

# ÜBUNGSAUFGABEN

## DATENANALYSE MIT R

### 1. ÜBUNGSAUFGABE

Lernen Sie R und den Umgang mit Variablen und Fällen anhand von Merkmalsausprägungen kennen

### 2. ÜBUNGSAUFGABE

Analysieren Sie die technischen Spezifikationen von Automobilen mittels uni-, bi- und multivariater Statistik

### 3. ÜBUNGSAUFGABE

Untersuchen Sie den Einfluss unterschiedlicher Behandlungsmethoden mittels Chi-Quadrat-Test

### 4. ÜBUNGSAUFGABE

Vergleichen Sie die Wirksamkeit von Orangensaft und Vitamin C mittels t-Test

### 5. ÜBUNGSAUFGABE

Mittels Faktorenanalyse identifizieren Sie die Big Five als Persönlichkeitsstrukturen

### 6. ÜBUNGSAUFGABE

Ein intuitiver Einstieg in das maschinelle Lernen mit Einblicken in ausgewählte Algorithmen



# 1. ÜBUNGSAUFGABE

## Einleitung

Über den Befehl ***iris*** können Sie in der R Konsole auf den IRIS Datensatz zugreifen. Die Bezeichnung ist auf die gleichnamige griechische Göttin des Regenbogens zurückzuführen, nach der ebenfalls die Pflanzengattung der Schwertlilien benannt ist. Im IRIS Datensatz finden Sie insgesamt 150 Beobachtungen mit jeweils vier Merkmalsvariablen von Schwertlilien. Der Datensatz enthält Informationen über die Breite und die Länge des Kelchblatts (Sepalum) sowie des Kronblatts (Petalum) in Zentimeter. Darüber hinaus bezieht sich der Datensatz auf drei verschiedene Arten der Schwertlilie (Setosa, Versicolor und Virginica), welche sich jeweils in Breite und Länge des Kelch- und Kronblatts unterscheiden. Ziel der Übungsaufgabe soll es sein, anhand der vier Merkmalsvariablen eine manuelle Identifikation der richtigen Schwertlilienart in der R Konsole zu ermöglichen.

	Kelchblatt		Kronblatt		
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	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	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

## Univariate Statistik (Teil 1)

Zunächst sollen der IRIS Datensatz, die darin enthaltenen Variablen und deren Merkmalsausprägungen möglichst genau beschrieben werden. Dazu können Sie auf den Befehl **summary(iris)** in der R Konsole zurückgreifen:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

(1) Vergleichen Sie die Längen von Kelch- und Kronblatt. Welches Blatt ist länger?

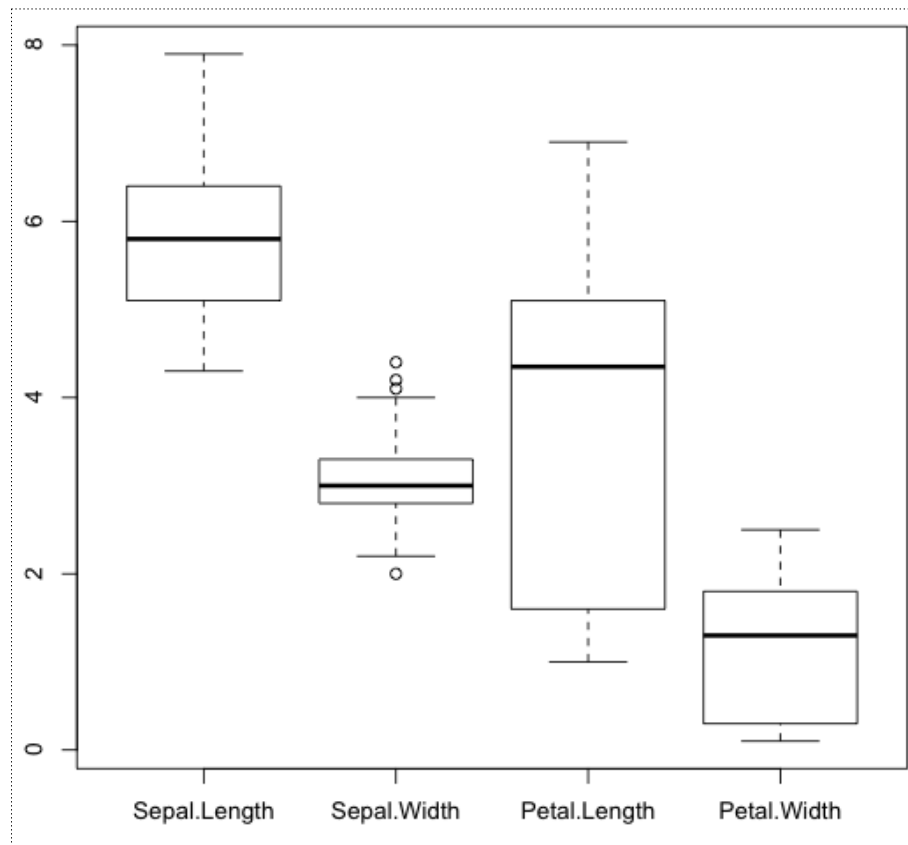


(2) Wie viele Fälle der verschiedenen Schwertlilienarten liegen jeweils vor?



## Univariate Statistik (Teil 2)

Die vier Merkmalsvariablen zur Breite und Länge von Kelch- und Kronblatt unterscheiden sich jeweils in ihrem Mittelwert als auch in ihrer minimalen und maximalen Merkmalsausprägung deutlich voneinander. Diese Unterschiede können mit dem Befehl **`boxplot(iris[1:4])`** in der R Konsole visualisiert werden. Dieser Befehl visualisiert nicht alle Variablen des IRIS Datensatzes, sondern lediglich die ersten vier Merkmalsvariablen für alle Schwertlilienarten zusammen:



**Hinweis:** Dass die Merkmalsvariablen unterschiedlich ausgeprägt sind, ist für eine korrekte Identifikation der verschiedenen Schwertlilienarten ausschlaggebend. Ein detaillierter Einblick in die unterschiedlichen Schwertlilienarten ist über diese Darstellung jedoch noch nicht möglich.

## Univariate Statistik (Teil 3)

Weil für die Identifikation der verschiedenen Schwertlilienarten ein gutes Verständnis des zugrundeliegenden Datensatzes unerlässlich ist, folgt noch eine weitere deskriptive Analyse. Diese soll verdeutlichen, dass die verschiedenen Schwertlilienarten jeweils eine unterschiedliche Breite und Länge von Kelch- und Kronblatt aufweisen. Hierzu müssen mit den Befehlen **`setosa <- subset(iris, Species=="setosa")`** sowie **`versicolor <- subset(iris, Species=="versicolor")`** und **`virginica <- subset(iris, Species=="virginica")`** geeignete Teildatensätze zu den jeweiligen Schwertlilienarten in der R Konsole erstellt werden.

Nun lassen sich über die erstellten Teildatensätze und in Verbindung mit den einzelnen Befehlen **`summary(setosa[c(1:4)])`** sowie **`summary(versicolor[c(1:4)])`** und **`summary`**

**(*virginica*[*c*(1:4)])** die deskriptiven Kennzahlen der einzelnen Schwertlilienarten in der R Konsole aufrufen. Diese Art der Darstellung und die dazugehörige Interpretation der deskriptiven Kennzahlen kennen Sie bereits aus Teil 2 der vorliegenden Übungsaufgabe.

(3) Welche Schwertlilienart hat die längsten Kelch- und Kronblätter?



(4) Welche Schwertlilienart hat das schmalste Kronblatt?

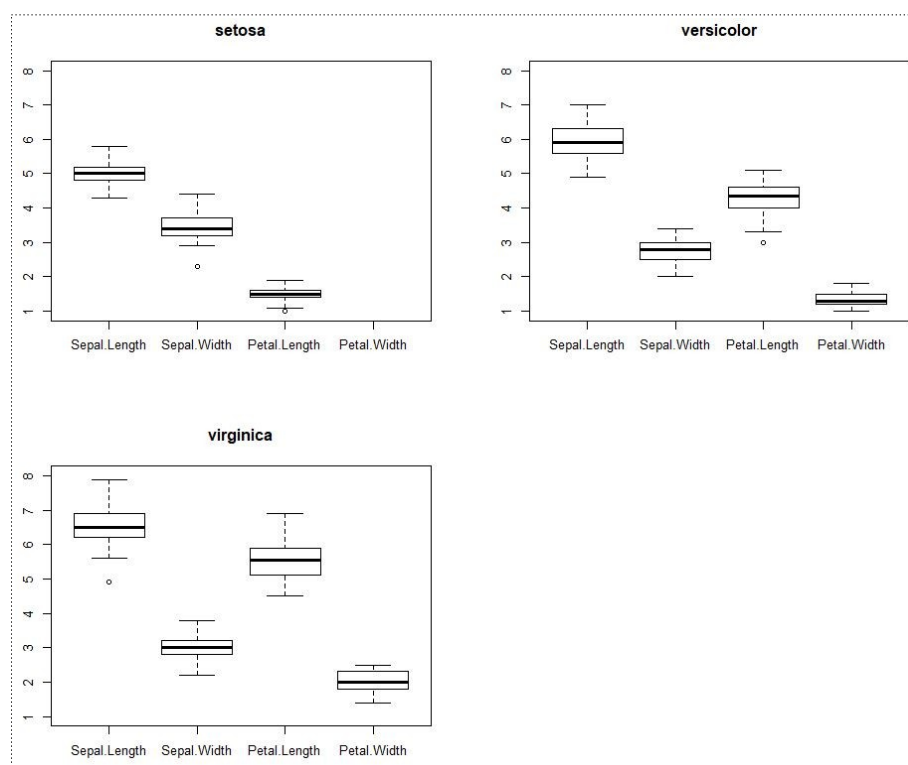


#### Univariate Statistik (Teil 4)

Die deskriptiven Kennzahlen aus Teil IV der vorliegenden Übungsaufgabe können natürlich auch wieder über eine Boxplot grafisch dargestellt werden. Hierzu können für die Schwertlilienarten Setosa, Versicolor und Virginica einzelne Boxplots erstellt werden, die über den etwas umfassenderen Befehl in der R Konsole zusammengefasst werden können:

