Welche Düngung wirkt am besten?

Auswertung eines Düngeversuchs mit Pflanzen: Varianzanalyse, Tukey-Test und Boxplot mit R

von Norbert Kessel

Manchmal muss man herausfinden, welche Art von Düngung das stärkste Wachstum bei Pflanzen bewirkt. Meist wird die Wirkung an Feldfrüchten untersucht, aber auch Forstpflanzen reagieren auf eine Zugabe von Dünger, der meist in mineralischer Form ausgebracht wird. Die Wirkung zeigt sich unter anderem in stärkerem Höhenwachstum, deshalb misst man die Höhen von Pflanzen und vergleicht die Werte miteinander. Das ist wichtig, denn je schneller eine junge Pflanze dem Äser des überall vorhandenen Rehwilds entwachsen ist, umso sicherer ist das Überleben.

Im folgenden Beispiel soll herausgefunden werden, ob eine Düngung einen *statistisch signifikanten Einfluss* auf das Höhenwachstum von jungen Buchen hat. Hierzu wurde eine kleine Versuchsanlage aufgebaut, in der die Wirkung von vier verschiedenen Arten der Düngung an jeweils 5 Pflanzen getestet wurde.

Die Zahl der Pflanzen ist in diesem Fall sicher sehr klein, aber für eine exemplarische Darstellung reicht es durchaus. In einer richtigen Versuchsanlage würde man für jede Variante nicht 5, sondern 50 Pflanzen untersuchen. Auch der Zeitraum, in dem gedüngt wird, ist wichtig, so wäre es interessant zu wissen, wieviel Jahre man düngen muss, um die Terminalknospe so hoch wachsen zu lassen, dass das Rehwild keinen Schaden mehr anrichten kann (bei Revieren, in denen nie Schnee liegt, wäre das ca. 1,30 m, in Revieren mit Schnee muss die maximale Schneehöhe im Winter dazu addiert werden).

Die folgende Tabelle fasst die Höhen der Buchen nach Ablauf von zwei Jahren zusammen. Insgesamt waren es vier Varianten (V1-V4). Eine Variante blieb ungedüngt, die anderen drei Varianten erhielten eine Gabe von Phosphor (P), Phosphor und Calcium (P+Ca) sowie Stickstoff, Phosphor und Calcium (N+P+Ca).

Variante	V1	V2	V3	V4
Düngung	ohne	P	P+Ca	N+P+Ca
	0,7	0,7	1,0	1,0
	0,8	1,4	1,3	1,6
	0,9	1,3	1,4	1,4
	1,0	1,4	1,6	1,6
	1,3	1,2	1,7	2,0

Die Höhe von Pflanzen (in m) nach Ablauf von zwei Jahren. V1 ist die unbehandelte Kontrollfläche, daneben sind die drei gedüngten Varianten.

Die Daten stammen aus einem Biometrie-Praktikum der Universität Freiburg (Institut für Forstliche Biometrie, mit freundlicher Genehmigung von Prof. Dr. D. Pelz).

Ein erster Blick auf die Mittelwerte

Ein erster Blick bei solchen Versuchen gilt immer den arithmetischen Mittelwerten, denn manchmal liefern diese schon Hinweise. Die Mittelwerte können mit dem Taschenrechner oder auch mit der Tabellenkalkulation (LibreOffice, OpenOffice, MS-Excel) berechnet werden.

Variante	V1	V2	V3	V4
arithmetische Mittelwerte	0,94	1,2	1,4	1,52

Der kleinste Mittelwert findet sich bei der Variante V1, der größte bei der Variante V4. Interessanterweise ist dieser Mittelwert um rund 50 % höher als der bei V1. Die anderen beiden Mittelwerte liegen dazwischen.

Die Frage ist nun, ob die Unterschiede zwischen den Varianten so groß sind, dass man von einem *signifikanten* Unterschied sprechen kann. Dass dabei die Streuung der Werte innerhalb der vier Gruppen eine große Rolle spielt, liegt auf der Hand: ist die Streuung sehr groß, dann liegt kein nachweisbarer Effekt vor – die Unterschiede wären dann zufällig entstanden und ließen sich nicht auf die Düngung zurückführen.

