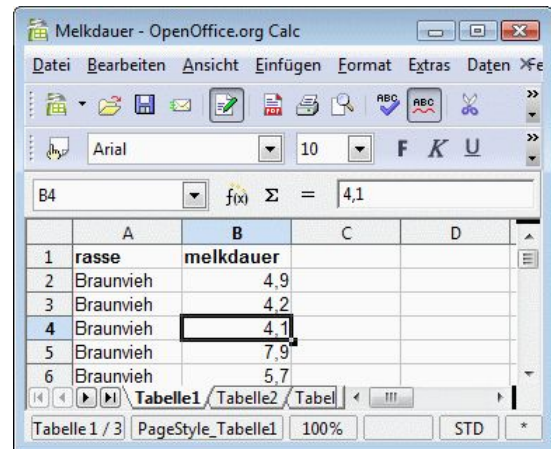
Sie können dieses Skriptum kochrezeptartig verwenden, indem Sie vor allem die gelb hinterlegten Teile mit Ihren eigenen Angaben in Übereinstimmung bringen. Für ein besseres Verständnis von "R" empfehlen wir aber, zuerst den Skriptenteil "Statistik-Sprache R" ([http://statedv.boku.ac.at/roberts\\_it-kurs-unterlagen/?i=Stat-R](http://statedv.boku.ac.at/roberts_it-kurs-unterlagen/?i=Stat-R)) durcharbeiten, den Sie ebenfalls aus Moodle herunterladen können. Für das Verständnis der statistischen Verfahren und für die Interpretation der Ergebnisse sind natürlich Vorlesungs- und Übungsskripten unentbehrlich.

## Wie bekomme ich eigene Daten in R hinein?

Im Ordner 'R' befindet sich ein Unterordner namens 'eigene R-Dateien'. Dort finden Sie bereits einige .XLS-Dateien mit den in diesem Skriptum verwendeten Daten.



Es gibt immer mehrere Möglichkeiten, eigene Daten einzugeben, wir empfehlen die Eingabe mittels eines Tabellenkalkulationsprogramms (z.B. Microsoft Excel oder OpenOffice Calc).



Achten Sie darauf, dass jede Spalte einen Spaltennamen *ohne Leerzeichen* und *ohne Umlaute* erhält!

In den Tabellenkalkulationsprogrammen ist bei deutscher oder österreichischer Landereinstellung von Windows das Dezimaltrennzeichen ein Komma! (Wichtig: In R und in den Online-Beispielen ein Dezimal-Punkt, in der Tabellenkalkulation i.d.R. ein Dezimal-Komma! Wenn Excel die Eingaben nicht rechtsbündig anzeigt, haben Sie *keine* Zahlenwerte eingegeben, mit denen gerechnet werden kann!)

Speichern Sie Ihre eigenen Daten im Ordner 'R\eigene R-Dateien', und achten Sie unbedingt beim Speichern darauf, dass Sie den Dateityp 'Excel 97/2000/XP (\*.XLS)' verwenden, und *NICHT* .XLSX oder .ODS!

Wenn Sie die aus Moodle heruntergeladene vorbereitete R-Version verwenden, können Sie .XLS-Dateien mit Hilfe der Funktion `read.xls()` einlesen. Beispiel:

```
> meine_daten <- read.xls( file.choose() )
```

Falls Sie R von <http://cran.at.r-project.org/> heruntergeladen und installiert haben, egal ob Windows, Linux oder Mac, können Sie `read.xls()` sowie die Funktionen `schiefe()`, `exzess()` und `stamtblatt()` *NICHT* verwenden. Sie können die Daten in diesem Fall im CSV-Format speichern und einlesen (Info dazu im Skriptenteil „Statistik-Sprache R“ ([http://statedv.boku.ac.at/roberts\\_it-kurs-unterlagen/?i=Stat-R](http://statedv.boku.ac.at/roberts_it-kurs-unterlagen/?i=Stat-R)) auf den Seiten 8 und 9), oder Daten mittels `data.frame()` direkt eingeben:

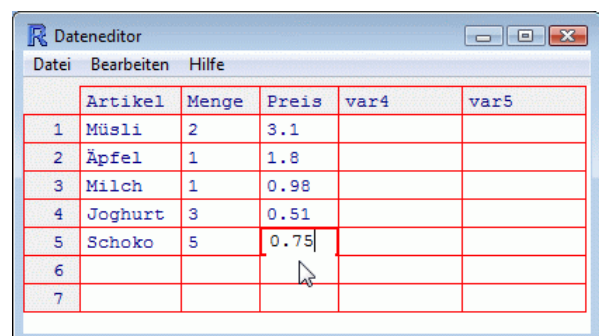
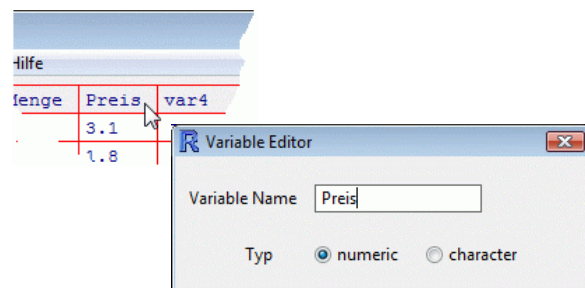
Daten direkt in Dataframes eingeben

```
> einkauf <- data.frame() Erzeugt einen neuen, leeren Dataframe.
```

```
> fix(einkauf) Ruft einen rudimentären Daten-Editor auf.
```

Die Spaltennamen lassen sich durch Anklicken umbenennen.

Beachten Sie, dass in R immer ein Punkt das Dezimal-Trennzeichen ist, *nicht* das Komma!



Im Hinblick auf die Diplomarbeit ist es sinnvoll, wenn Sie sich mit den verschiedenen Möglichkeiten des Datei-Imports (aus .XLS- oder .CSV-Dateien) auseinandersetzen. Falls Sie den Import aber nicht hinbekommen, tippen Sie Ihre Daten bitte auf diese Weise ein.

Der Aufruf des Dateneditors mittels `fix()` eignet sich auch hervorragend zur Kontrolle der mittels `read.xls()` eingelesenen Daten!

## Beschreibende Statistik

Bevor man mit einer statistischen Analyse beginnt, ist es immer wichtig, sich einen Überblick über die Daten zu verschaffen. Dazu dienen einerseits Grafiken, andererseits Kennzahlen.

Die Datei *Milchleistung.xls* enthält Daten über die Milchleistung (in kg) von 40 Rindern. Die Abbildung zeigt einen Ausschnitt der Daten.

	A
1	menge
2	5700
3	4200
4	3700
5	5000

```
> milch <- read.xls("Milchleistung.xls")
bzw.
> milch <- read.xls( file.choose() )
> milch
  menge
1  5700
2  4200
3  3700
4  5000
5  4300
....
```

Erläuterung: Die Datei 'Milchleistung.xls' wird eingelesen und in einem sogenannten DataFrame namens 'milch' abgelegt. Durch Eingabe von 'milch' können Sie den Inhalt des DataFrame anzeigen lassen. Ein DataFrame kann mehrere Spalten enthalten. Will man auf eine bestimmte Spalte zugreifen, kann man das mittels Dollarzeichen und Spaltenname tun, z.B. 'milch\$menge'. Dabei ist auch auf die Groß- und Kleinschreibung zu achten (R ist case sensitive!). Weiters sind Leerzeichen, Umlaute (äöüß) oder Sonderzeichen in Spaltennamen nicht zulässig.

Die Funktion read.xls() funktioniert NUR im vorbereiteten R-Paket!

Mac- und Linux-User:  
'Arbeitsbereich' > 'Gespeicherten Arbeitsbereich laden ...' , Ordner 'eigene R-Dateien (Mac)' öffnen, Datei 'Statistik-Uebungen.RData' öffnen.  
'Verschiedenes' > 'Arbeitsverzeichnis wechseln', Ordner 'eigene R-Dateien (Mac)' öffnen.  
> milch <- read.csv("Milchleistung.csv") bzw.  
> milch <- read.csv2("Milchleistung.csv")  
(Details in: "Statistik-Sprache R")

Alternative:  
> milch <- data.frame()  
> fix(milch)

### Kennzahlen

#### Mittelwert (mean)

```
> mean(milch$menge)
[1] 4210
```

#### Standardabweichung (standard deviation)

```
> sd(milch$menge)
[1] 608.824
```

#### Varianz (variance)

```
> var(milch$menge)
[1] 370666.7
```

#### Quantile (inkl. besonderer Quantile: Quartile und Median)

```
> quantile(milch$menge, c(0.10, 0.25, 0.50, 0.75, 0.90))
 10%  25%  50%  75%  90%
3400 3775 4200 4600 4910
```

### Zusammenfassung

```
> summary(milch$menge)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2900   3775   4200   4210   4600   5700
```

#### Schiefe (skewness)

```
> schiefe(milch$menge)
[1] 0.1039435
```

Interpretation:  $g_1 > 0$  – die Verteilung ist linkssteil. Der Wert liegt allerdings sehr nahe bei Null.

#### Exzess (kurtosis)

```
> exzess(milch$menge)
[1] -0.3618902
```

Interpretation:  $g_2 < 0$  – die Verteilung ist stumpfer als die Standardnormalverteilung. Auch dieser Wert liegt sehr nahe bei Null.