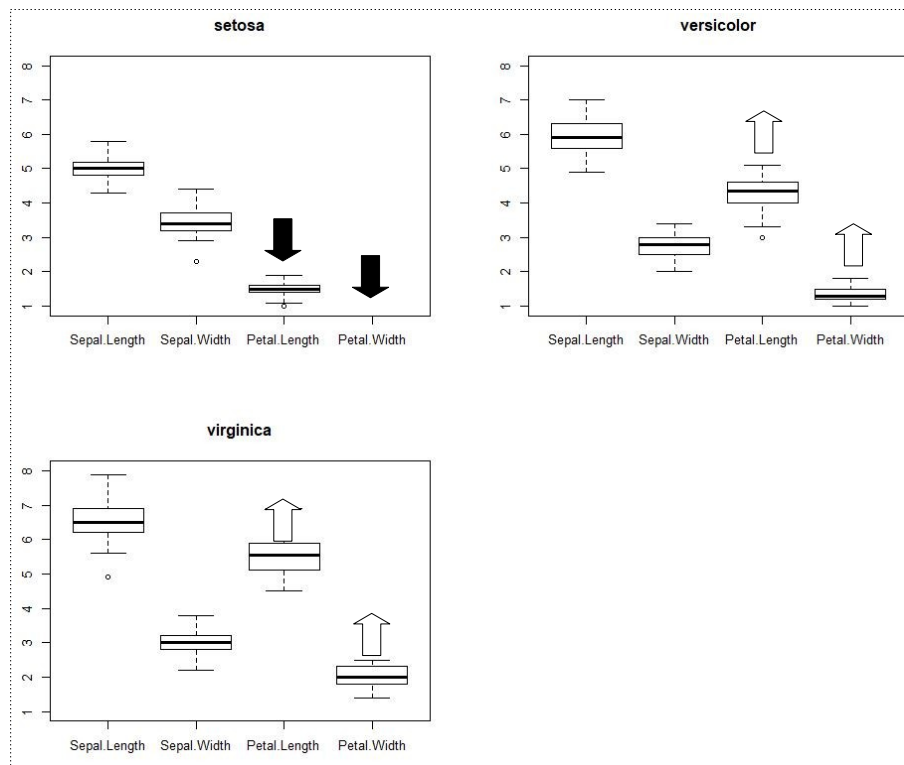
```
par(mfrow=c(2,2))  
boxplot(setosa[c(1:4)], main="setosa", ylim=c(1,8))  
boxplot(versicolor[c(1:4)], main="versicolor", ylim=c(1,8))  
boxplot(virginica[c(1:4)], main="virginica", ylim=c(1,8))
```

Der par-Befehl fasst die Boxplots in einer zwei Reihen und zwei Spalten großen Grafik zusammen. Über den main-Befehl erhält jede Boxplot eine Überschrift und der ylim-Befehl definiert die anzuzeigenden Werte der Y-Achse:



## Mustererkennung (Teil 1)

Nun geht es darum die unterschiedlichen Muster in den Teildatensätzen zu den Schwertlilienarten Setosa, Versicolor und Virginica zu erkennen. Die grafische Darstellung aus Teil 4 der vorliegenden Übungsaufgabe stellt die zugrundeliegenden Muster grafisch gegenüber, so dass diese auch augenscheinlich erkannt werden können. Beispielsweise handelt es sich bei Setosa um die einzige Schwertlilienart mit einem Kronblatt, dessen Breite kleiner dem Wert 1 ist. Dieses Muster unterscheidet Setosa deutlich von Versicolor und Virginica. Auch ist die Länge des Kronblattes bei Setosa mit Werten zwischen 1 und 2 deutlich kleiner ausgeprägt als bei Versicolor (Werte zwischen 3 und 5) und Virginica (Werte zwischen 4,5 und 7).



## Mustererkennung (Teil 2)

Diese erkennbaren Unterschiede lassen sich nutzen, um in der R Konsole die Bedingungen für einen Teildatensatz zu programmieren, der ausschließlich aus Setosa Schwertlilien besteht. Wenn die Bedingungen nicht präzise genug programmiert sind, werden dem Teildatensatz entweder nicht alle Setosa Schwertlilien zugeordnet, oder aber Versicolor und Virginica werden dem Teildatensatz fälschlicherweise zugeordnet. Ein möglicher Befehl in der R Konsole, der die im vorherigen Teil der vorliegenden Übungsaufgabe erkannten Muster von Setosa in entsprechende Bedingungen überführt, lautet: **`identification <- subset(iris, Petal.Length>="1" & Petal.Length<="2" & Petal.Width <="1", select=c(Species))`**. Abschließend kann über den Befehl `summary(identification)` in der R Konsole überprüft werden, wie viele von den insgesamt 50 Setosa Schwertlilien erkannt wurden und ob möglicherweise Falschzuordnungen vorlagen.

	Species
setosa	: 50
versicolor	: 0
virginica	: 0

Mit dem oben genannten Befehl konnten in der R Konsole demnach 50 von 50 Setosa Schwertlilien korrekt identifiziert werden. Dies entspricht einer Erkennungsrate von 100 Prozent.

### **Mustererkennung (Teil 3)**

Jetzt sind Sie an der Reihe. Nutzen Sie die Ihnen bekannten Muster der Schwertlilienarten und die Befehle für "größer gleich ( $\geq$ )" und "kleiner gleich ( $\leq$ )" auf die entsprechenden Variablen in der R Konsole an, um die jeweils 50 ausstehenden Schwertlilien für Versicolor und Virginica korrekt zu identifizieren.

(5) Beschreiben Sie das Muster der Versicolor Schwertlilien und definieren Sie über den oben gezeigten subset-Befehl entsprechende Bedingungen zur Identifikation in der R Konsole:



(6) Beschreiben Sie das Muster der Versicolor Schwertlilien und definieren Sie über den oben gezeigten subset-Befehl entsprechende Bedingungen zur Identifikation in der R Konsole:



### **Lösungen**

- (1) Das Kelchblatt ist stets (min., max. und mean) länger als das Kronblatt.
- (2) Im gesamten Datensatz ist jede Schwerlilienart 50-mal vertreten.
- (3) Virginica.
- (4) Setosa.
- (5) Hier ist Kreativität gefragt :-)
- (6) Hier ist Kreativität gefragt :-)

*The End*

## 2. ÜBUNGSAUFGABE

### Einleitung

In R finden Sie den MTCARS Datensatz, dessen Abkürzung für Motor Trend Cars steht und auf Daten aus dem Magazin "US Motor Trend" aus dem Jahr 1974 basiert. Dargestellt werden 32 verschiedene Automobile, u.a. der "Mazda RX4" oder ein "Lincoln Continental" und jeweils zehn dazugehörige Merkmalsvariablen dieser Automobile. Zu den Merkmalsvariablen zählen u.a. "cyl" für Cylinders und "wt" für Weight. So kann bspw. ein Ausschnitt aus dem MTCARS Datensatz in R aussehen, wenn man diesen in der R Konsole über den Befehl **mtcars** aufruft:

	Cylinders					Weight					
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1

In den einzelnen Spalten finden Sie die Merkmalsvariablen und jede Zeile entspricht einem Automobil. Tipp: Diese Anordnung an Variablen innerhalb einer Tabelle ist generell für viele statistischen Analysevorhaben zu empfehlen. Mit der vorliegenden Übungsaufgabe werden Sie Schritt für Schritt durch den Prozess der Datenanalyse mit R geführt, um Ihnen einen Einblick in die vielfältigen Anwendungsfelder der Statistik zu geben.

### Univariate Statistik (Teil 1)

Zunächst werden wir den MTCARS Datensatz deskriptiv analysieren. Dabei geht es primär um eine Beschreibung der vorliegenden Variablen und deren Merkmalsausprägungen. Hierbei lässt sich nicht nur jedes einzelne Automobil entlang der oben dargestellten Tabelle beschreiben, sondern auch Detailfragen in Bezug auf mehrere Automobile beantworten. Eine Frage könnte zum Beispiel sein, wie viele Zylinder (Cylinders) die Automobile im Jahr 1974 durchschnittlich hatten. Über den Befehl **summary(mtcars)** lässt sich in der R Konsole eine erste systematische Zusammenfassung aller Merkmalsvariablen aufrufen, hier exemplarisch dargestellt für die Cylinders:

		cyl	
Minimumwert	→	Min. :4.000	
		1st Qu.:4.000	
		Median :6.000	
		Mean :6.188	← Durchschnittswert
		3rd Qu.:8.000	
Maximumwert	→	Max. :8.000	

Zusätzlich werden im vorliegenden Beispiel noch das erste und dritte Quartil sowie der Median ausgewiesen. Diese statistischen Kennwerte werden im Rahmen des Seminars / der Vorlesung näher erläutert und dienen u.a. der Erstellung einer grafischen Boxplot, welche Sie für einzelne Merkmalsvariablen in der R Konsole bspw. über den Befehl **`boxplot(mtcars$cyl)`** aufrufen können.