Auswertung mit einer Varianzanalyse und dem Tukey-Test

Liegen mehrere arithmetische Mittelwerte vor, dann werden sie oft mit der sogenannten Varianzanalyse ausgewertet, ein Rechenverfahren, das es seit mehr als 100 Jahren gibt. Der Englische Name lautet 'Analysis of Variance' (abgekürzt AOV). Wie der Name schon sagt, werden die Varianzen (Streuungen) analysiert. Im Deutschen verwendet man auch den Begriff Streuungszerlegung.

Der in einem anderen Aufsatz beschriebene t-Test macht etwas ähnliches, vergleicht aber nur *zwei* arithmetische Mittelwerte miteinander.

Die Varianzanalyse kann durchaus mit einem Taschenrechner gerechnet werden, allerdings muss man sehr sorgfältig sein und die einzelnen Schritte penibel auf einem Blatt notieren. Gibt man im Taschenrechner eine falsche Zahl ein, ist dieser Fehler kaum zu rekonstruieren. Verwendet man stattdessen eine Tabellenkalkulation zur Speicherung der Daten, ist es sicherer, da man die einmal erfassten Daten speichern, ausdrucken und kontrollieren kann.

Eine Varianzanalyse kann nur ermitteln, *ob* es Unterschiede zwischen den Varianten gibt. Um herauszufinden, welche Variante(n) sich signifikant von anderen unterscheidet, muss ein zweiter Test verwendet werden (hier: Tukey).

Datenspeicherung mit einer Tabellenkalkulation

Um die Daten in R zu verarbeiten, müssen sie zuerst aus einer Datei gelesen und in den Hauptspeicher des Computers kopiert werden. Hier wird dazu eine Datei verwendet, die mit einer Tabellenkalkulation erstellt wurde, gängige Programme dazu sind: LibreOffice, OpenOffice und MS-Excel.

Die folgende Abbildung zeigt den Inhalt der Datei mit den Daten, die in zwei Spalten angeordnet sind. In der ersten Spalte sind die Bezeichnungen der Varianten (V1-V4), in der zweiten Spalte die 20 Messwerte, die bereits oben vorgestellt wurden. Die Datei hat hier den Namen 'AOV_Daten.xls'.

	A	B
1	V1	0,7
2	V1	0,8
3	V1	0,9
4	V1	1
5	V1	1,3
6	V2	0,7
7	V2	1,4
8	V2	1,3
9	V2	1,4
10	V2	1,2
11	V3	1
12	V3	1,3
13	V3	1,4
14	V3	1,6
15	V3	1,7
16	V4	1
17	V4	1,6
18	V4	1,4
19	V4	1,6
20	V4	2

Die Daten für die Varianzanalyse (Ausschnitt)

Verarbeitung der Daten in R

Die zum Lesen der Daten nötige Funktion read.xlsx() ist nicht Bestandteil der Standard-Installation von R. Sie gehört zu einem Paket mit dem Namen 'XLSX' und muss (einmal) zusätzlich aus dem Internet von einem Server heruntergeladen werden.

Installation und Laden des zusätzlichen Pakets in R

Mit der folgenden Anweisung, in R ausgeführt, installiert man das Paket 'XLSX' auf dem Rechner:

```
install.packages("xlsx")
```

Man muss angeben, von welchem Server das Paket heruntergeladen werden soll, nach den guten Erfahrungen mit der Universität Münster wird diese Quelle für den Download empfohlen. Sie kann aus der angezeigten Liste ausgewählt werden.

Nachdem das Paket heruntergeladen ist, muss es noch in den Hauptspeicher geladen werden, hierzu verwendet man die Funktion:

```
library("xlsx")
```

Danach kann die Funktion 'read.xlsx()' zum Lesen von Daten aus einer Datei verwendet werden.

Aufruf der Funktion

Mit der folgenden Anweisung liest man die Daten und speichert sie in eine Variable:

```
AOV_Daten <- read.xlsx("/Users/norbertkessel/Desktop/Buecher2019/KesselStatistik/AOV_Daten.xls",
sheetName="Sheet1", header=FALSE)
```

Zur Erläuterung:

- AOV_Daten ist der Name einer Variablen (er ist frei wählbar)
- die Funktion read.xlsx() liest die Daten aus einer Datei, dabei muss der Pfad zu der Datei sowie der Dateiname angegeben werden
- Sheet1 ist der Name des Arbeitsblattes in der Datei
- header=FALSE bedeutet, dass es in der Datei keine Spaltenüberschriften gibt.