"schiefe" und "exzess" sind benutzerdefinierte Funktionen und *nicht* standardmäßig in R enthalten. Sie können sich den Quellcode durch Eingabe des Funktionsnamens (ohne Klammern und Parameter) anzeigen lassen.

```
> schiefe
function(x) {
  m3 <- mean((x-mean(x))^3)
  m3/(sd(x)^3)
}

> exzess
function(x) {
  m4 <- mean((x-mean(x))^4)
  m4/(sd(x)^4)-3
}
```

## Diagramme

### Stamm-und-Blatt-Diagramm (stem-and-leaf diagram)

```
> stamtblatt(milch$menge)
1 | 2: represents 1200
leaf unit: 100
      n: 40
  1   2* | 9
      30 |
  3   t  | 23
  6   f  | 445
 10   s  | 6677
 13   3* | 899
 17   40 | 0111
(8)   t  | 22222333
 15   f  | 4555
 11   s  | 666
  8   4* | 8999
  4   50 | 01
  2   t  | 3
      f  |
  1   s  | 7
```

Die Funktion wählt selbständig eine Darstellung. Diese können Sie aber durch Angabe von Parametern ändern:  
 unit ....leaf unit, as a power of 10 (e.g., 100, .01); omit to let the function choose the unit.

m .....number of parts (1, 2, or 5) into which each stem should be divided; omit to let the function choose the number of parts/stem.

```
> stamtblatt(milch$menge, m=2, unit=100)
```

Falls Sie nicht den vorbereiteten R-Ordner verwenden (also alle Mac- und Linux-UserInnen), müssen Sie den Workspace 'Statistik-Uebung.RData' laden, sonst steht die Funktion stamtblatt() nicht zur Verfügung.

Unter bestimmten Umständen treten bei der Funktion stamtblatt Rundungsfehler auf und die Werte mancher Blätter sind nicht korrekt. Verwenden Sie in diesem Fall die Funktion stem:

```
> stem(milch$menge)
```

Im Gegensatz zur Darstellung im Vorlesungs- und Übungsskriptum ist die Tiefe hier links angegeben. Die Tiefe der leeren "Blätter" ist jeweils dieselbe wie die jenes angrenzenden Blattes, das näher beim Rand liegt. Die Verteilung sieht nahezu symmetrisch aus. Dies erklärt auch, warum der Wert für die Schiefe sehr nahe bei Null liegt.

### Histogramm (histogram)

```
> hist(milch$menge)

> hist(milch$menge, col="orange",
      xlab="Milchleistung [kg]",
      ylab="Häufigkeit",
      main="Milchleistung von Rindern")
```

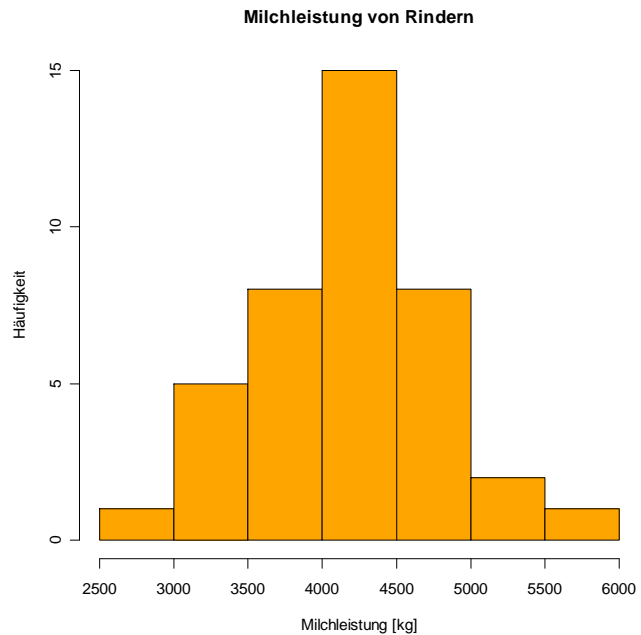
"hist" zeichnet ein relativ farbloses Histogramm.

Mit den Parametern

```
col .. color, Farbe (Eingabe als Text oder Zahlenwert)
xlab .. x label, Beschriftung der x-Achse
ylab .. y label, Beschriftung der y-Achse
main .. Titel
```

kann das Histogramm zusätzlich gestaltet werden. Sie sollten sich angewöhnen, die Achsen eines Diagramms immer korrekt und vollständig mit Einheiten zu beschriften.

Aus dem Histogramm ist auch zu erkennen, dass die Verteilung nahezu symmetrisch ist.

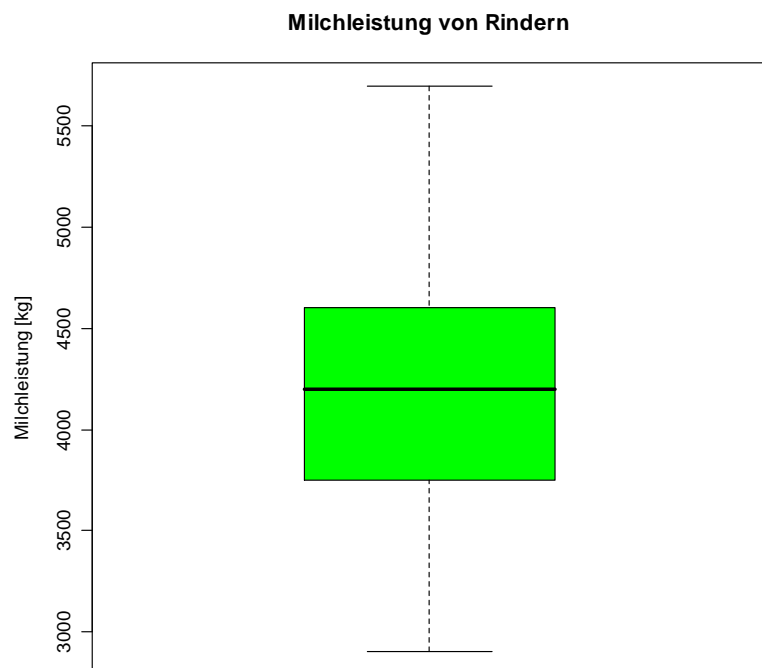


### Kastendiagramm (boxplot)

```
> boxplot(milch$menge)

> boxplot(milch$menge,
      col="green",
      ylab="Milchleistung [kg]",
      main="Milchleistung von Rindern")
```

Aus dem Kastendiagramm ist erkennbar, dass es keine Ausreißer gibt.



## Wie bekomme ich die Ergebnisse aus R in die Textverarbeitung?

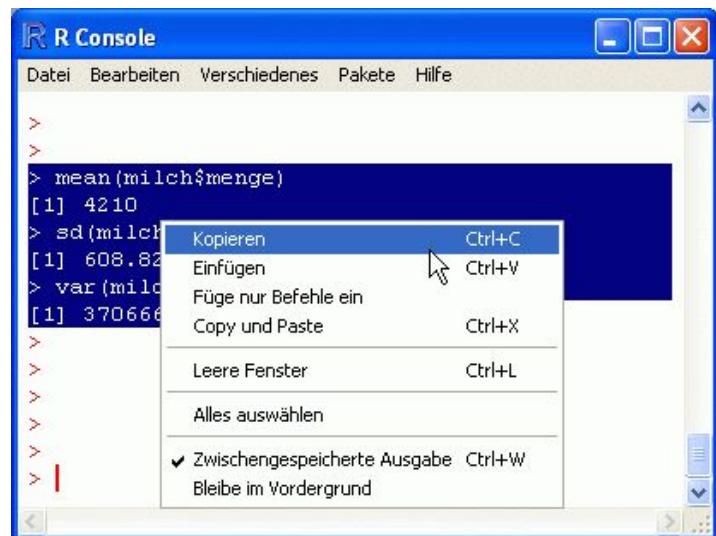
Wenn Sie eine R-Sitzung beenden, werden zwar auf Wunsch die erstellten Objekte wie Vektoren, Dataframes, selbst erstellte Funktionen etc. sowie die eingetippten Befehle gespeichert (wenn Sie beim Beenden "Workspace sichern?" mit "Ja" beantworten) und bei der nächsten Sitzung wieder bereitgestellt, die Ergebnisse sind dann aber nicht mehr vorhanden.

Sie sollten daher Ihre Ergebnisse, gleich nachdem Sie sie erhalten haben, mittels Zwischenablage in ein Textverarbeitungsdocument kopieren:

Laden Sie die 'Vorlage für Abgabedatei' aus Moodle herunter und öffnen Sie diese Datei mittels Textverarbeitungsprogramm (z.B. MS Word oder OpenOffice Writer). In den Benutzerräumen der Boku steht Ihnen OpenOffice Writer zur Verfügung. Tragen Sie bitte Ihren Namen, Matrikelnummer, Gruppe, Beispielnummer, Beispielcode, Semester vollständig in die Kopfzeile ein und speichern Sie die Datei genau so wie es in der 'Anleitung zu Abgabe der EDV-Beispiele' beschrieben ist. Bei elektronischer Abgabe müssen Dateityp und Dateiname präzise den Vorgaben entsprechen, sonst wird die Abgabedatei nicht akzeptiert.

Markieren Sie die relevanten Ergebnisse in der R Console, und kopieren Sie den markierten Text in die Zwischenablage (z.B. mit der Tastenkombination Strg+C, oder mit dem Menüpunkt 'Bearbeiten' > 'Kopieren', oder Sie klicken mit der rechten Maustaste auf den markierten Block und wählen den Punkt 'Kopieren' aus dem Kontextmenü.

(Auch wenn Sie im Forum Fragen stellen, z.B. wegen einer Fehlermeldung, kopieren Sie bitte auf diese Weise nicht nur die vollständige Fehlermeldung, sondern prinzipiell auch immer Ihre Daten und alles, was Sie bisher getippt haben, mit ins Forum. Nur auf diese Weise können Probleme sinnvoll nachvollzogen werden, anderenfalls sind die ForenteilnehmerInnen, von denen Sie ja eine Antwort erhoffen, auf Herumraten angewiesen.)



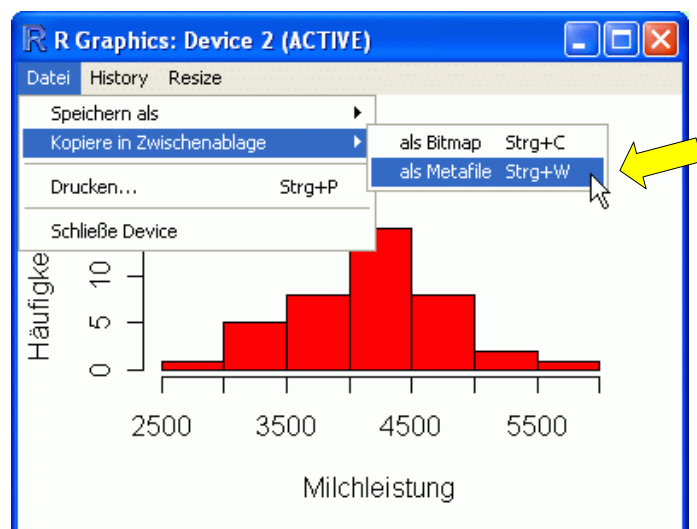
Im Textverarbeitungsdocument können Sie den Inhalt der Zwischenablage mit der Tastenkombination Strg+V oder mit 'Bearbeiten' > 'Einfügen' an der Cursorposition einfügen.