## **Univariate Statistik (Teil 2)**

(1) Wie ist das durchschnittliche Gewicht der Automobile im Jahr 1974?



-----

(2) Wie viele Gänge hatten die Automobile im Jahr 1974 maximal?



-----

(3) Was war der niedrigste Benzinverbrauch der Automobile im Jahr 1974?



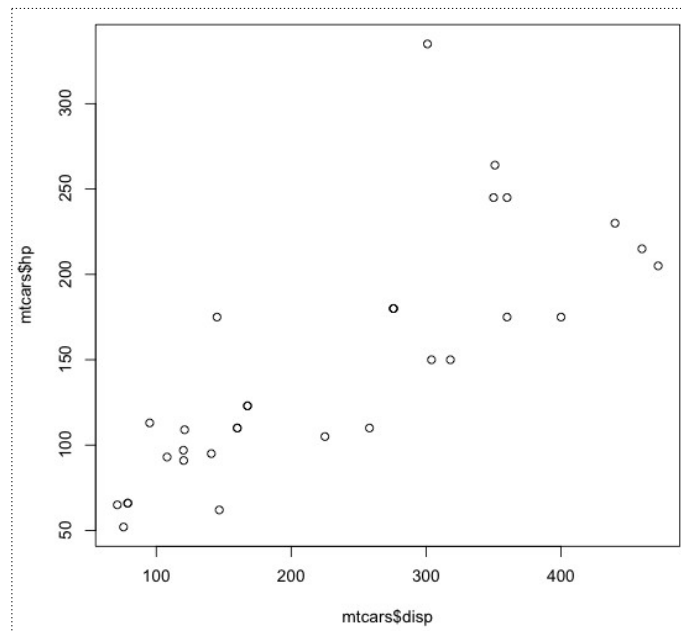
-----

Hinweis: Wenn Sie erklärende Details zu den einzelnen Merkmalsvariablen des MTCARS Datensatzes aufrufen möchten, können Sie den Befehl **`help(mtcars)`** in die R Konsole eingeben.

## **Bivariate Statistik (Teil 1)**

Als nächstes beschäftigen wir uns mit den kausalen Zusammenhängen zwischen den einzelnen Merkmalsvariablen. Ein sehr bekannter und physikalisch gut nachvollziehbarer Zusammenhang besteht bspw. zwischen dem Hubraum (Displacement) und dem PS Kennwert (Horsepower) eines Automobils. Hierzu schauen wir uns zunächst das Zusammenspiel beider Merkmalsvariablen über die Befehle **`plot(mtcars$disp, mtcars$hp)`** bzw. **`plot(mtcars$hp~mtcars$disp)`** in der R Konsole an:

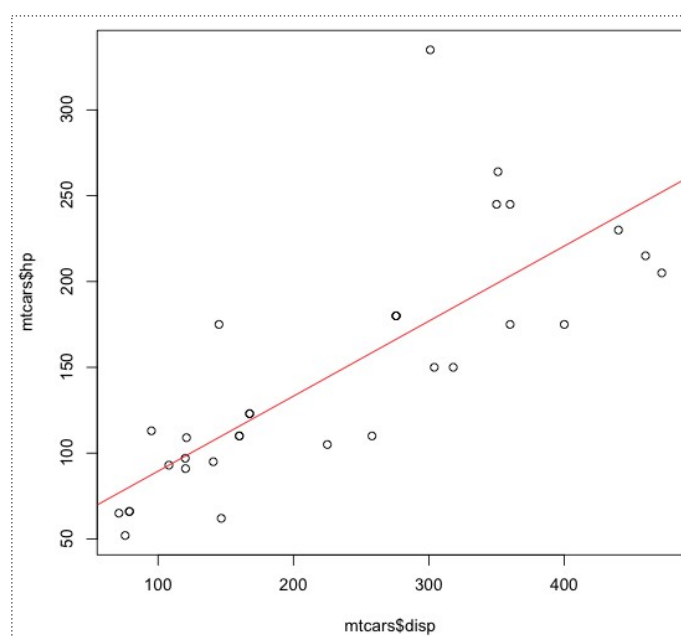




Hinweis: Über den Befehl ***plot(mtcars)*** können Sie innerhalb der R Konsole auch eine grafische Darstellung aller Zusammenhänge im MTCARS Datensatz aufrufen.

## Bivariate Statistik (Teil 2)

Bei der vorherigen Grafik zur bivariaten Analyse des Hubraums und dem PS Kennwert eines Automobils scheint zunächst ein positiver Zusammenhang vorzuliegen: Mit steigendem Hubraum nimmt auch der PS Kennwert zu. Diesen positiven Zusammenhang können wir mittels einer linearen Funktion verdeutlichen, die sich aus dem positiven Zusammenhang zwischen Hubraum und PS Kennwert ergibt. Der vorherigen Grafik lässt sich diese lineare Funktion in der R Konsole über den Befehl ***abline(lm(mtcars\$hp~mtcars\$disp), col="red")*** hinzufügen:



### Bivariate Statistik (Teil 3)

Die genaue Stärke des positiven Zusammenhangs ist u.a. definiert über die Steigung der linearen Funktion. Bei zwei Merkmalsausprägungen lässt sich die Steigung über eine lineare Regression bestimmen, welche im Seminar / in der Vorlesung näher erläutert wird. In Bezug auf die obigen Grafiken geht es insbesondere um die Linearität des Zusammenhangs, welche mittels Korrelation exakt ermittelt werden kann. Um den sogenannten Korrelationskoeffizienten aufzurufen, bedient man sich in der R Konsole des Befehls **`cor(mtcars$disp, mtcars$hp)`**. Ausgewiesen wird bspw. der exakte Wert der Korrelation zwischen dem Hubraum und dem PS Kennwert eines Automobils:

( - 1.00)	<b>perfekt negativ</b>	
(+ 1.00)	<b>perfekt positiv</b>	→ [1] 0.7909486
(± 0.00)	<b>kein Zusammenhang</b>	

Im vorliegenden Beispiel beträgt der exakte Wert der Korrelation + 0.79, dies bedeutet, dass ein starker linearer Zusammenhang vorliegt.

### Bivariate Statistik (Teil 4)

Wenn Sie alle Zusammenhänge zwischen den Merkmalsvariablen auf einmal aufrufen möchten, so können Sie dies in der R Konsole über den Befehl **`cor(mtcars)`** erreichen. Hier sehen Sie einen Auszug der somit erzeugten Korrelationstabelle, in der jede Korrelation einer Merkmalsvariablen mit sich selbst einem perfekt positiven Zusammenhang von + 1.00 entspricht. Alle weiteren positiven und negativen Zusammenhänge lassen sich entsprechend der Korrelationstabelle entnehmen:

	mpg	cyl	disp	hp
mpg	1.0000000	-0.8521620	-0.8475514	-0.7761684
cyl	-0.8521620	1.0000000	0.9020329	0.8324475
disp	-0.8475514	0.9020329	1.0000000	0.7909486
hp	-0.7761684	0.8324475	0.7909486	1.0000000
drat	0.6811719	-0.6999381	-0.7102139	-0.4487591
wt	-0.8676594	0.7824958	0.8879799	0.6587479
qsec	0.4186840	-0.5912421	-0.4336979	-0.7082234
vs	0.6640389	-0.8108118	-0.7104159	-0.7230967
am	0.5998324	-0.5226070	-0.5912270	-0.2432043
gear	0.4802848	-0.4926866	-0.5555692	-0.1257043
carb	-0.5509251	0.5269883	0.3949769	0.7498125

Hinweis: Entlang der einzelnen Korrelationen der Merkmalsvariablen mit sich selbst spiegeln sich die anderen Korrelationen. Manche Werte sind also doppelt in der Korrelationstabelle vertreten und müssen lediglich einmalig betrachtet werden.

## Bivariate Statistik (Teil 5)

(4) Bestimmen Sie den exakten Zusammenhang zwischen dem Benzinverbrauch (Miles per Gallon) eines Automobils und dem Gewicht (Weight):



(5) Bestimmen Sie den exakten Zusammengang zwischen der Anzahl an Gängen (Gear) eines Automobils und der Schaltungsart (Automatic / Manual):



(6) Rufen Sie in der R Konsole eine Grafik auf, die den Zusammenhang zwischen der Beschleunigung ( $\frac{1}{4}$  Mile Time) eines Automobils und dessen Hubraum (Displacement) darstellen kann und interpretieren Sie diese Grafik.