Anschließend kann man sich die Daten anzeigen lassen, indem man den Namen der Variablen eingibt und die Enter-Taste drückt:

```
AOV_Daten
```

führt zur Anzeige der folgenden Liste:

```
> AOV_Daten
   X1  X2
1  V1 0.7
2  V1 0.8
3  V1 0.9
4  V1 1.0
5  V1 1.3
6  V2 0.7
7  V2 1.4
8  V2 1.3
9  V2 1.4
10 V2 1.2
11 V3 1.0
12 V3 1.3
13 V3 1.4
14 V3 1.6
15 V3 1.7
16 V4 1.0
17 V4 1.6
18 V4 1.4
19 V4 1.6
20 V4 2.0
```

Die fortlaufenen Zahlen in der ersten Spalte dienen zur Nummerierung der 20 Datensätze. In der zweiten Spalte stehen die Abkürzungen der Varianten, in der dritten Spalte die Höhen-der Pflanzen.

Auswertung der Daten

Nun können die Daten ausgewertet werden. Die dazu nötige Funktion hat den Namen aov(). Es gibt verschiedene Arten, die Funktion aov() aufzurufen. Weit verbreitet ist die folgende, die alle ermittelten Werte in einer weiteren Variablen speichert, die dann mit der Funktion summary() ausgewertet wird.

In diesem Beispiel bekommt diese zusätzliche Variable den Namen 'varianzanalyse':

```
varianzanalyse <-aov(X2 ~X1, data=AOV_Daten)
```

Zur Erläuterung:

- X2 und X1 sind die beiden Variablen, dabei ist X2 (Höhe) die sogenannte abhängige Variable und X1 die unabhängige
- zwischen den beiden Variablen muss die sogenannte Tilde (~)stehen

Zur Auswertung wird die Funktion summary() folgendermaßen aufgerufen:

```
summary(varianzanalyse)
```

Sie liefert die folgende zusammenfassende Übersicht:

```
             Df Sum Sq Mean Sq F value Pr(>F)
X1            3 0.9655  0.3218   3.731 0.0331 *
Residuals    16 1.3800  0.0862
```

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

Zur Erläuterung: Die wichtigste Information ist ‚Pr(>F)‘: dort steht die Zahl **0,0331.** Liegt dieser Wert unter 0,05 (so wie hier), dann gibt es *signifikante* Unterschiede zwischen den Gruppen, was mit einem Asterisk symbolisiert wird.

Anmerkungen zu den anderen Informationen in der Übersicht:

- Df ist die Abkürzung für 'degrees of freedom', die Freiheitsgrade, die bei der manuellen Auswertung unten benötigt werden
- Sum Sq steht für 'sum of squares', die Summe der quadrierten Abweichungen
- Mean Sq sind die mittleren Abweichungsquadrate, die berechnet werden, indem man Sum Sq durch Df dividiert, auch diese Informationen werden für die manuelle Berechnung der Varianzanalyse unten benötigt.
- F-Value ist der errechnete Prüfwert, der bei manueller Auswertung mit einem theoretischen Tabellenwert verglichen wird (s.u.).

Welche Gruppen unterscheiden sich signifikant?

Die Varianananalyse hat ergeben, *dass* es Unterschiede in der Wirkung der untersuchten Düngungsvarianten gibt. Aber *welche* Variante unterscheidet sich von den anderen? Um das herauszufinden wird der sogenannte Tukey-Test gerechnet. Er vergleicht jede Variante mit jeder anderen.

Den Test kann man folgendermaßen aufrufen:

```
TukeyHSD(aov(X2 ~X1, data=AOV_Daten))
```

In den Klammern hinter der Funktion TukeyHSD() steht im Grunde das, was bei der Funktion aov() auch zu schreiben war: die Namen der Variablen, von einer Tilde getrennt, und der Name der Datenquelle.