Formatieren Sie diese Ergebnisse im Textverarbeitungsprogramm mit der Schriftart 'Courier New', damit die Werte korrekt untereinander ausgerichtet stehen.

Bei Diagrammen im Grafikfenster von R wählen Sie den Menüpunkt 'Datei' > 'Kopiere in Zwischenablage' > 'als Metafile'.

(Metafile ist ein Vektorformat und braucht daher wenig Speicherplatz und kann beliebig skaliert werden. Verwenden Sie für die Abgabe unbedingt ein Vektorgrafikformat.

Pixel-/Bitmap-Formate wie z.B. .PNG, .BMP, .TIFF, .JPEG werden bei der Konversion ins .RTF-Format anders codiert, was i.d.R. den Speicherbedarf stark erhöht.)



## t-Tests

### Vergleich eines Mittelwerts mit einer Konstanten

Die Frage, ob die mittlere Milchleistung kleiner oder gleich 4000 kg ist, kann mit Hilfe eines t-Tests beantwortet werden. Die zugehörige Nullhypothese lautet: "Der theoretische Mittelwert für die Milchleistung ist kleiner oder gleich 4000 kg." -  $H_0: \mu \leq 4000$  kg). Die Alternativhypothese lautet: "Der theoretische Mittelwert für die Milchleistung ist größer als 4000 kg." Das Risiko 1. Art wird mit  $\alpha = 0.05$  festgelegt.

```
> t.test(milch$menge, mu=4000, alternative="greater", conf.level=0.95)
```

```
One Sample t-test

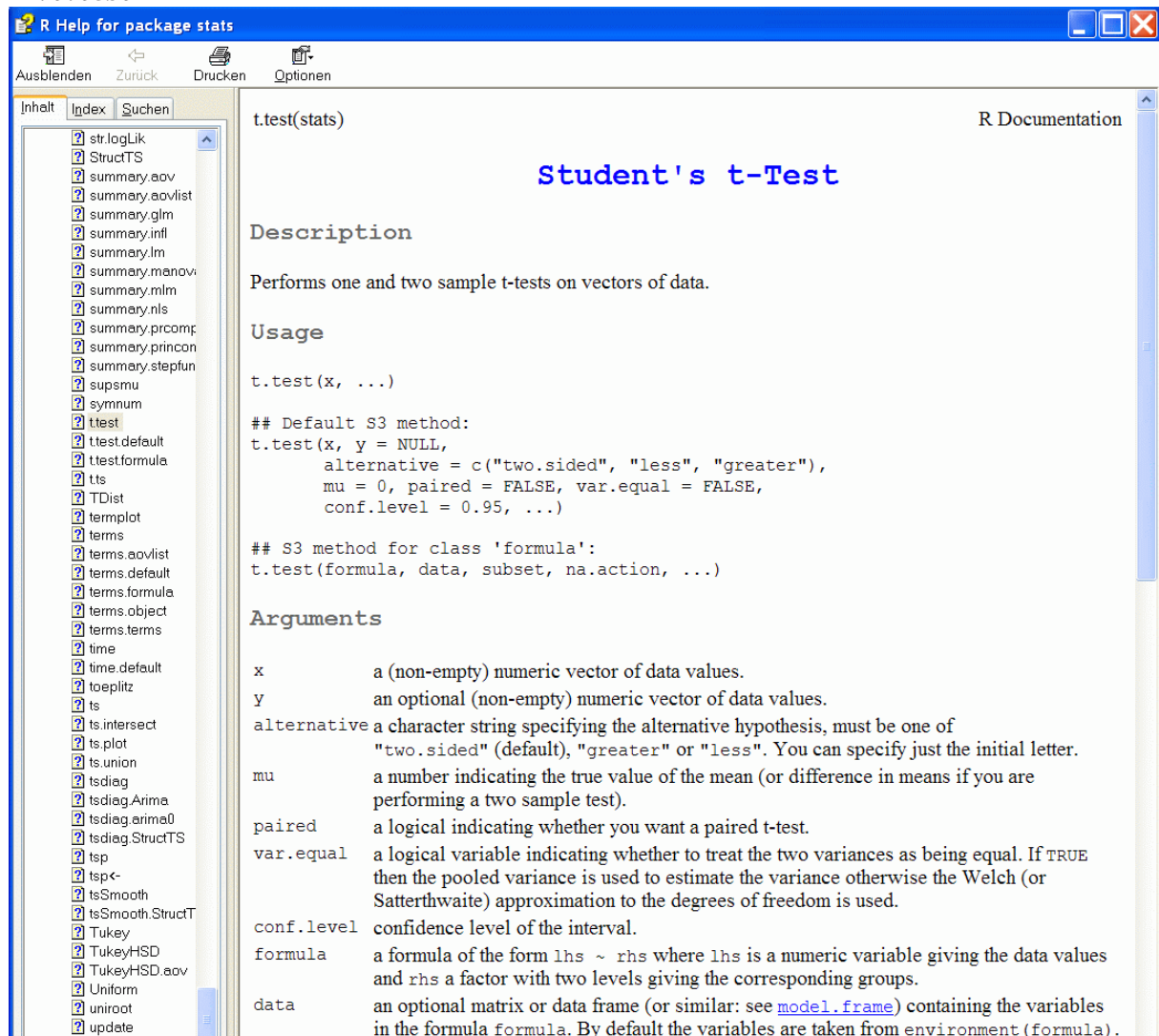
data:  milch$menge
t = 2.1815, df = 39, p-value = 0.01762
alternative hypothesis: true mean is greater than 4000
95 percent confidence interval:
 4047.808      Inf
sample estimates:
mean of x
 4210
```

t .....Wert der Teststatistik des t-Tests  
df .....Anzahl der Freiheitsgrade (degrees of freedom)  
p-value.....p-Wert

Der p-Wert (0.01762) ist kleiner als das gewählte Risiko 1. Art  $\alpha = 0.05$ ; daher ist die Nullhypothese abzulehnen. Der theoretische Mittelwert liegt also über 4000 kg.

Die Abbildung zeigt einen Ausschnitt der Hilfe für den t-Test: Es werden die Bedeutung der einzelnen Parameter und ggf. Alternativen erklärt. Es ist hier wichtig zu beachten, dass nicht die Hypothese selber, sondern die Alternativhypothese angegeben wird. Außerdem kann man den Konfidenzoeffizienten für das Konfidenzintervall wählen. Sie müssen jedoch beachten, dass das Konfidenzintervall immer dem Test entsprechend einseitig oder zweiseitig ist.

```
> ?t.test
```



**Student's t-Test**

**Description**  
Performs one and two sample t-tests on vectors of data.

**Usage**  
t.test(x, ...)

**Arguments**

- x a (non-empty) numeric vector of data values.
- Y an optional (non-empty) numeric vector of data values.
- alternative a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter.
- mu a number indicating the true value of the mean (or difference in means if you are performing a two sample test).
- paired a logical indicating whether you want a paired t-test.
- var.equal a logical variable indicating whether to treat the two variances as being equal. If TRUE then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.
- conf.level confidence level of the interval.
- formula a formula of the form lhs ~ rhs where lhs is a numeric variable giving the data values and rhs a factor with two levels giving the corresponding groups.
- data an optional matrix or data frame (or similar: see [model.frame](#)) containing the variables in the formula formula. By default the variables are taken from environment (formula).

Um ein zweiseitiges 90%-iges Konfidenzintervall für den Mittelwert zu berechnen, muss neuerlich ein t-Test durchgeführt werden.

```
> t.test(milch$menge, mu=4000, alternative="two.sided", conf.level=0.90)
```

One Sample t-test

```
data: milch$menge
t = 2.1815, df = 39, p-value = 0.03524
alternative hypothesis: true mean is not equal to 4000
90 percent confidence interval:
 4047.808 4372.192
sample estimates:
mean of x
 4210
```

Damit erhalten wir als Grenzen für ein zweiseitiges 90%-iges Konfidenzintervall für die mittlere Milchleistung 4047.808 und 4372.192 kg.

### Vergleich zweier Mittelwerte für unabhängige Stichproben

In einem Aufzuchtversuch wurden zwei Gruppen von Kälbern Futtermittel mit unterschiedlicher Proteinzugabe verabreicht. Die Datei *Kaelber.xls* enthält die entsprechenden Daten für die mittlere Tageszunahme (in kg).