## Multivariate Statistik (Teil 1)

Neben der bivariaten Analyse von jeweils zwei Merkmalsvariablen ermöglicht die multivariate Analyse die gleichzeitige Berücksichtigung von mehr als zwei Merkmalsvariablen zur Bestimmung eines Zusammenhangs. Wenn bspw. die Fragen geklärt werden sollen, welche Merkmalsvariablen den stärksten Einfluss auf den Benzinverbrauch (Miles per Gallon) eines Automobils haben und wie stark diese Einflüsse ausgeprägt sind, wenn sie nicht (wie zuvor in der bivariaten Analyse) alleine auf diesen einwirken, empfiehlt sich ein Regressionsmodell. Ein Regressionsmodell, welches schrittweise und automatisch die Merkmalsvariablen mit dem stärksten Einfluss bestimmt, kann in der R Konsole über den Befehl **`step(lm(data = mtcars, mpg ~ .), trace=0)`** generiert werden:

```
Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Coefficients:
(Intercept)          wt          qsec          am
      9.618       -3.917        1.226        2.936
```

Demnach hat das Gewicht (Weight) eines Automobils einen negativen Einfluss, dessen Beschleunigung ( $\frac{1}{4}$  Mile Time) einen positiven Einfluss und die Schaltungsart (Automatic / Manual) ebenfalls einen positiven Einfluss auf den Benzinverbrauch (Miles per Gallon).

## Multivariate Statistik (Teil 2)

Neben der Bestimmung der Merkmalsvariablen mit dem stärksten Einfluss lassen sich mit einem Regressionsmodell noch weitere Informationen gewinnen. Hierzu muss der zuvor ausgeführte Befehl in der R Konsole modifiziert werden: **`regression <- step(lm(data = mtcars, mpg ~ .), trace=0)`**. Dadurch wird die Regression in R als neues Objekt angelegt und kann über den bereits bekannten Befehl **`summary(regression)`** weiter analysiert werden:

```
Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Residuals:
    Min       1Q   Median       3Q      Max
-3.4811 -1.5555 -0.7257  1.4110  4.6610

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6178     6.9596   1.382 0.177915
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
qsec          1.2259     0.2887   4.247 0.000216 ***
am            2.9358     1.4109   2.081 0.046716 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11
```

Hier interessieren insbesondere die Verteilung der Residuen (Residuals), der Determinationskoeffizient (Multiple R-squared) sowie der Standardfehler (Std. Error), wie sie im Seminar / in der Vorlesung erläutert werden.

## Multivariate Statistik (Teil 3)

(7) Welche unabhängigen Variablen tragen entsprechend in einem schrittweisen Regressionsmodell zur Beschleunigung (¼ Mile Time) bei?



(8) Welche unabhängige Variable hat den stärksten Einfluss auf die Beschleunigung?



## Lösungen

- (1) 3.217 Tonnen.
- (2) 5.000 Gänge.
- (3) 33.90 Miles per Gallon.
- (4) - 0.867 (negativer Zusammenhang).
- (5) + 0.794 (positiver Zusammenhang).
- (6) `plot(mtcars$disp, mtcars$qsec)` i.V.m. `abline(lm(mtcars$qsec~mtcars$disp))`.
- (7) `step(lm(data=mtcars, qsec~.), trace=0)`.
- (8) Weight mit + 1.475.

*The End*

### 3. ÜBUNGSAUFGABE

#### Einleitung

In Ihrem Verantwortungsbereich liegen zwei Patientengruppen (A und B) mit identischem Krankheitsbild, von denen die Patientengruppe A nach einer neuen Methode und die Patientengruppe B mit der konventionellen Methode behandelt wird. In jeder Patientengruppe liegt die Anzahl (beobachtete Werte) der erfolgreichen und nicht erfolgreichen Behandlungen für jeweils 100 Patientinnen und Patienten vor (siehe Tabelle):

	Patientengruppe A (neue Methode)	Patientengruppe B (konventionelle Methode)
Behandlung erfolgreich	70	55
Behandlung nicht erfolgreich	30	45

Ihre Alternativhypothese soll lauten: Mit der neuen Methode werden mehr Patientinnen und Patienten erfolgreich behandelt als mit der konventionellen Methode.

**Hinweis:** In R können Sie diese Datenmatrix über die Befehle ***erfolgreich <- cbind(70,55)*** in Verbindung mit ***nicht\_erfolgreich <- cbind(30, 45)*** und schließlich ***matrix <- rbind(erfolgreich, nicht\_erfolgreich)*** anlegen. Mit dem Befehl ***matrix*** können Sie sich diese Datenmatrix in R entsprechend anzeigen lassen.

#### Hypothesenaufstellung

(1) Ausgehend von der oben genannten Hypothese lautet die Nullhypothese:



-----

(2) Beim Chi-Quadrat-Test wird die Nullhypothese überprüft. Was muss dann gelten?



-----

### Erwartete Werte berechnen

(3) Wie lauten die erwarteten Werte, wenn die Nullhypothese gilt? Berücksichtigen Sie dabei, dass in der Ausgangssituation insgesamt 63 % (125 / 200) erfolgreich und 37 % (75 / 200) nicht erfolgreich behandelt werden:

	Patientengruppe A (neue Methode)	Patientengruppe B (konventionelle Methode)
Behandlung erfolgreich	➡	➡
Behandlung nicht erfolgreich	➡	➡

### Chi-Quadrat-Wert berechnen

(4) Ausgehend von den beobachteten und den erwarteten Werten können Sie Chi-Quadrat zur Überprüfung Ihrer Nullhypothese ermitteln. Benutzen Sie hierfür die folgende Formel für beobachtete Werte und erwartete Werte:

$$\chi^2 = \sum \frac{(\text{beobachteter} - \text{erwarteter Wert})^2}{\text{erwarteter Wert}}$$



### Abgleich mit Chi-Quadrat-Verteilungstabelle

(5) Wenn der von Ihnen ermittelte Wert größer ist als der Referenzwert in der Chi-Quadrat-Tabelle, hier ausgehend von einem zu 95 % zutreffenden Ergebnis und einem Freiheitsgrad (df) von 1, können Sie die Nullhypothese ablehnen.

Freiheitsgrade	1 - α					
	00,85	00,90	00,95	00,975	00,99	00,995
1	02,07	02,71	03,84	05,02	06,63	07,88
2	03,79	04,61	05,99	07,38	09,21	10,60
3	05,32	06,25	07,81	09,35	11,34	12,84
4	06,74	07,78	09,49	11,14	13,28	14,86
5	08,12	09,24	11,07	12,83	15,09	16,75
(...)	(...)	(...)	(...)	(...)	(...)	(...)



Nullhypothese abgelehnt [ ]      Nullhypothese nicht abgelehnt [ ]

Hinweis: Über **chisq.test(matrix)** kann man in R den exakten p-Wert ermitteln!

## Lösungen

- (1) Mit der neuen Methode werden nicht mehr Patientinnen und Patienten erfolgreich behandelt als mit der konventionellen Methode.
- (2) Dann sollte zwischen Patientengruppe A und Patientengruppe B kein Unterschied im Behandlungserfolg vorliegen.
- (3) A (erfolg.): 63; A (nicht erfolg.): 37; B (erfolg.): 63; B (nicht erfolg.): 37.
- (4)  $(70-63)^2/63 + (55-63)^2/63 + (30-37)^2/37 + (45-37)^2/37 = 4,8$ .
- (5) Nullhypothese abgelehnt.

*The End*



## 4. ÜBUNGSAUFGABE

### Einleitung

In dieser Übungsaufgabe geht es um Meerschweinchen - im TOOTHGROWTH Datensatz finden Sie insgesamt drei Variablen, diese beziehen sich auf die Länge der Zähne (Tooth Length) von insgesamt 60 Meerschweinchen, die Art eines Nahrungsergänzungsmittels (Supplement Type) und die tatsächlich verabreichte Dosis des Nahrungsergänzungsmittels (Dose in mg/day). Den TOOTHGROWTH Datensatz rufen Sie in der R Konsole ganz klassisch über den Befehl ***ToothGrowth*** auf:

	Tooth Length		Dose in mg/day
	len	supp	dose
1	4.2	VC	0.5
2	11.5	VC	0.5
3	7.3	VC	0.5
4	5.8	VC	0.5
5	6.4	VC	0.5
6	10.0	VC	0.5
7	11.2	VC	0.5
8	11.2	VC	0.5
9	5.2	VC	0.5
10	7.0	VC	0.5

Verabreicht wurde den Meerschweinchen entweder pures Vitamin C, im Datensatz als VC ausgewiesen, oder natürlicher Orangensaft, im Datensatz als OJ (Orange Juice) ausgewiesen. Mittels statistischer Analysen soll nun die Frage beantwortet werden, welches Nahrungsergänzungsmittel besser für das Wachstum der Zähne von Meerschweinchen ist.

### Univariate Statistik (Teil 1)

Zunächst werden wir den TOOTHGROWTH Datensatz mit den bereits bekannten deskriptiven Analyseverfahren im Detail betrachten. Über den Befehl ***summary*** (***ToothGrowth***) lässt sich in der R Konsole wieder eine erste systematische Zusammenfassung aller Merkmalsvariablen aufrufen, hier exemplarisch dargestellt für die Tooth Length:

	len	
Minimale Länge →	Min. : 4.20	
	1st Qu.:13.07	
	Median :19.25	
	Mean :18.81	← Durchschnittliche Länge
	3rd Qu.:25.27	
Maximale Länge →	Max. :33.90	

Zusätzlich werden im vorliegenden Beispiel erneut das erste und dritte Quartil sowie der Median ausgewiesen, welche u.a. für die grafische Erstellung einer Boxplot notwendig sind. Diese lässt sich für die Tooth Length über den Befehl **`boxplot(ToothGrowth$len)`** aufrufen. Beachten Sie bei dem vorliegenden Beispiel, dass einzelne Variablen in einem Datensatz gezielt über den Befehl `$` adressiert werden können. Über den aus anderen Analysen bereits bekannten Befehl **`boxplot(ToothGrowth)`** erzeugen Sie hingegen Boxplots für alle im Datensatz enthaltenen Variablen. Dies ist für die Variablen Supplement Type und Dose in mg/day aufgrund der Merkmalsausprägungen allerdings nicht sinnvoll.