Hier sind die daraufhin angezeigten Informationen:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = X2 ~ X1, data = AOV_Daten)
$X1
      diff       lwr       upr     p adj
V2-V1 0.26 -0.27141085 0.7914108 0.5173015
V3-V1 0.46 -0.07141085 0.9914108 0.1022772
V4-V1 0.58  0.04858915 1.1114108 0.0300540
V3-V2 0.20 -0.33141085 0.7314108 0.7080727
V4-V2 0.32 -0.21141085 0.8514108 0.3446966
V4-V3 0.12 -0.41141085 0.6514108 0.9153864
```

Wichtig sind die Werte in der letzten Spalte: wenn sie unter 0,05 liegen, gibt es signifikante Unterschiede. Hier ist das nur im Vergleich der Varianten V4 mit V1 der Fall, die anderen Varianten unterscheiden sich nicht signifikant.

Schlussfolgerung

Die beiden Varianten V1 und V4 unterscheiden sich signifikant voneinander. Das wurde bereits nach dem ersten Blick auf die arithmetischen Mittelwerte der beiden Gruppen vermutet.

Zum Vergleich: Varianzanalyse von Hand

Hier soll gezeigt werden, wie man früher eine Varianzanalyse mit dem Taschenrechner gerechnet hat. Hierzu werden zuerst die arithmetischen Mittelwerte jeder Variante (Gruppe) berechnet. Danach werden drei Arten von Streuungen ermittelt: 1. Innerhalb der Gruppen, 2. zwischen den Gruppenmittelwerten und 3. die Gesamtstreuung. Schließlich wird der F-Wert berechnet, der mit einem Tabellenwert verglichen wird.

Hier die Übersicht der vier Varianten mit den Summen und den arithmetischen Mittelwerten.

Variante	V1	V2	V3	V4	
Düngung	ohne	P	P+Ca	N+P+Ca	
	0,7	0,7	1,0	1,0	
	0,8	1,4	1,3	1,6	
	0,9	1,3	1,4	1,4	
	1,0	1,4	1,6	1,6	
	1,3	1,2	1,7	2,0	
Summen	4,7	6,0	7,0	7,6	25,3
Mittelwerte	0,94	1,2	1,4	1,52	1,265

In der letzten Spalte finden sich die Gesamtsumme aller Werte und der Gesamtmittelwert.

Berechnung der Varianzen

1. Varianz innerhalb der Gruppen ($SQ_{innerhalb}$)

Für jede der vier Gruppen werden die Differenzen der Einzelwerte zum Mittelwert berechnet, quadriert und summiert.

Für die erste Datenreihe (V1) ergibt sich die folgende Rechenanweisung:

$V1: (0,7-0,94)^2 + (0,8-0,94)^2 + (0,9-0,94)^2 + (1,0-0,94)^2 + (1,3-0,94)^2$
$= 0,0576 + 0,0196 + 0,0016 + 0,0056 + 0,1296$
$= 0,214$

Für die zweite Datenreihe (V2):

$V2: (0,7-1,2)^2 + (1,4-1,2)^2 + (1,3-1,2)^2 + (1,4-1,2) + (1,2-1,2)^2$
$= 0,25 + 0,04 + 0,01 + 0,04 + 0$
$= 0,34$

Für die dritte Datenreihe (V3):

$V3: (1,0-1,4)^2 + (1,3-1,4)^2 + (1,4-1,4)^2 + (1,6-1,4)^2 + (1,7-1,4)^2$
$= 0,16 + 0,01 + 0 + 0,04 + 0,09$
$= 0,3$

Für die vierte Datenreihe (V4):

$V4: (1,0-1,52)^2 + (1,6-1,52)^2 + (1,4-1,52)^2 + (1,6-1,52)^2 + (2,0-1,52)^2$
$= 0,27 + 0,006 + 0,014 + 0,006 + 0,230$
$= 0,526$

In der folgenden Übersicht sind die Varianzen der vier Varianten zusammengefasst.

Variante	V1	V2	V3	V4	Summe
Varianz	0,214	0,34	0,3	0,526	1,38

Addiert man die vier Werte erhält man die Summe aller quadrierten Abweichungen (s. letzte Spalte):

$0,214 + 0,34 + 0,3 + 0,526 = 1,38$

man bezeichnet diesen Wert auch als

$SQ_{innerhalb} = 1,38$

2. Varianz zwischen den Gruppenmittelwerten

Man berechnet die Differenzen der vier Gruppenmittelwerte zu dem Gesamtmittelwert (1,265), quadriert und summiert sie:

$((0,94 - 1,265)^2 + (1,2 - 1,265)^2 + (1,4 - 1,265)^2 + (1,52 - 1,265)^2)*5)$

= (0,105 + 0,004 + 0,018 + 0,065) * 5
= 0,192 * 5
= 0,960

$SQ_{zwischen}$ = 0,960

3. Gesamt-Varianz

Zur Berechnung der Gesamt-Varianz benötigt man noch die Summen der quadrierten Einzelwerte ($\sum x^2$), für die Variante 1 ergibt sich die folgende Rechenanweisung:

$= 0,7^2 + 0,8^2 + 0,9^2 + 1,0^2 + 1,3^2$

$= 0,49+0,64+0,81+1+1,69$

$= 4,63$

Hier eine Zusammenfassung der vier Varianten mit der Gesamtsumme:

Variante	V1	V2	V3	V4	Gesamt
$\sum x^2$	4,63	7,54	10,1	12,08	34,35

Die Rechenanweisung zur Ermittlung der Gesamtvarianz sieht folgendermaßen aus:

$$\text{Summe aller quadrierten Werte} - \frac{(\text{Summe der Werte})^2}{\text{Anzahl der Werte}}$$

$$= 34,35 - \frac{25,3^2}{20}$$

$$= 2,3455$$

Somit ergibt sich für die Gesamtvarianz ein Wert von

SQ_{gesamt} = 2,3455

Kontrolle

Zur Kontrolle kann man die Gesamtstreuung berechnen, indem man die Werte von $SQ_{innerhalb}$ und $SQ_{zwischen}$ addiert:

$$SQ_{zwischen} + SQ_{innerhalb} = SQ_{gesamt}$$

$$0,960 + 1,38 = 2,34$$

Der Wert von SQ_{gesamt} ist am einfachsten zu berechnen, deshalb berechnet man man $SQ_{innerhalb}$ oft aus (SQ_{gesamt} - $SQ_{zwischen}$).

Berechnung des F-Wertes

Nachdem die Varianzen berechnet sind, wird mit den gerade ermittelten Varianzen der sogenannte F-Wert berechnet. Dieser wird anschließend mit einem Tabellenwert verglichen. Ist der berechnete Wert größer als der Tabellenwert, dann wird die sogenannte H_0-Hypthese auf Gleichheit verworfen. Das würde dann bedeuten, dass es signifikante Unterschiede gibt.