	A	B
1	futter	zunahme
2	1	1,36
3	1	1,32
4	1	1,24
5	1	1,42
6	1	1,38
	1	1,24
	1	1,4
	1	1,43
	1	1,12
	1	1,48
	1	1,45
	1	1,35
	1	1,56
	1	1,43
	1	1,26
	2	1,31
18	2	1,21
19	2	1,28

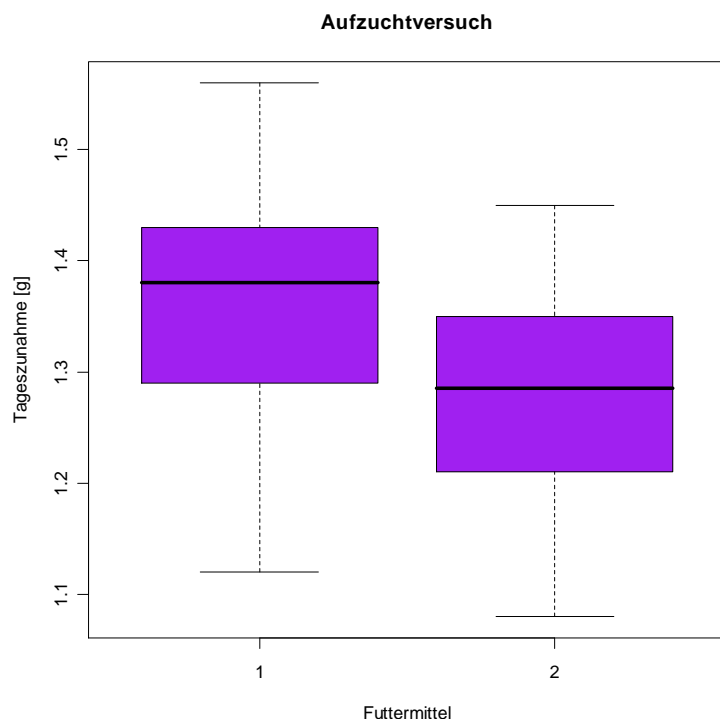
```
> kaelber <- read.xls( file.choose() )
> kaelber
  futter  zunahme
1      1    1.36
2      1    1.32
..     .    ....
15     1    1.26
16     2    1.31
..     .    ....
29     2    1.30
```

Achten Sie auf die exakte Schreibweise der Spaltennamen, auch Groß- und Kleinschreibung ist zu beachten! Weiters sind Leerzeichen, Umlaute oder Sonderzeichen in Spaltennamen nicht zulässig. In der Tabellenkalkulation ist das Dezimalzeichen (abhängig von den Ländereinstellungen in Windows) i.d.R. das Komma, nicht der Punkt.

### Kastendiagramm (boxplot)

Mit Hilfe eines gruppierten Kastendiagramms verschafft man sich einen Überblick über die Daten. Man erkennt, dass sich die mittleren Tageszunahmen (Median) unterscheiden und die Streuungen (Kastenlängen = Interquartilabstände) ähnlich groß sind.

```
> boxplot(kaelber$zunahme ~
kaelber$futter,
xlab="Futtermittel",
ylab="Tageszunahme [g]",
main="Aufzuchtversuch",
col="purple")
```



## Levene-Test

Die Frage, ob die Tageszunahme je nach Futtermittel unterschiedlich ist, kann zum Beispiel mit Hilfe eines t-Tests beantwortet werden. Um die Parameter richtig zu wählen, ist es notwendig zu wissen, ob die Varianzen für beide Gruppen gleich sind. Die Varianzhomogenität kann mit Hilfe des Levene-Tests überprüft werden. Die dazugehörige Nullhypothese lautet: "Die Varianzen beider Gruppen sind gleich." Die Alternativhypothese lautet: "Die Varianzen beider Gruppen sind ungleich." Das Risiko 1. Art wird wie üblich mit  $\alpha = 0.05$  festgesetzt.

```
> levene.test(kaelber$zunahme, kaelber$futter)
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  1  0.1102 0.7425
      27
```

Df .....Anzahl der Freiheitsgrade  
(degrees of freedom)  
F value .....Wert der Teststatistik  
Pr(>F) .....p-Wert

Da der p-Wert (0.7425) über  $\alpha = 0.05$  liegt, muss die Hypothese gleicher Varianzen beibehalten werden.

## t-Test

Die Nullhypothese für den t-Test lautet: "Die theoretische mittlere Tageszunahme ist für beide Futtermittel gleich." –  $H_0: \mu_1 = \mu_2$ . Die Alternativhypothese lautet: "Die theoretische mittlere Tageszunahme ist für beide Futtermittel ungleich." Das Risiko 1. Art wird mit  $\alpha = 0.05$  festgelegt.

```
> t.test(kaelber$zunahme ~
kaelber$futter, alternative =
"two.sided", paired = FALSE,
var.equal = TRUE, conf.level
= 0.95)
```

Auch hier ist es wieder wichtig, die richtige Alternativhypothese und das gewünschte Risiko 1. Art zu wählen. Da es sich um zwei unabhängige Stichproben und nicht um gepaarte Beobachtungen handelt, ist bei 'paired' der Wert 'FALSE' zu wählen. Je nachdem, ob die Varianzen gleich sind oder nicht, wird 'var.equal' auf 'TRUE' oder auf 'FALSE' gesetzt.

Die Tilde '~' bedeutet, dass die Werte für zunahme nach dem Faktor futter gruppiert werden. Dies ist notwendig, weil jede Zeile des DataFrames einem eigenen Fall, hier also jeweils einem anderen Kalb entspricht.

Two Sample t-test

```
data: kaelber$zunahme by kaelber$futter
t = 2.1612, df = 27, p-value = 0.03971
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.004327426 0.166720193
sample estimates:
mean in group 1 mean in group 2
 1.362667      1.277143
```

Der p-Wert (0.03971) liegt unter dem gewählten Risiko 1. Art  $\alpha = 0.05$ , daher muss die Hypothese gleicher mittlerer Tageszunahmen bei beiden Futtermitteln verworfen werden.

## Vergleich zweier Mittelwerte für gepaarte Beobachtungen

Bei einem Forschungsprojekt über die Schadstoffbelastung der Luft wurde an 21 Messpunkten an zwei Terminen (März/Juli) die Anzahl an Asbestfasern pro Kubikmeter Luft gemessen. Die Messwerte sind in der Datei *Asbest.xls* enthalten.

```
> asbest <- read.xls( file.choose() )
> asbest
  maerz juli
1    450  530
2    670  780
3    650  530
4    480  510
5    460  550
6    580  610
..    ...  ...
```

	A	B
1	maerz	juli
2	450	530
3	670	780
	650	530
	480	510
	460	550
	580	610
	250	450
	320	440
	750	550
	600	780
	200	380
	540	710
	520	500

Bei gepaarten Beobachtungen müssen die unterschiedlichen Messwerte als getrennte Variable, d.h. Spalten, eingegeben werden. Die beiden zusammengehörenden Beobachtungen müssen dabei in einer Zeile stehen.

Achten Sie auf die exakte Schreibweise der Spaltennamen, auch Groß- und Kleinschreibung ist zu beachten!

Weiters sind Leerzeichen und Umlaute in Spaltennamen nicht zulässig.

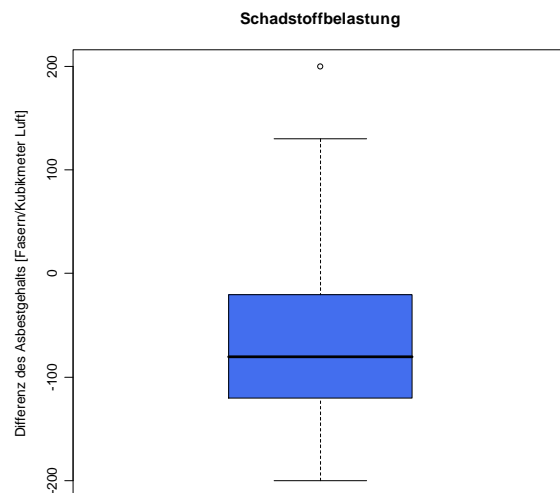
In der Tabellenkalkulation ist das Dezimalzeichen (abhängig von den Ländereinstellungen in Windows) i.d.R. das Komma, nicht der Punkt.

### Kastendiagramm (boxplot)

Mit Hilfe eines Kastendiagramms der Differenzen bekommt man einen Überblick über die Daten.

```
> boxplot( asbest$maerz - asbest$juli,
  ylab="Differenz des Asbestgehalts
[Fasern/Kubikmeter Luft]",
  main="Schadstoffbelastung",
  col="royalblue2" )
```

Aus der Darstellung ist erkennbar, dass die Belastung an den meisten Messpunkten im Juli höher ist als jene im März. (Die Differenzen März-Juli sind überwiegend negativ.)



### t-Test

Die Frage, ob sich die Schadstoffbelastung im März und im Juli unterscheidet, kann z.B. mit Hilfe eines t-Tests für gepaarte Beobachtungen beantwortet werden. Die Nullhypothese lautet: "Der theoretische mittlere Asbestgehalt ist an beiden Terminen gleich." –  $H_0: \mu_{\text{maerz}} = \mu_{\text{juli}}$ . Die Alternativhypothese lautet: "Der theoretische mittlere Asbestgehalt ist an beiden Terminen ungleich." Das Risiko 1. Art wird mit  $\alpha = 0.05$  festgelegt.

```
> t.test(asbest$maerz, asbest$juli, alternative = "two.sided", paired=TRUE,
  conf.level = 0.95)
```

Paired t-test

```
data: asbest$maerz and asbest$juli
t = -2.3205, df = 20, p-value = 0.031
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -107.606441  -5.726892
sample estimates:
mean of the differences
 -56.66667
```

Der p-Wert (0.031) liegt unter dem Risiko 1. Art  $\alpha = 0.05$ , daher muss die Hypothese gleicher Schadstoffbelastung verworfen werden. Der mittlere Asbestgehalt an den beiden Terminen ist daher nicht gleich.

## Varianzanalyse (ANalysis Of VAriance, ANOVA)

### Einfache Varianzanalyse

Bei der Schlachtleistungsprüfung wird u.a. auch die Nettozunahme in Gramm pro Tag (Schlachtkörpergewicht/Mastdauer) festgestellt. Die Werte für Rinder verschiedener Rassen sind in der Datei *Tageszunahme.xls* angegeben.

```
> mast <- read.xls( file.choose() )
> mast$rasse <- factor(mast$rasse)
> mast
```

	rasse	zunahme
1	Blondvieh	784
2	Blondvieh	747
3	Blondvieh	831
4	Blondvieh	780
5	Blondvieh	810
6	Blondvieh	824
7	Blondvieh	844
8	Blondvieh	790
9	Fleckvieh	705
10	Fleckvieh	735
11	Fleckvieh	768
12	Fleckvieh	750
13	Fleckvieh	781
14	Fleckvieh	715
15	Fleckvieh	721
16	Fleckvieh	733
17	Grauvieh	734
18	Grauvieh	756
..	...	...