## Univariate Statistik (Teil 2)

(1) Wie vielen Meerschweinchen wurde Orangensaft verabreicht?



-----

(2) Welche Dosis wurde den Meerschweinchen durchschnittlich verabreicht?



-----

(3) Welche Interpretationsprobleme ergeben sich in Anbetracht der Variablen Supplement Type und Dose in mg/day, wenn Sie von der durchschnittlichen Dosis sprechen?

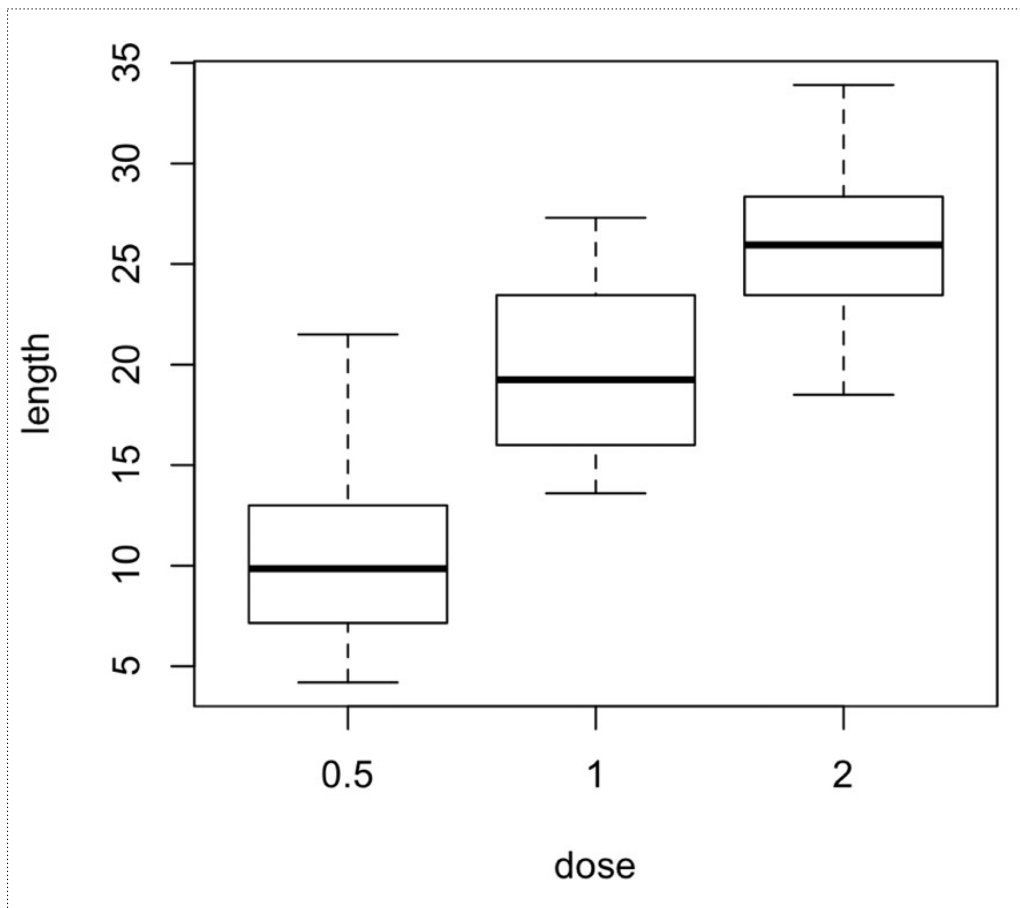


-----

**Hinweis:** Wenn Sie erklärende Details zu den einzelnen Merkmalsvariablen des Datensatzes aufrufen möchten, können Sie den Befehl **`help(ToothGrowth)`** in die R Konsole eingeben.

## Bivariate Statistik (Teil 1)

Wir haben bereits gesehen, dass die allgemeinen Boxplots für den vorliegenden Datensatz aufgrund der einzelnen Merkmalsausprägungen der Variablen Supplement Type und Dose in mg/day nicht sinnvoll sind. Stattdessen lässt sich in der R Konsole über ergänzende Parameter der Befehl **`boxplot(ToothGrowth$len ~ ToothGrowth$dose, xlab = "dose", ylab = "length")`** eingeben, der eine besondere Boxplot erzeugt, welche die Länge der Zähne der Meerschweinchen in Abhängigkeit von der verabreichten Dosis abbildet. Darüber hinaus können Labels für die X-Achse sowie die Y-Achse, wie im vorliegenden Befehl dargestellt, frei vergeben werden:



## Subsets interpretieren (Teil 1)

Die Meerschweinchen erhalten entweder Vitamin C oder Orangensaft als Nahrungsergänzungsmittel. Jeweils 30 von insgesamt 60 Meerschweinchen erhalten eines der beiden Nahrungsergänzungsmittel. Demnach sollten eigentlich zwei Datensätze vorliegen; Einer für die Meerschweinchen die Vitamin C verabreicht bekommen und einer für die Meerschweinchen die Orangensaft verabreicht bekommen. Dazu modellieren wir den vorliegenden Datensatz um und erzeugen über den Befehl **`vc <- subset(ToothGrowth, supp=="VC")`** ein Subsample für die Vitamin C Meerschweinchen und über den Befehl **`oj <- subset(ToothGrowth, supp=="OJ")`** ein Subsample für die Orangensaft Meerschweinchen. Die Subsamples werden als einzelne Objekte in R angelegt, die über die Befehle **`vc`** oder **`oj`** aufgerufen werden können. Eine deskriptive Zusammenfassung des Subsamples der Vitamin C Meerschweinchen kann bspw. über den Befehl **`summary(vc[c(1,3)])`** aufgerufen werden, wobei nun die Variable zum Supplement Type (in R Konsole: **`vc[c(2)]`**) ausgelassen werden kann:

**summary(vc[c(1)])**

**summary(vc[c(3)])**

len		dose	
Min.	: 4.20	Min.	:0.500
1st Qu.	:11.20	1st Qu.	:0.500
Median	:16.50	Median	:1.000
Mean	:16.96	Mean	:1.167
3rd Qu.	:23.10	3rd Qu.	:2.000
Max.	:33.90	Max.	:2.000

### **Subsets interpretieren (Teil 2)**

(4) Welche Zahnlänge haben die Vitamin C Meerschweinchen durchschnittlich?



(5) Bei welchem Nahrungsergänzungsmittel tritt die längste Zahnlänge auf?



Die vorliegenden deskriptiven Befunde müssen nun statistisch validiert werden. Zum einen lassen sich hierfür weiterführende grafische Analysen aufrufen und zum anderen kann überprüft werden, ob es sich bei dem festgestellten Unterschied zwischen den arithmetischen Mitteln (MEAN) der Vitamin C Meerschweinchen und der Orangensaft Meerschweinchen um einen statistisch signifikanten Unterschied handelt, oder ob im vorliegenden Datensatz lediglich eine zufällige Differenz enthalten ist, die keine Rückschlüsse auf die Art des verwendeten Nahrungsergänzungsmittels zulässt.