Neben den Varianzen werden die sogenannten Freiheitsgrade gebraucht. Die folgende Tabelle fasst die Werte zusammen:

Streuung	Summe der Quadrate SQ	Freiheitsgrade FG	Mittlere Quadrate MQ = SQ/FG
zwischen	0,9655	3	0,3218
innerhalb	1,38	16	0,08625
gesamt	2,345	19	0,123

Erläuterung:

- in der zweiten Spalte finden sich die zuvor berechneten Streuungswerte
- in der dritten Spalte sind die Freiheitsgrade, sie berechnen sich folgendermaßen:
  - zwischen = (Anzahl der Gruppen -1) = 4 - 1 = 3
  - innerhalb = (Anzahl der Messwerte - 4) = 20-4 =16
  - gesamt = (Anzahl der Messwerte - 1) = 20 - 1 = 19
- in der vierten Spalte sind die MQ-Werte, sie werden berechnet, indem man die SQ-Werte durch die Zahl der Freiheitsgrade dividiert

Auch bei der Zahl der Freiheitsgrade gilt, dass man die Werte überprüfen kann, indem man rechnet:

$FG_{zwischen} + FG_{innerhalb} = FG_{gesamt}$

Schranken der t-, χ^2- und F-Verteilung für P = 0,95 (α = 0,05)

FG ν	t einseitig	t zweiseitig	χ^2	F obere Schranken 1	2	3	4	5	6	7	8	9	10	12	15
1	6,31	12,71	3,84	161	200	216	225	230	234	237	239	241	242	244	246
2	2,92	4,30	5,99	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,39	19,40	19,41	19,43
3	2,35	3,18	7,81	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70
4	2,13	2,78	9,49	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86
5	2,02	2,57	11,07	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62
6	1,94	2,45	12,59	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94
7	1,89	2,36	14,07	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51
8	1,86	2,31	15,51	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22
9	1,83	2,26	16,92	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01
10	1,81	2,23	18,31	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85
11	1,80	2,20	19,68	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72
12	1,78	2,18	21,03	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62
13	1,77	2,16	22,36	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53
14	1,76	2,14	23,68	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46
15	1,75	2,13	25,00	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40
16	1,75	2,12	26,30	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35
17	1,74	2,11	27,59	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31
18	1,73	2,10	28,87	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27
19	1,73	2,09	30,14	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23
20	1,72	2,09	31,41	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20
21	1,72	2,08	32,67	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18
22	1,72	2,07	33,92	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15
23	1,71	2,07	35,17	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13

Die Schranken der F-Verteilung.
Hervorgehoben ist der Wert für 3,16 Freiheitsgrade, er beträgt: 3,24
Quelle: Sachs, L. Angewandte Statistik, Springer Verlag, 6. Auflage (Ausschnitt)
Die vollständige Tabelle findet sich auch im Internet:
https://link.springer.com/content/pdf/bfm%3A978-3-662-21613-2%2F1.pdf
(Suche nach: Schranken der F-Verteilung)

$3 + 16 = 19$

Die genannten Freiheitsgrade finden sich auch in der von R berechneten Übersicht, dort sind sie mit 'df' bezeichnet (degrees of freedom).

Der berechnete F-Wert

$F_{ber.} = MQ_{zwischen} / MQ_{innerhalb}$

$F_{ber.} = 0{,}3218 / 0{,}08625$

$F_{ber.} = 3{,}73$

Dieser Wert muss mit dem F-Wert aus der Tabelle verglichen werden, der abhängig von den Freiheitsgraden abgelesen wird. Die Pfeile markieren den Einstieg in die Tabelle: zuerst oben mit 3, dann links mit 16 abzulesen.

Der F-Wert aus der Tabelle (s. unten)

$F_{Tab\ (3,\ 16)} = 3{,}24$

Fazit

Der berechnete Wert (3,73) ist größer als der Tabellenwert (3,24). Somit liegen signifikante Unterschiede zwischen den Gruppenmittelwerten vor.

Boxplot

Graphische Darstellung von Daten mit einem Boxplot

Die graphische Darstellung von Zahlen ist oft hilfreich, wenn man sich einen ersten Eindruck verschaffen möchte. Oft handelt es sich um Messwerte, von denen man wissen möchte, ob es eine Häufung um den Mittelwert gibt und ob Ausreißer vorhanden sind.

Sind nur wenige Daten vorhanden, lässt sich das einfach überprüfen, aber bei mehreren hundert Messwerten oder verschiedenen Datenreihen, wird es kompliziert. Hier kommt der Boxplot ins Spiel, der die Daten anschaulich beschreibt. Hierzu teilt er die Daten auf, ermittelt den Median und sucht nach eventuellen Ausreißern.

Ein erstes Beispiel mit wenigen Daten

Mit den folgenden Zahlen soll ein erster Boxplot erstellt werden: 3, 5, 6, 7, 9.

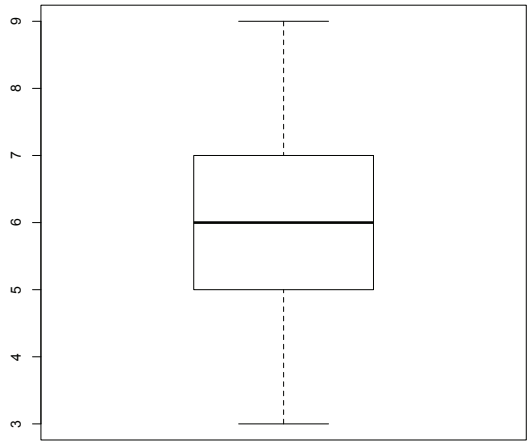
Um mit den Daten in R zu arbeiten, müssen sie zunächst in eine Variable kopiert werden (der Name der Variablen ist frei wählbar, Groß- und Kleinschreibung ist zu beachten):