	A	B	C
1	rasse	zunahme	
2	Blondvieh	784	
3	Blondvieh	747	
4	Blondvieh	831	
5	Blondvieh	780	
6	Blondvieh	810	
7	Blondvieh	824	
8	Blondvieh	844	
9	Blondvieh	790	
10	Fleckvieh	705	
11	Fleckvieh	735	
12	Fleckvieh	768	
13	Fleckvieh	750	
14	Fleckvieh	781	
15	Fleckvieh	715	
16	Fleckvieh	721	
17	Fleckvieh	733	
18	Fleckvieh	734	
19	Fleckvieh	756	
20	Fleckvieh	706	
21	Fleckvieh	780	
22	Fleckvieh	768	
23	Fleckvieh	731	
24	Fleckvieh	741	
25	Fleckvieh	763	
26	Fleckvieh	809	
27	Fleckvieh	780	
28	Fleckvieh	830	
29	Fleckvieh	773	
30	Fleckvieh	798	
31	Fleckvieh	805	
32	Fleckvieh	802	
33	Fleckvieh	815	
34	Fleckvieh	641	
35	Schwarzbunte	728	
36	Schwarzbunte	654	
37	Schwarzbunte	678	
38	Schwarzbunte	708	
39	Schwarzbunte	730	
40	Schwarzbunte	692	
41	Schwarzbunte	675	

Die Dateneingabe für die Varianzanalyse bereitet immer wieder Verständnisprobleme. Beachten Sie bitte, dass die gemessenen Werte alle untereinander in einer Spalte stehen müssen, und die Faktorstufen in der Spalte daneben.

Die Funktion `factor()` wandelt die entsprechende Spalte in einen Faktor um. (Falls die Werte des Faktors als Text eingegeben, ist das nicht notwendig, falls die Werte des Faktors numerisch sind, müssen sie unbedingt in einen Faktor umgewandelt werden.)

Sie können mit `str(mast)` überprüfen, ob die Spalten als Faktor oder als Zahlen (`num` bzw. `int`) interpretiert werden.

Achten Sie auf die exakte Schreibweise der Spaltennamen, auch Groß- und Kleinschreibung ist zu beachten. Weiters sind Leerzeichen und Umlaute in Spaltennamen nicht zulässig.

In der Tabellenkalkulation ist das Dezimalzeichen (abhängig von den Ländereinstellungen in Windows) i.d.R. das Komma, nicht der Punkt.

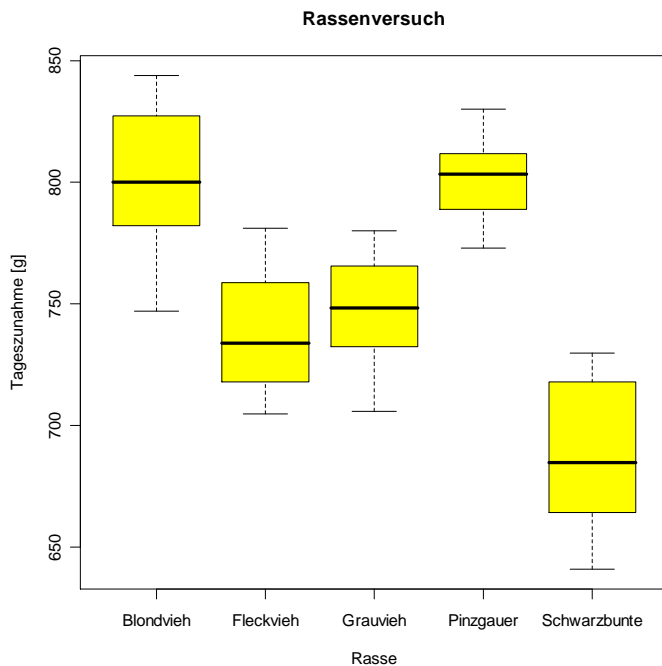
### Kastendiagramm (boxplot)

Ein gruppiertes Kastendiagramm gibt einen Überblick über die Daten.

```
> boxplot(mast$zunahme ~ mast$rasse, xlab="Rasse",
ylab="Tageszunahme [g]", main="Rassenversuch", col="yellow")
```

Die Darstellung zeigt, dass die Verteilungen der einzelnen Gruppen ungefähr symmetrisch sind und es keine Ausreißer gibt. Die Varianzen sind etwa gleich groß, es bestehen aber deutliche Unterschiede bezüglich der mittleren Tageszunahmen.

Die Frage, ob die mittlere Tageszunahme für alle Rassen gleich ist, kann z.B. mit Hilfe einer einfachen Varianzanalyse beantwortet werden.



## Levene-Test

Voraussetzung für die Varianzanalyse ist, dass die Varianzen für alle Gruppen gleich sind. Dies kann wieder mit Hilfe eines Levene-Tests überprüft werden. Die Nullhypothese lautet: "Die Varianzen aller Gruppen sind gleich." Die Alternativhypothese lautet: "Die Varianzen von mindestens zwei Gruppen sind unterschiedlich." Das Risiko 1. Art wird wie üblich mit  $\alpha = 0.05$  festgesetzt.

```
> levene.test(mast$zunahme, mast$rasse)
Levene's Test for Homogeneity of Variance
  Df F value Pr(>F)
group 4  1.0848  0.379
  35
```

Falls Sie nicht den vorbereiteten R-Ordner verwenden (also alle Mac- und Linux-UserInnen), müssen Sie den Workspace 'Statistik-Uebung.RData' laden, sonst steht die Funktion `levene.test()` nicht zur Verfügung.

Da der p-Wert (0.379) über  $\alpha = 0.05$  liegt, muss die Hypothese gleicher Varianzen beibehalten werden.

## ANOVA

Die Nullhypothese für die einfache Varianzanalyse lautet: "Die theoretische mittlere Tageszunahme ist für alle Rassen gleich." Die Alternativhypothese lautet: "Die theoretische mittlere Tageszunahme unterscheidet sich für mindestens zwei Rassen." Das Risiko 1. Art wird mit  $\alpha = 0.05$  festgelegt. `lm()` steht für *linear model*.

```
> anova(lm(mast$zunahme ~ mast$rasse))
```

Analysis of Variance Table

```
Response: mast$zunahme
      Df Sum Sq Mean Sq F value    Pr(>F)
mast$rasse  4  72692   18173  24.722 8.853e-10 ***
Residuals  35  25729     735
```

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

lm ..... linear model  
Df..... Anzahl der Freiheitsgrade (degrees of freedom)  
Sum Sq..... Quadratsumme (sum of squares)  
Mean Sq Mittlere Quadratsumme (mean sum of squares)  
F value..... Wert der Teststatistik  
Pr(>F) ..... p-Wert  
Die Zeile 'mast\$rasse' enthält die Werte für den Faktor, 'Residuals' für den Fehler.

Da der p-Wert ( $8.853e-10 = 8.853 \cdot 10^{-10}$ ) unter dem Risiko 1. Art  $\alpha = 0.05$  liegt, muss die Hypothese verworfen werden. Es besteht also ein Unterschied in den Tageszunahmen der verschiedenen Rassen. Die 'Signif. Codes' geben an, auf welchem Niveau der Test signifikant ist. Die Mittlere Quadratsumme des Fehlers (735) entspricht der Fehlervarianz  $s_e^2$ .

## Zweifache (zweifaktorielle) Varianzanalyse

Die Datei *Roggenversuch.xls* enthält Daten über den Kornertrag (in dt/ha) verschiedener Roggensorten, einmal unter Beigabe von Steinmehl zur verwendeten Gülle, einmal ohne.

```
> rogggen <- read.xls( file.choose() )
> rogggen$sorte <- factor(rogggen$sorte)
> rogggen$steinmehl <- factor(rogggen$steinmehl)
> rogggen
  sorte  steinmehl  ertrag
1 EhoKurz      mit    42.0
2 EhoKurz      mit    45.2
3 EhoKurz      mit    46.3
4 EhoKurz      mit    44.7
5 EhoKurz     ohne    41.6
6 EhoKurz     ohne    43.7
7 EhoKurz     ohne    41.5
8 EhoKurz     ohne    40.8
9   Motto      mit    39.9
10  Motto      mit    42.4
   ...      ...      ...
```

	A	B	C	D
1	sorte	steinmehl	ertrag	
2	EhoKurz	mit		42
3	EhoKurz	mit		45,2
4	EhoKurz	mit		46,3
5	EhoKurz	mit		44,7
6	EhoKurz	ohne		41,6
7	EhoKurz	ohne		43,7
8	EhoKurz	ohne		41,5
9	EhoKurz	ohne		40,8
10	Motto	mit		39,9
11	Motto	mit		42,4
12	Motto	mit		41,8
13	Motto	mit		40,4
14	Motto	mit		42,6
15	Motto	mit		43,7
16	Motto	mit		41,8
17	Motto	mit		42,4
18	Motto	mit		39,5
19	Motto	mit		41,3
20	Motto	mit		39,6
21	Motto	mit		41,9
22	Motto	mit		37,4
23	Motto	mit		39
24	Motto	mit		36,2
25	Kustro	ohne		37,4

Die Dateneingabe für die Varianzanalyse bereitet immer wieder Verständnisprobleme. Beachten Sie bitte, dass die gemessenen Werte alle untereinander in einer Spalte stehen müssen, und die Faktorabstufungen in der Spalte daneben.

In der Tabellenkalkulation ist das Dezimalzeichen (abhängig von den Ländereinstellungen in Windows) i.d.R. das Komma, nicht der Punkt. Achten Sie auf die exakte Schreibweise der Spaltennamen, auch Groß- und Kleinschreibung ist zu beachten! Weiters sind Leerzeichen und Umlaute (äöüß) in Spaltennamen nicht zulässig.