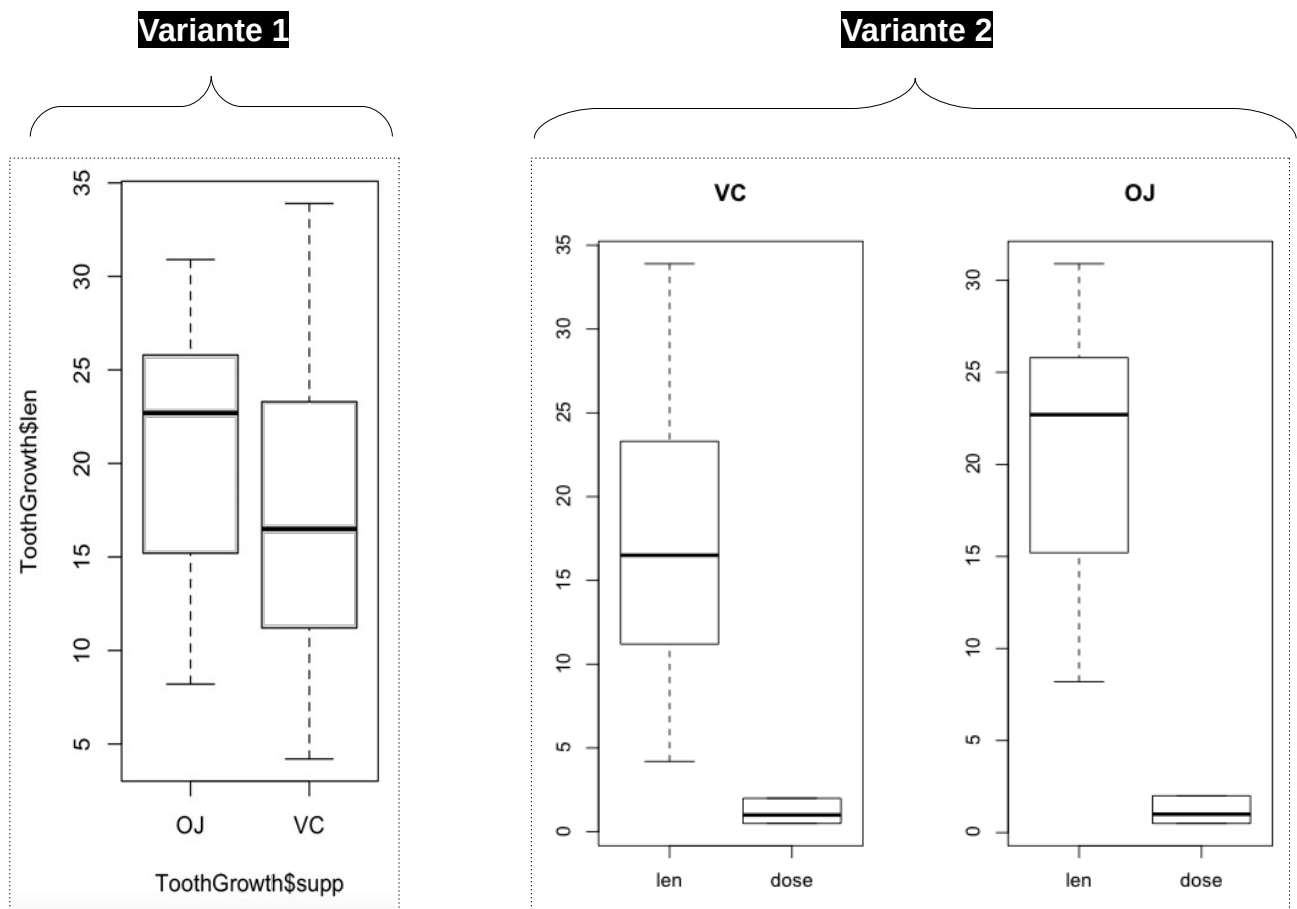
### **Bivariate Statistik (Teil 2)**

Sie haben nun zwei Möglichkeiten Boxplots für die Unterschiede zwischen den Vitamin C Meerschweinchen und den Orangensaft Meerschweinchen in der R Konsole aufzurufen: Zum einen über eine leichte Modifikation des im vorherigen Aufgabenteil dargestellten Befehls, indem Sie die Variable **ToothGrowth\$dose** gegen **ToothGrowth\$supp** austauschen.

Zum anderen über den klassischen Boxplot Befehl bezogen auf die Subsamples für die Vitamin C Meerschweinchen (**`vc[c(1,3)]`**) und die Orangensaft Meerschweinchen (**`oj[c(1,3)]`**). Hierzu weisen Sie R an, zwei ansonsten getrennte Boxplots in einer Grafik zusammenzuführen:

```
par(mfrow=c(1,2))  
boxplot(vc[c(1,3)], main="VC")  
boxplot(oj[c(1,3)], main="OJ")
```

Rufen Sie beide Varianten auf und vergleichen Sie diese miteinander:



### Bivariate Statistik (Teil 3)


Die Boxplots deuten an, dass Orangensaft als Nahrungsergänzungsmittel besser zum Wachstum der Zähne von Meerschweinchen beiträgt als Vitamin C. Diese Vermutung soll abschließend mittels eines Signifikanztests überprüft werden. Dieser lässt sich in R als sogenannter t-Test über den Befehl **`t.test(ToothGrowth$len ~ ToothGrowth$supp)`** aufrufen:

```

Welch Two Sample t-test

data: ToothGrowth$len by ToothGrowth$supp
t = 1.9153, df = 55.309, p-value = 0.06063 ← Signifikanz
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1710156  7.5710156
sample estimates:
mean in group OJ mean in group VC
    20.66333      16.96333

```

  
**Vergleich der Mittelwerte**

Der Signifikanztest bestätigt, dass bei einem Signifikanzniveau von  $p = .060$  ein nicht zufälliger Unterschied zwischen dem arithmetischen Mittel (16.96) der Vitamin C Meerschweinchen und dem arithmetischen Mittel (20.66) der Orangensaft Meerschweinchen besteht. Orangensaft scheint demnach bei dem gegebenen Signifikanzniveau zu durchschnittlich längeren Zähnen bei den Meerschweinchen zu führen als Vitamin C.

### **Lösungen**

- (1) 30 Meerschweinchen erhielten Orangensaft.
- (2) Die durchschnittliche Dosis beträgt 1,16.
- (3) Bei der bisherigen Auswertung kann nicht zwischen Vitamin C und Orangensaft Meerschweinchen unterschieden werden.
- (4) Die durchschnittliche Zahnlänge bei Vitamin C beträgt 16,96.
- (5) Vitamin C, da  $33,90 > 30,90$  Orangensaft.

*The End*

## 5. ÜBUNGSAUFGABE

### Einleitung

Wenn Sie in R über die Befehle ***install.packages("psych", dependencies=TRUE)*** sowie ***library(psych)*** das entsprechende Package installieren, steht Ihnen ein Auszug aus dem Synthetic Aperture Personality Assessment (SAPA) zur Verfügung. In diesem Datensatz sind fünf Persönlichkeitsstrukturen hinterlegt, welche auf Basis einer Selbstauskunft erhoben wurden. Die fünf Persönlichkeitsstrukturen sind unterteilt in: Agreeableness (Verträglichkeit), Conscientiousness (Gewissenhaftigkeit), Extraversion (Extraversion), Neuroticism (Neurotizismus) und Openness (Offenheit), jeweils durch ihren Anfangsbuchstaben mit mehreren Variablen in Verbindung mit drei soziodemographischen Variablen im Datensatz hinterlegt. Dieser kann in R über den Befehl ***bfi*** aufgerufen werden:

	Agreeableness					Conscientiousness					Extraversion					
	A1	A2	A3	A4	A5	C1	C2	C3	C4	C5	E1	E2	E3	E4	E5	N1
61617	2	4	3	4	4	2	3	3	4	4	3	3	3	4	4	3
61618	2	4	5	2	5	5	4	4	3	4	1	1	6	4	3	3
61620	5	4	5	4	4	4	5	4	2	5	2	4	4	4	5	4
61621	4	4	6	5	5	4	4	3	5	5	5	3	4	4	4	2
61622	2	3	3	4	5	4	4	5	3	2	2	2	5	4	5	2
61623	6	6	5	6	5	6	6	6	1	3	2	1	6	5	6	3
	N2	N3	N4	N5	01	02	03	04	05	gender	education		age			
61617	4	2	2	3	3	6	3	4	3		1			NA	16	
61618	3	3	5	5	4	2	4	3	3		2			NA	18	
61620	5	4	2	3	4	2	5	5	2		2			NA	17	
61621	5	2	4	1	3	3	4	3	5		2			NA	17	
61622	3	4	4	3	3	3	4	3	3		1			NA	17	
61623	5	2	2	3	4	3	5	6	1		2			3	21	
	Neuroticism					Openness										

Über den Befehl ***help(bfi)*** können Sie zunächst weitere Hintergrundinformationen zum SAPA sowie das zugrundeliegende Codebuch des Datensatzes mit den dazugehörigen Fragen an die Teilnehmerinnen und Teilnehmer einsehen. Die nachfolgenden Fragen sollen Ihnen bei der ersten Orientierung im Datensatz sowie beim Umgang mit den Variablen behilflich sein:

(1) Wie sind die Fragen zu den Persönlichkeitsstrukturen skaliert?



(2) Wie lautet die dazugehörige Frage zur Variable N2?



(3) Mit welcher Merkmalsausprägung sind Frauen im Datensatz hinterlegt?



Hinweis: Über den Befehl **dim(bfi)** können Sie zusätzlich die Größe (hier: Dimension) des BFI Datensatzes einsehen, um neben der Anzahl an Variablen auch die Anzahl an Teilnehmerinnen und Teilnehmern einsehen zu können.

### Univariate Statistik (Teil 1)

Die Interpretationsmöglichkeiten der deskriptiven Analyse sollen am Beispiel der Variablen Agreeableness und Neuroticism verdeutlicht werden. Dazu wird der Output des Befehls **summary(bfi[c(1:5)])** mit dem Output des Befehls **summary(bfi[c(16:20)])** verglichen. Dabei fällt auf, dass die Variablen mindestens 11 und maximal 36 Fehlwerte beinhalten und das arithmetische Mittel bei Agreeableness durchschnittlich höher ausfällt als bei Neuroticism:

A1		A2		A3		A4		A5	
Min.	:1.000	Min.	:1.000	Min.	:1.000	Min.	:1.0	Min.	:1.00
1st Qu.	:1.000	1st Qu.	:4.000	1st Qu.	:4.000	1st Qu.	:4.0	1st Qu.	:4.00
Median	:2.000	Median	:5.000	Median	:5.000	Median	:5.0	Median	:5.00
Mean	:2.413	Mean	:4.802	Mean	:4.604	Mean	:4.7	Mean	:4.56
3rd Qu.	:3.000	3rd Qu.	:6.000	3rd Qu.	:6.000	3rd Qu.	:6.0	3rd Qu.	:5.00
Max.	:6.000	Max.	:6.000	Max.	:6.000	Max.	:6.0	Max.	:6.00
NA's	:16	NA's	:27	NA's	:26	NA's	:19	NA's	:16