```
daten <- c(3, 5, 6, 7, 9)
```

Zunächst soll ein Boxplot mit diesen Daten angefertigt werden, im Anschluss daran wird erläutert, was er anzeigt. Die folgende Anweisung

```
boxplot(daten)
```

erzeugt diesen Boxplot:



Zur Erläuterung:

- der Kasten (die Box) in der Mitte umfasst 50 % der Daten
- die waagerechte dicke Linie in der Box steht für den Median, das ist der Wert, der die Datenmenge halbiert, hier ist er bei 6. Die eine Hälfte der Daten liegt oberhalb, die andere Hälfte der Daten unterhalb dieser Linie
- die nach oben und unten geführten Linien stehen für das untere und obere Quartil der Daten

Warum Median und nicht arithmetischer Mittelwert?

Der arithmetische Mittelwert ist uns vertrauter, aber er wird im Boxplot nicht angezeigt, warum ist das so? Der arithmetische Mittelwert ist gegenüber Ausreißern empfindlich, der Median dagegen zeigt sich von Ausreißern unbeeindruckt. Zur Veranschaulichung ändert man die zuvor verwendeten Daten ab (man kann die alte Anweisung mit den Cursor-Tasten anzeigen lassen und modifizieren):

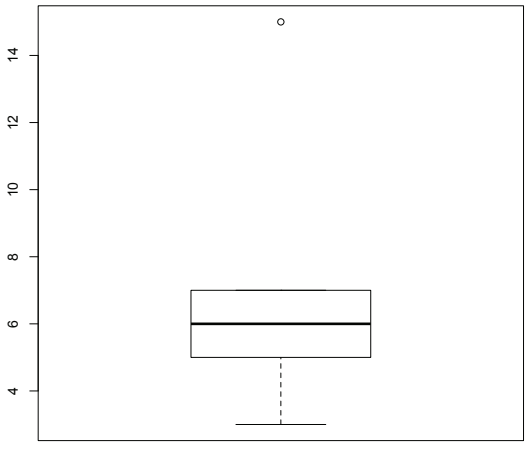
```
daten <- c(3, 5, 6, 7, 15)
```

Die 9 wurde durch 15 ersetzt.

Lässt man nun den Boxplot anzeigen,