### Kastendiagramm (boxplot)

Ein gruppiertes Kastendiagramm verschafft einen Überblick über die Daten.

```
> boxplot(rogggen$ertrag ~ rogggen$sorte
* rogggen$steinmehl, ylab="Ertrag
[dt/ha]", main="Roggenversuche",
col="green4")
```

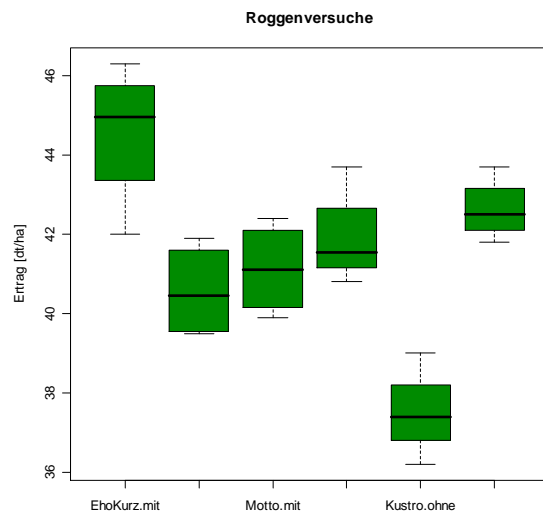
Die Tilde '~' bedeutet, dass die Werte für ertrag nach den Faktoren sorte und steinmehl gruppiert werden. Zwischen den beiden Faktoren steht das Malzeichen '\*'.

Die Fragen nach signifikanten Unterschieden im Kornertrag zwischen den Sorten bzw. bei Zugabe von Steinmehl, sowie nach signifikanten Wechselwirkungen zwischen den Faktoren Sorte und Steinmehl in Bezug auf den Ertrag, können mit Hilfe einer zweifachen (zweifaktoriellen) Varianzanalyse beantwortet werden.

### Levene-Test

Um die Gleichheit der Varianzen mittels Levene-Test überprüfen zu können, ist es notwendig zuerst einen neuen Gruppenfaktor einzuführen.

```
> rogggen$gruppierung <-
factor(10*as.numeric(rogggen$sorte)+as.numer
ic(rogggen$steinmehl))
```



Die Nullhypothese lautet: "Die Varianzen aller Faktorstufenkombinationen sind gleich." Die Alternativhypothese lautet: "Die Varianzen von mindestens zwei Faktorstufenkombinationen sind unterschiedlich." Das Risiko 1. Art wird wie üblich mit  $\alpha = 0.05$  festgesetzt.

```
> levene.test(rogggen$ertrag,
rogggen$gruppierung)
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  5  0.3908 0.8486
      18
```

Für den neuen Gruppenfaktor wählen wir sorte als Zehnerstelle und steinmehl als Einerstelle. Dazu müssen die "Werte" allerdings erst mittels der Funktion 'as.numeric'. in Zahlen umgewandelt werden.

## ANOVA

Die entsprechenden Nullhypothesen für die Varianzanalyse lauten:

$H_{\text{Sorte}}$ : Die Sorte hat keinen Einfluss auf den theoretischen mittleren Ertrag.

$H_{\text{Steinmehl}}$ : Die Zugabe von Steinmehl zur Gülle hat keinen Einfluss auf den theoretischen mittleren Ertrag.

$H_{\text{Steinmehl:Sorte}}$ : Es gibt keine Wechselwirkungen zwischen den beiden Faktoren Sorte und Steinmehl.

Die Alternativhypothesen lauten: "Sorte und Zugabe von Steinmehl zur Gülle haben Einfluss auf den theoretischen mittleren Ertrag und es gibt Wechselwirkungen zwischen den beiden Faktoren." Das Risiko 1. Art wird mit  $\alpha = 0.05$  festgelegt. `lm()` steht für *linear model*.

```
> anova(lm(roggen$ertrag ~ rogggen$steinmehl * rogggen$sorte))
```

Analysis of Variance Table

Response: rogggen\$ertrag

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
roggen\$steinmehl	1	11.900	11.900	7.3731	0.014181 *
roggen\$sorte	2	73.091	36.545	22.6424	1.218e-05 ***
roggen\$steinmehl:roggen\$sorte	2	25.556	12.778	7.9168	0.003414 **
Residuals	18	29.052	1.614		

'+'.....Modell ohne Wechselwirkungen  
'\*'. ....Modell mit Wechselwirkungen  
'\*' steht kurz für das Modell 'sorte + steinmehl + sorte: steinmehl'

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Die p-Werte für alle drei Hypothesen liegen unter dem gewählten Risiko 1. Art  $\alpha = 0.05$ , daher müssen alle Hypothesen verworfen werden. Es haben daher sowohl die Zugabe von Steinmehl zur Gülle als auch die Sorte Einfluss auf den Ertrag und es gibt Wechselwirkungen zwischen diesen beiden Faktoren. Die Mittlere Quadratsumme des Fehlers (1.614) entspricht der Fehlervarianz  $s_e^2$ .

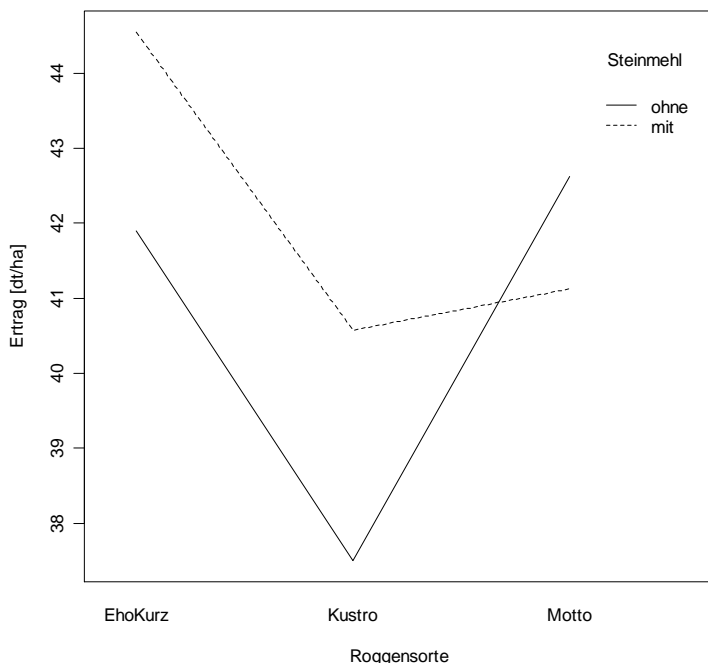
## Profildigramm

Mittels Profildigramm können die Mittelwerte für alle Faktorstufenkombinationen vergleichend dargestellt werden.

```
> interaction.plot(roggen$sorte, rogggen$steinmehl, rogggen$ertrag,
main="Profildigramm von Ertrag",
ylab="Ertrag [dt/ha]", xlab="Roggensorte",
trace.label="Steinmehl")
```

Um die Interpretation zu erleichtern, sollte der Faktor mit der geringeren Anzahl an Abstufungen als letzter gewählt werden, da dieser die Anzahl an separaten Linien bestimmt. Mit 'trace.label' kann die Beschriftung der Legende geändert werden.

Profildigramm von Ertrag



Das Profildigramm zeigt die Mittelwerte über den Ertrag der drei Roggensorten EhoKurz, Kustro und Motto sowohl bei Zugabe von Steinmehl zur Gülle als auch ohne Zugabe von Steinmehl. Die Darstellung zeigt insbesondere, wo die Wechselwirkung zwischen Sorte und Steinmehl anfällt: während bei EhoKurz und Kustro der Ertrag bei Beigabe von Steinmehl höher ist als ohne, ist es bei der Sorte Motto gerade umgekehrt. Im Falle von nichtsignifikanten Wechselwirkungen wäre zu erwarten, dass die Linien im Profildigramm parallel oder zumindest annähernd parallel verlaufen und sich nicht wie in diesem Fall eindeutig überkreuzen.

## Regression

In der Datei *Luzernegrasmischungen.xls* wird der Ertrag von Luzernegrasmischungen (in dt/ha) zusammen mit dem verwendeten Grasanteil (in %) angegeben.

```
> gras <- read.xls( file.choose() )
> gras
  grasanteil ertrag
1          0  116.8
2          5  117.1
3         10  115.4
4         15  118.3
5         20  118.8
6         25  124.6
7         30  120.5
8         35  122.7
.         ..  .....
```

	A	B
1	grasanteil	ertrag
2	0	116.8
3	5	117.1
4	10	115.4
5	15	118.3
6	20	118.8
7	25	124.6
8	30	120.5
9	35	122.7
10		

### Lineare Regression

Mittels linearer Regression kann der Ertrag als lineare Funktion des Grasanteils beschrieben werden. Das Modell lautet:  $\text{Ertrag} = \alpha + \beta * \text{Grasanteil} + \varepsilon$ .

```
> modell <- lm(ertrag ~ grasanteil, data=gras)
> summary(modell)
```

Eine andere Möglichkeit, anzugeben, auf welche Spalten welches DataFrames man sich beziehen will, besteht in der Angabe des DataFrames mit dem Parameter 'data'

```
Call:
lm(formula = gras$ertrag ~ gras$grasanteil)
```

lm ... linear model

```
Residuals:
    Min       1Q   Median       3Q      Max
-2.3321 -1.0786 -0.3179  0.5786  3.7821
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  115.67500    1.29638   89.229 1.33e-10 ***
gras$grasanteil  0.20571    0.06198    3.319  0.0160 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.008 on 6 degrees of freedom
Multiple R-squared:  0.6474,    Adjusted R-squared:  0.5886
F-statistic: 11.02 on 1 and 6 DF,  p-value: 0.01602
```

Aus der Tabelle 'Coefficients' können die Werte für die Regressionsparameter  $\alpha$  und  $\beta$  abgelesen werden.

$\alpha$  (Intercept - Achsenabschnitt) = 115.67500

$\beta$  (gras\$grasanteil) = 0.20571

Das Modell lautet also:  $\text{Ertrag} = 115.67500 + 0.20571 * \text{Grasanteil}$

In diesen beiden Zeilen stehen auch die t-Statistiken und die p-Werte zur jeweiligen Hypothese "Parameter=0":

- $H_\alpha: \alpha = 0$      $p = 1.33 * 10^{-10}$      $< \alpha = 0.05$ , daher wird  $H_\alpha$  abgelehnt:  $\alpha \neq 0$ .
- $H_\beta: \beta = 0$      $p = 0.0160$      $< \alpha = 0.05$ , daher wird  $H_\beta$  abgelehnt:  $\beta \neq 0$ .