Mean  
4.208

**Differenz (!)**

N1		N2		N3		N4		N5	
Min.	:1.000	Min.	:1.000	Min.	:1.000	Min.	:1.000	Min.	:1.00
1st Qu.	:2.000	1st Qu.	:2.000	1st Qu.	:2.000	1st Qu.	:2.000	1st Qu.	:2.00
Median	:3.000	Median	:4.000	Median	:3.000	Median	:3.000	Median	:3.00
Mean	:2.929	Mean	:3.508	Mean	:3.217	Mean	:3.186	Mean	:2.97
3rd Qu.	:4.000	3rd Qu.	:5.000	3rd Qu.	:4.000	3rd Qu.	:4.000	3rd Qu.	:4.00
Max.	:6.000	Max.	:6.000	Max.	:6.000	Max.	:6.000	Max.	:6.00
NA's	:22	NA's	:21	NA's	:11	NA's	:36	NA's	:29



Mean  
3.164



Mit **`agreeableness_sum <- (bfi$A1+bfi$A2+bfi$A3+bfi$A4+bfi$A5)/5`** kann in R ein Summenscore für Agreeableness angelegt werden. Das arithmetische Mittel kann über den Befehl **`summary(agreeableness_sum)`** aufgerufen werden. Über **`neuroticism_sum <- (bfi$N1+bfi$N2+bfi$N3+bfi$N4+bfi$N5)/5`** in Verbindung mit **`summary(neuroticism_sum)`** kann ein Summenscore als Vergleichswert für Neuroticism aufgerufen werden. Die unterschiedlichen arithmetischen Mittel scheinen darauf hinzudeuten, dass sich das Antwortverhalten bei Agreeableness und Neuroticism deutlich voneinander zu unterscheiden scheinen.

### **Univariate Statistik (Teil 2)**

(4) Welche Agreeableness Variable unterscheidet sich im arithmetischen Mittel deutlich von den anderen Agreeableness Variablen?



-----

(5) Wie lautet jeweils der Summenscore für Openness und Extraversion?



-----

(6) Wie viele Variablen haben bei Conscientiousness ein niedrigeres arithmetisches Mittel als der dazugehörige Summenscore?



-----

(7) Formulieren Sie in eigenen Worten, welchen Einfluss solche “Ausreißer“ bei den Mittelwerten auf die Korrelationen der Variablen innerhalb einer Persönlichkeitsstruktur nehmen können:



-----

## Bivariate Statistik (Teil 1)

Auch für die bivariate Analyse sollen zunächst die Variablen für Agreeableness und Neuroticism betrachtet werden. Dies geschieht in R über die Befehle **`cor(bfi[c(1:5)], use="complete.obs")`** für Agreeableness und **`cor(bfi[c(16:20)], use="complete.obs")`** für Neuroticism. Dabei fällt auf, dass die Variable A1 die niedrigsten Korrelationskoeffizienten mit den anderen Agreeableness Variablen aufweist. Dies ist möglicherweise darauf zurückzuführen, dass auch das arithmetische Mittel dieser Variable von dem der anderen Agreeableness Variablen abgewichen ist. Bei den Variablen für Neuroticism liegen die Korrelationskoeffizienten näher beieinander und liegen alle bei Werten  $> 0.30$ . Auf der nachfolgenden Seite wird eine Möglichkeit vorgestellt, wie man diesen Unterschied in Summe für alle Variablen identifizieren und abbilden kann.

**Variable A1  
korreliert am  
schwächsten**



	A1	A2	A3	A4	A5
A1	1.0000000	-0.3416242	-0.2682817	-0.1483927	-0.1826790
A2	-0.3416242	1.0000000	0.4867503	0.3352432	0.3877875
A3	-0.2682817	0.4867503	1.0000000	0.3621720	0.5051762
A4	-0.1483927	0.3352432	0.3621720	1.0000000	0.3067003
A5	-0.1826790	0.3877875	0.5051762	0.3067003	1.0000000

	N1	N2	N3	N4	N5
N1	1.0000000	0.7057205	0.5559276	0.3985620	0.3768102
N2	0.7057205	1.0000000	0.5454604	0.3902320	0.3523076
N3	0.5559276	0.5454604	1.0000000	0.5180478	0.4279769
N4	0.3985620	0.3902320	0.5180478	1.0000000	0.3975710
N5	0.3768102	0.3523076	0.4279769	0.3975710	1.0000000

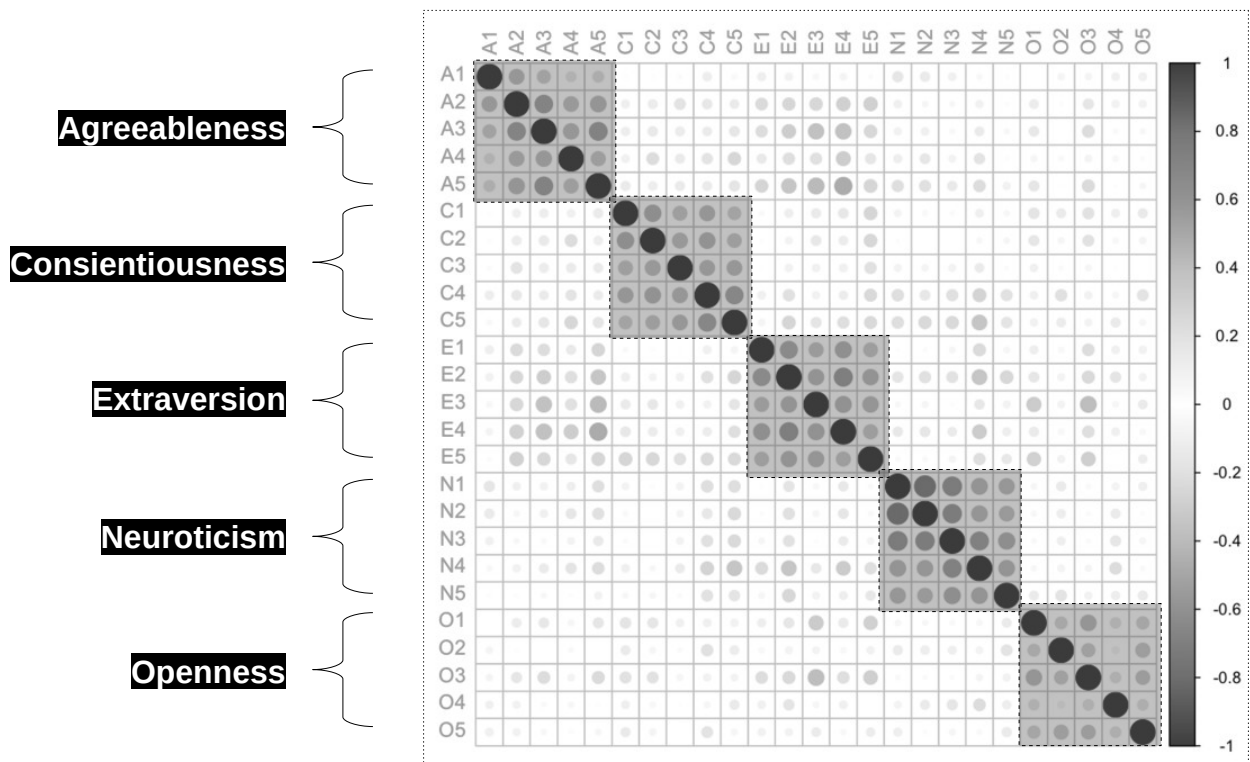
**Variablen korrelieren  
> 0.30 miteinander**

(8) Analysieren Sie entsprechend die Korrelationen bei Conscientiousness, Extraversion und Openness. In welcher Persönlichkeitsstruktur korrelieren die Variablen am schwächsten miteinander?



## Bivariate Statistik (Teil 2)

Um die bivariate Analyse visuell zu unterstützen, kann über die Befehle **`install.packages("corrplot")`** sowie **`library(corrplot)`** das gleichnamige Package aufgerufen werden. Dieses Package visualisiert die Korrelationskoeffizienten, sofern man diese in R als neues Objekt angelegt hat. Dies geschieht über den Befehl **`cor_matrix <- cor(bfi[c(1:25)], use="complete.obs")`** in Verbindung mit dem Befehl **`corrplot(cor_matrix)`**. Für eine bessere Darstellung ist der entsprechende Output nachfolgend ohne die zugrundeliegende Farbgebung für positive (blau) und negative (rot) Korrelationskoeffizienten ausgewiesen:



(9) Die Variablen von Agreeableness korrelieren darüber hinaus besonders mit...