```
boxplot(daten)
```

ergibt sich folgendes Bild



Zur Erläuterung:

- oben erscheint ein kleiner Kreis, er steht für einen vermuteten Ausreißer
- der Median ist unverändert 6.

Zum Vergleich kann man den arithmetischen Mittelwert leicht mit der folgenden Anweisung ermitteln:

```
mean(daten)
```

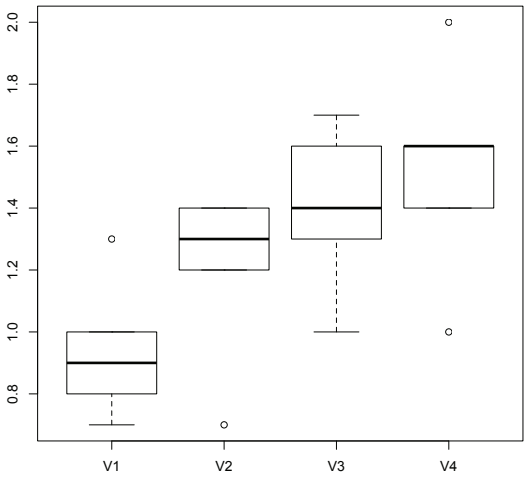
Für die erste Datenreihe (3, 5, 6, 7, 9) ergibt sich ein Mittelwert von 6, für die zweiten Datenreihe (3, 5, 6, 7, 15) ein Mittelwert von 7,2.

Ein Beispiel mit den Daten aus der Varianzanalyse

Bei der oben vorgestellten Varianzanalyse wurden von R Namen für die Variablen festgelegt (X1 bzw. X2).

Mit der folgenden Anweisung lassen sich diese Daten grafisch darstellen:

```
boxplot(X2 ~X1, data=AOV_Daten)
```



Zur Erläuterung:

- Der Median der Variante V1 ist deutlich kleiner als die anderen drei Mediane
- die Variante V4 hat nach oben und unten Ausreißer.