'Residual standard error' entspricht der Wurzel aus der Fehlervarianz.  $s_\varepsilon = 2.008$ .

## Streudiagramm (Scatterplot)

Das Streudiagramm dient zur grafischen Beurteilung, wie gut die Modellgerade den Zusammenhang zwischen den beiden Variablen beschreibt.

```
> plot(gras$grasanteil, gras$ertrag, xlab="Grasanteil [%]", ylab="Ertrag [dt/ha]",
main="Erträge von Luzernegrasmischungen")
> abline(modell)
```

Weiters sollen das Konfidenzband und das Vorhersageband eingezeichnet werden. Dazu müssen zuerst zusätzliche x-Werte generiert werden.

```
> hilfswerte <- data.frame(grasanteil=
seq(min(gras$grasanteil),
max(gras$grasanteil), length=100))
```

Die x-Werte müssen unbedingt wieder denselben Variablenamen bekommen wie im Modell.

Für jeden x-Wert werden die zugehörigen y-Werte und Grenzen der Konfidenz- und Vorhersageintervalle berechnet, die zusammen die entsprechenden Kurven ergeben.

```
> konfidenzband <- predict(modell, hilfswerte, interval="conf")
> vorhersageband <- predict(modell, hilfswerte, interval="pred")
```

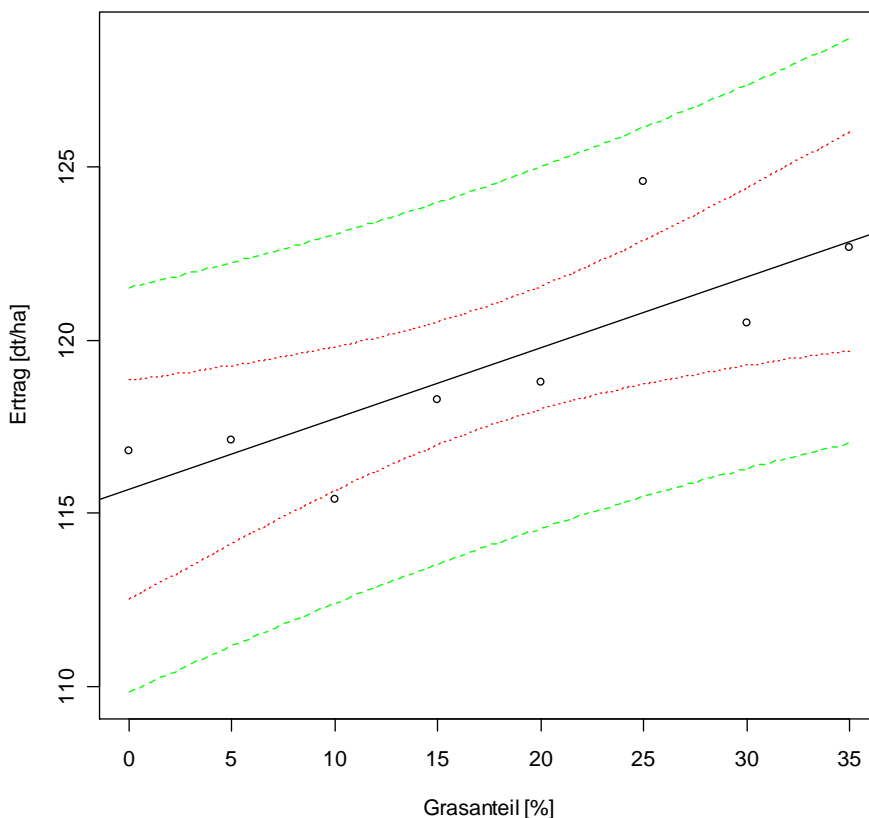
Diese Kurven müssen nun nur noch in das Diagramm eingetragen werden.

```
> plot(gras$grasanteil, gras$ertrag, xlab="Grasanteil [%]", ylab="Ertrag [dt/ha]",
main="Erträge von Luzernegrasmischungen",
ylim=c(min(vorhersageband[,2]),max(vorhersageband[,3])))
> abline(modell)
```

```
> matlines(hilfswerte$grasanteil, konfidenzband,
lty=c("blank", "dotted", "dotted"), col=c("black", "red",
"red"))
> matlines(hilfswerte$grasanteil, vorhersageband,
lty=c("blank", "dashed", "dashed"), col=c("black", "green",
"green"))
```

Ein Auszug aus der Hilfe für die Option Linienart (line type) – lty: Line types can either be specified as an integer (0=blank, 1=solid, 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash) or as one of the character strings "blank", "solid", "dashed", "dotted", "dotdash", "longdash", or "twodash", where "blank" uses 'invisible lines' (i.e., does not draw them).

**Erträge von Luzernegrasmischungen**



## Modellqualität

Die Modellqualität kann mit Hilfe des Bestimmtheitsmaßes (Multiple R-squared in der vorletzten Zeile der Ausgabe) beurteilt werden. Je näher dieser Wert bei 1 liegt, desto besser passt das Modell. Der Wert Multiple R-squared = 0.6474 besagt, dass 64.7% der Variabilität der y-Werte durch das Modell erklärt werden. 0.6474 ist moderat, was auch aus dem Streudiagramm ersichtlich ist: 2 Werte werden nicht gut durch das Modell (Gerade) beschrieben.

In der letzten Zeile der Ausgabe stehen die Ergebnisse des Vergleichs zwei verschieden komplexer Modelle:

$$H_0: y_i = \alpha + \varepsilon \equiv y_i = \alpha + \beta * x + \varepsilon$$

Der dazugehörige p-Wert (0.01602) besagt, dass das komplexere Modell – und damit die lineare Anpassung – besser passt.

## Voraussetzungen

Voraussetzung für die lineare Regression ist, dass die Varianzen der Fehlerterme gleich sind. Aus dem Streudiagramm ist erkennbar, dass die Abweichungen der Beobachtungen von der Modellgeraden (also die Residuen) überall gleich groß sind, sodass kein offensichtlicher Einwand dagegen besteht.

## Konfidenzintervall und Vorhersageintervall

Ein beidseitig beschränktes 95%-iges Konfidenzintervall für den erwarteten Ertrag bei einem Grasanteil von 20% soll berechnet werden.

```
> konfidenzintervall <- data.frame(grasanteil=20)
> predict(modell, int="c", level=0.95, newdata=konfidenzintervall)
      fit      lwr      upr
[1,] 119.7893 118.0110 121.5676
```

Damit erhalten wir als Grenzen eines 95%-igen Konfidenzintervalls für den erwarteten Ertrag bei einem Grasanteil von 20% 118.0 und 121.6 dt/ha.

Um ein nach oben beschränktes 90%-iges Vorhersageintervall für einen (gemessenen) Ertrag bei einem Grasanteil von 25% berechnen zu können, muss  $\alpha$  für das Konfidenzintervall verdoppelt werden.

```
> vorhersageintervall <- data.frame(grasanteil=25)
> predict(modell, int="p", level=0.80, newdata=vorhersageintervall)
      fit      lwr      upr
[1,] 120.8179 117.6788 123.9570
```

Damit erhalten wir als obere Grenze eines nach oben beschränkten 90%-igen Vorhersageintervalls für einen (gemessenen) Ertrag bei einem Grasanteil von 25% 124.0 dt/ha.

## Kreuztabellen (Kontingenztafeln)

In einer Marktuntersuchung einer Konsumentenschutzorganisation wurden 434 Testkäufe von Frischobst getätigt. Unter anderem soll geklärt werden, ob ein Zusammenhang zwischen Verkaufsform (Wochenmarkt, Einzelhändler oder Supermarkt) und Obstqualität (in drei Stufen: gut – mittel – schlecht) besteht. Die Datei *Verkaufsformen.xls* gibt an, wie oft jede Kombination von Verkaufsform und Qualität beobachtet wurde.

```
> verkauf <- read.xls( file.choose() )
> verkauf
      wochenmarkt haendler supermarkt
gut          65          69          30
mittel       27          82          73
schlecht     33          13          42
```

	A	B	C	D
1		wochenmarkt	haendler	supermarkt
2	gut	65	69	30
3	mittel	27	82	73
4	schlecht	33	13	42

Da es sich um zwei Merkmale handelt, können die Daten in einer Kreuztabelle dargestellt werden.

### $\chi^2$ -Test

Mittels  $\chi^2$ -Test kann die Nullhypothese "Verkaufsform und Qualität sind unabhängig" gegen die Alternativhypothese "Verkaufsform und Qualität sind abhängig" getestet werden. Das Risiko 1. Art wird mit  $\alpha = 0.05$  festgelegt.

```
> chisq.test(verkauf)

Pearson's Chi-squared test

data:  verkauf
X-squared = 56.0578, df = 4, p-value = 1.95e-11
```

```
> chisq.test(verkauf)$expected
      wochenmarkt haendler supermarkt
gut          47.23502 61.97235  54.79263
mittel       52.41935 68.77419  60.80645
schlecht     25.34562 33.25346  29.40092
```

In der Tabelle ist die erwartete Häufigkeit angegeben. Der p-Wert der dazugehörigen Teststatistik ( $1.95 \cdot 10^{-11}$ ) liegt unter dem Risiko 1. Art  $\alpha = 0.05$ , weshalb die Nullhypothese abgelehnt werden muss. Es besteht also ein Zusammenhang zwischen Verkaufsform und Qualität.

### Voraussetzungen

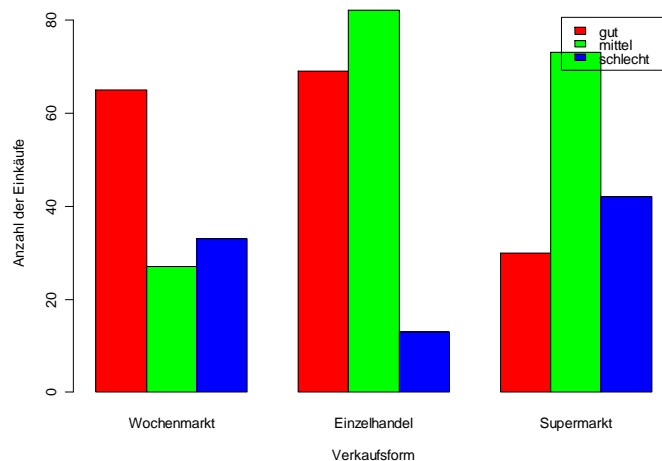
Die erwartete Zellhäufigkeit soll beim Chiquadrat-Test mindesten 5 betragen (alternativ: die minimal erwartete Zellhäufigkeit ist größer als 2 und etwa 50% sind größer als 5); andernfalls sollte man durch Zusammenfassung von Faktorstufen höhere Häufigkeiten erzwingen. Aus der oben angegebenen Tabelle ist ersichtlich, dass die minimal erwartete Zellhäufigkeit 25.3 beträgt, die Voraussetzung ist also erfüllt.

### Balkendiagramm

Der gefunden Zusammenhang kann mittels Balkendiagramm veranschaulicht werden.

```
> barplot(as.matrix(verkauf),
beside=TRUE, legend=TRUE,
xlab="Verkaufsform",
ylab="Anzahl der Einkäufe",
col=c("red", "green", "blue"))
```

Bei Zutreffen der Nullhypothese erwartet man ein gleichartiges Bild der Säulengruppen. Deutliche Abweichungen – wie in diesem Fall – stehen dagegen im Widerspruch zur Nullhypothese. Man erkennt gut, dass beim Händler schlechte Qualität eine geringere Rolle spielt, wohingegen im Supermarkt gute Qualität eindeutig unterrepräsentiert ist.



## Nichtparametrische Verfahren

Zum t-Test und zur Varianzanalyse gibt es auch ein dazugehöriges nichtparametrisches Verfahren. Diese werden dann benutzt, wenn notwendige Voraussetzungen nicht erfüllt sind. Zwei wichtige nichtparametrisches Verfahren sind der Wilcoxon-Rangsummentest und der Kruskal-Wallis-Test.

### Wilcoxon-Rangsummentest

Für die Herdenbucheintragung wurde bei 13 Milchkühen der Rinderrasse "Braunvieh" und bei 11 Kühen der Rasse "Fleckvieh" eine Melkbarkeitsprüfung durchgeführt. Dabei wurde unter anderem die Melkdauer (in min) der Kühe pro Tag ermittelt. Die Ergebnisse sind in der Datei *Melkdauer.xls* angegeben.

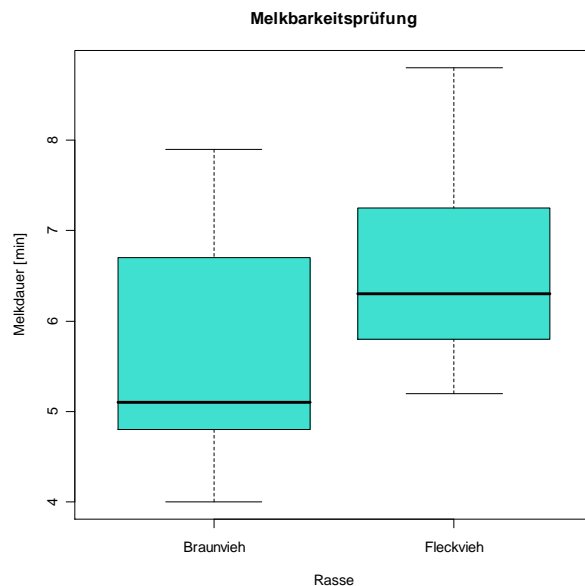
```
> melk <- read.xls( file.choose() )
> melk
      rasse melkdauer
1 Braunvieh      4.9
2 Braunvieh      4.2
3 Braunvieh      4.1
. ...           ...
```

	A	B
1	rasse	melkdauer
2	Braunvieh	4,9
3	Braunvieh	4,2
4	Braunvieh	4,1
5	Braunvieh	7,9
6	Braunvieh	5,7
7	Braunvieh	5,1
8	Braunvieh	5,6
9	Braunvieh	4
10	Braunvieh	7,1
11	Braunvieh	4,8
12	Braunvieh	4,9
13	Braunvieh	7,5
14	Braunvieh	6,7
15	Fleckvieh	6,1
16	Fleckvieh	6,8

### Kastendiagramm (boxplot)

Mittels gruppiertem Kastendiagramm gewinnt man einen Eindruck von den Daten.

```
> boxplot(melk$melkdauer ~ melk$rasse,
          xlab="Rasse", ylab="Melkdauer [min]",
          main="Melkbarkeitsprüfung",
          col="turquoise")
```



### Wilcoxon-Rangsummentest

Die Frage, ob ein Unterschied in der Melkbarkeit je nach Rasse besteht, kann z.B. mit einem Wilcoxon-Rangsummentest beantwortet werden. Die Nullhypothese lautet: "Die Verteilungen der Melkdauer sind für Braunvieh und Fleckvieh gleich." Die Alternativhypothese lautet: "Die Verteilungen der Melkdauer sind für Braunvieh und Fleckvieh ungleich." Das Risiko 1. Art wird mit  $\alpha = 0.05$  festgelegt.

```
> wilcox.test(melk$melkdauer ~ melk$rasse)

Wilcoxon rank sum test with continuity correction

data: melk$melkdauer by melk$rasse
W = 37.5, p-value = 0.05217
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(4.9, 4.2, 4.1, 7.9, 5.7, 5.1, 5.6, :
cannot compute exact p-value with ties
```

Der p-Wert der dazugehörigen Teststatistik (0.05217) liegt über dem Risiko 1. Art  $\alpha = 0.05$ , weshalb die Nullhypothese beibehalten werden muss. Die Verteilung der Melkdauer ist also für beide Rassen gleich.

## Kruskal-Wallis-Test

Um festzustellen, ob das Fach "Statistik" den Studierenden verschiedener Studienrichtungen unterschiedlich schwer fällt, wurden 27 Studierende der drei Studienrichtungen Agrarwissenschaften (AW), Lebensmittel- und Biotechnologie (LBT) und Umwelt- und Bioressourcenmanagement (UBRM) nach ihrer in Statistik erzielten Leistung (in % der erreichbaren Punkte) befragt. Die Ergebnisse sind in der Datei *Statistikergebnisse.xls* zusammengefasst.

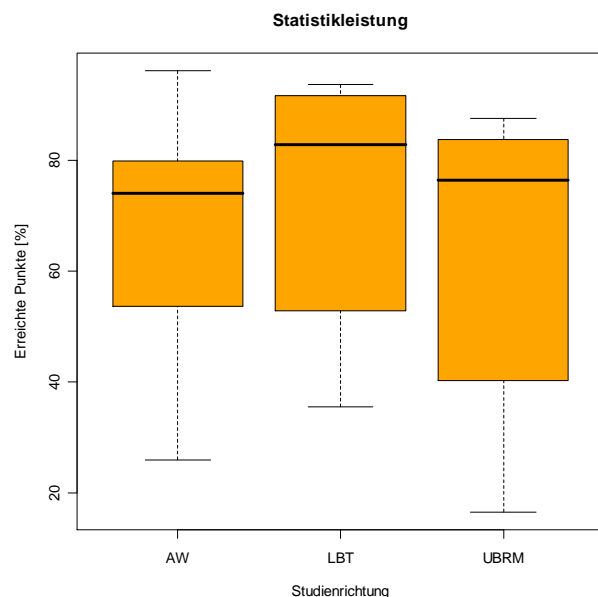
	A	B
1	studium	ergebnis
2	AW	78,3
3	AW	88,2
4	AW	96,2
5	AW	46,3
6	AW	25,9
7	AW	79,9
8	AW	73,2
9	AW	67,2
10	AW	74,8
11	AW	53,6
12	LBT	91,7
13	LBT	82,8
14	LBT	47,5
15	LBT	78,4
16	LBT	92,5
17	LBT	52,9
18	LBT	85,7
19	LBT	93,6
20	LBT	35,5
21	UBRM	83,6
22	UBRM	51,1
23	UBRM	16,5
24	UBRM	79,2
25	UBRM	87,5
26	UBRM	29,5
27	UBRM	73,6
28	UBRM	83,8
29		

```
> leistung <- read.xls( file.choose() )
> leistung
  studium ergebnis
1      AW      78.3
2      AW      88.2
.      ..      ....
10     AW      53.6
11     LBT      91.7
.      ..      ....
19     LBT      35.5
.      ..      ....
27     UBRM     83.8
```

### Kastendiagramm (boxplot)

Mit Hilfe eines gruppierten Kastendiagramms kann man die Statistikleistung der Studierenden anschaulich darstellen.

```
> boxplot(leistung$ergebnis ~
leistung$studium,
xlab="Studienrichtung", ylab="Erreichte
Punkte [%]", main="Statistikleistung",
col="orange")
```



### Kruskal-Wallis-Test

Die Frage, ob ein Zusammenhang zwischen Studienrichtung und Statistikleistung besteht, kann z.B. mittels Kruskal-Wallis-Test beantwortet werden. Die dazugehörige Nullhypothese lautet: "Die Verteilungen der Statistikleistung sind für alle drei Studienrichtungen gleich." Die Alternativhypothese lautet: "Mindestens zwei Verteilungen sind unterschiedlich." Das Risiko 1. Art wird mit  $\alpha = 0.05$  festgelegt.

```
> kruskal.test(leistung$ergebnis ~ leistung$studium)
```

Kruskal-Wallis rank sum test

```
data: leistung$ergebnis by leistung$studium
Kruskal-Wallis chi-squared = 1.1921, df = 2, p-value = 0.551
```

Der p-Wert der dazugehörigen Teststatistik (0.551) liegt über dem Risiko 1. Art  $\alpha = 0.05$ , weshalb die Nullhypothese beibehalten werden muss. Die Verteilung der Statistikleistung ist also für alle drei Studienrichtungen gleich.