### Faktorenanalyse (Teil 1)

Sowohl die deskriptive als auch die bivariate Analyse deuten darauf hin, dass die jeweiligen Variablen der individuellen Persönlichkeitsstrukturen diesen eindeutig zugeordnet werden können. Die Faktorenanalyse bestätigt diese Zuordnung über eine Minimierung des Residuums (MR-Werte) der Variablen innerhalb einer Persönlichkeitsstruktur. Dafür werden zunächst über `items <- bfi[1:25]` sowie `items <- items[complete.cases(items),]` nur relevante Variablen ohne Fehlwerte ausgewählt, da diese das Ergebnis verfälschen könnten. Die Faktorenanalyse wird in R über den Befehl `fa(items, nfactors=5, rotate="varimax")` aufgerufen. Die Anzahl an vorgegebenen Faktoren entspricht dabei der Anzahl an (bisher für den Datensatz nur theoretisch begründeten) individuellen Persönlichkeitsstrukturen, welche in Spalten von MR1 bis MR5 ausgegeben werden:

```
Factor Analysis using method = minres
Call: fa(r = items, nfactors = 5, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
```

	MR2	MR1	MR3	MR5	MR4	h2	u2	com
A1	0.11	0.04	0.02	-0.43	-0.08	0.20	0.80	1.2
A2	0.03	0.21	0.14	0.63	0.06	0.46	0.54	1.4
A3	0.01	0.32	0.11	0.65	0.06	0.54	0.46	1.5
A4	-0.07	0.20	0.23	0.44	-0.11	0.30	0.70	2.2
A5	-0.12	0.39	0.09	0.54	0.07	0.47	0.53	2.1
C1	0.01	0.07	0.55	0.04	0.21	0.35	0.65	1.3
C2	0.09	0.03	0.65	0.10	0.12	0.45	0.55	1.2
C3	-0.03	0.02	0.56	0.11	-0.01	0.32	0.68	1.1
C4	0.24	-0.06	-0.63	-0.04	-0.11	0.48	0.52	1.4
C5	0.29	-0.18	-0.56	-0.05	0.04	0.44	0.56	1.8
E1	0.04	-0.57	0.03	-0.10	-0.06	0.35	0.65	1.1
E2	0.24	-0.68	-0.10	-0.11	-0.04	0.55	0.45	1.4
E3	0.02	0.54	0.08	0.26	0.28	0.44	0.56	2.1
E4	-0.12	0.65	0.10	0.31	-0.07	0.54	0.46	1.6
E5	0.04	0.50	0.31	0.09	0.21	0.41	0.59	2.2
N1	0.79	0.08	-0.05	-0.22	-0.08	0.68	0.32	1.2
N2	0.75	0.03	-0.03	-0.19	-0.01	0.61	0.39	1.1
N3	0.73	-0.06	-0.07	-0.03	0.00	0.54	0.46	1.0
N4	0.59	-0.35	-0.18	0.01	0.08	0.51	0.49	1.9
N5	0.54	-0.16	-0.04	0.10	-0.15	0.35	0.65	1.4
O1	0.00	0.21	0.12	0.06	0.50	0.32	0.68	1.5
O2	0.18	0.00	-0.10	0.08	-0.47	0.27	0.73	1.4
O3	0.03	0.31	0.08	0.13	0.60	0.47	0.53	1.7
O4	0.22	-0.19	-0.02	0.16	0.37	0.25	0.75	2.7
O5	0.09	-0.01	-0.06	-0.01	-0.53	0.30	0.70	1.1

Die Persönlichkeitsstrukturen Agreeableness, Conscientiousness, Extraversion und Neuroticism weisen innerhalb ihrer Spalten die höchsten MR-Werte auf

Bei der Persönlichkeitsstruktur Openness fallen die MR-Werte insgesamt am geringsten aus

## Faktorenanalyse (Teil 2)

Ausgehend von dem zuvor dargestellten Output in R scheinen die Variablen den fünf individuellen Persönlichkeitsstrukturen entsprechend zugeordnet werden zu können. Darüber hinaus weist der Output in R Gütekriterien für die Faktorenanalyse aus. Dabei wird die Hypothese getestet, dass fünf Faktoren für die individuellen Persönlichkeitsstrukturen ausreichend sind. Als ein wichtiger Testindikator wird der RMSR-Wert ausgegeben. Liegt dieser bei einem Wert  $< 0.08$ , so können die zugrundeliegenden Variablen fünf und nicht mehr Faktoren zugeordnet werden. Demnach scheinen die fünf individuellen Persönlichkeitsstrukturen hinreichend über die insgesamt 25 Variablen abgebildet werden zu können:

The root mean square of the residuals (RMSR) is 0.03  
The df corrected root mean square of the residuals is 0.04

Hinweis: Wenn die Anzahl an Faktoren zunächst nicht bekannt ist, kann das Within Groups Sum of Squares Verfahren herangezogen werden. Dieses kann in R über den Befehl **fa.parallel(items)** aufgerufen werden und legt im vorliegenden Datensatz nahe, dass unter Umständen sechs Faktoren besser geeignet sein könnten.

(10) Bewerten Sie die möglichen sechs Faktoren über eine Faktorenanalyse:



## Lösungen

- (1) Von 1 (very inaccurate) bis 6 (very accurate).
- (2) Get irritated easily.
- (3) 2.
- (4) A1 weist mit 2,413 ein deutlich niedrigeres arithmetisches Mittel auf.
- (5) 3,863 für Openness und 3,791 für Extraversion.
- (6) Insgesamt zwei Variablen: C4 und C5.
- (7) Der Korrelationskoeffizient könnte niedriger sein.
- (8) Die Variablen bei Conscientiousness und Extraversion korrelieren ähnlich wie bei Agreeableness miteinander. Die niedrigsten Korrelationskoeffizienten weisen die Variablen bei Openness auf.
- (9) ...den Variablen der Extraversion.
- (10) Bei einer Faktorenanalyse mit sechs Faktoren lässt sich kein sechster Faktor auf Grundlage der MR-Werte identifizieren. Diese liegen in der Spalte MR6 nahe 00.00.

*The End*

## Einleitung

Über den Befehl ***iris*** können Sie in der R Konsole auf den IRIS Datensatz zugreifen. Die Bezeichnung ist auf die gleichnamige griechische Göttin des Regenbogens zurückzuführen, nach der ebenfalls die Pflanzengattung der Schwertlilien benannt ist. Im IRIS Datensatz finden Sie insgesamt 150 Beobachtungen mit jeweils vier Merkmalsvariablen von Schwertlilien. Der Datensatz enthält Informationen über die Breite und die Länge des Kelchblatts (Sepalum) sowie des Kronblatts (Petalum) in Zentimeter. Darüber hinaus bezieht sich der Datensatz auf drei verschiedene Arten der Schwertlilie (Setosa, Versicolor und Virginica), welche sich jeweils in Breite und Länge des Kelch- und Kronblatts unterscheiden. Ziel der Übungsaufgabe soll es sein, anhand der vier Merkmalsvariablen eine manuelle Identifikation der richtigen Schwertlilienart in der R Konsole zu ermöglichen.

	Kelchblatt		Kronblatt		
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	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	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

## Univariate Statistik (Teil 1)

Zunächst sollen der IRIS Datensatz, die darin enthaltenen Variablen und deren Merkmalsausprägungen möglichst genau beschrieben werden. Dazu können Sie auf den Befehl **summary(iris)** in der R Konsole zurückgreifen:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

(1) Vergleichen Sie die Längen von Kelch- und Kronblatt. Welches Blatt ist länger?

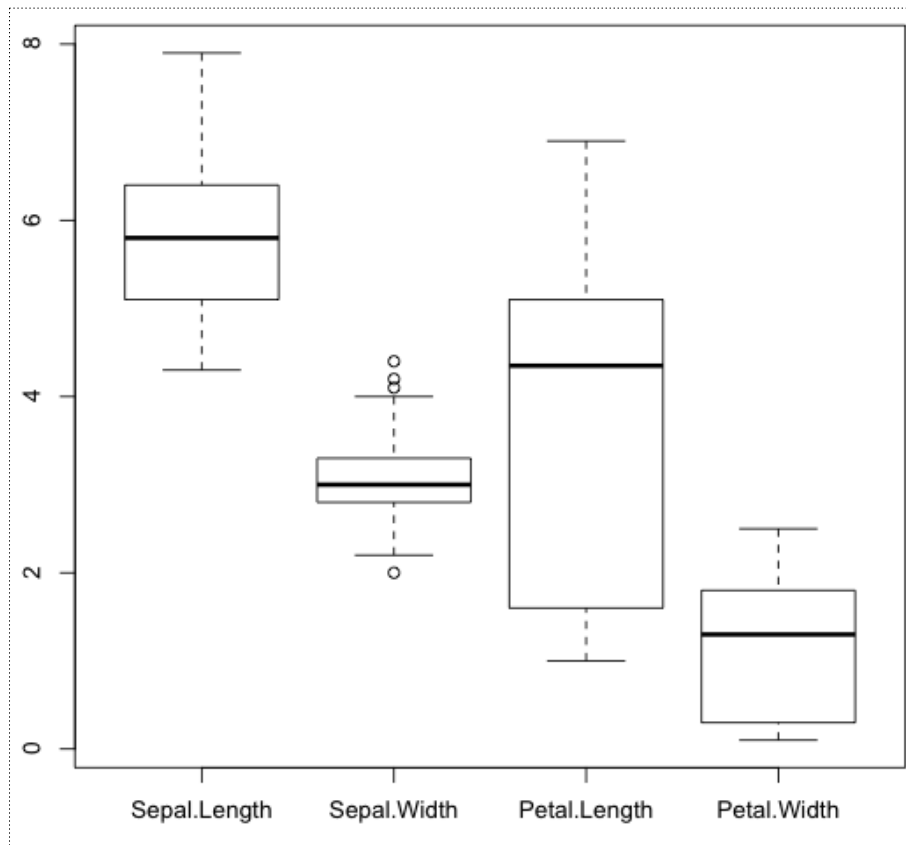


(2) Wie viele Fälle der verschiedenen Schwertlilienarten liegen jeweils vor?



## Univariate Statistik (Teil 2)

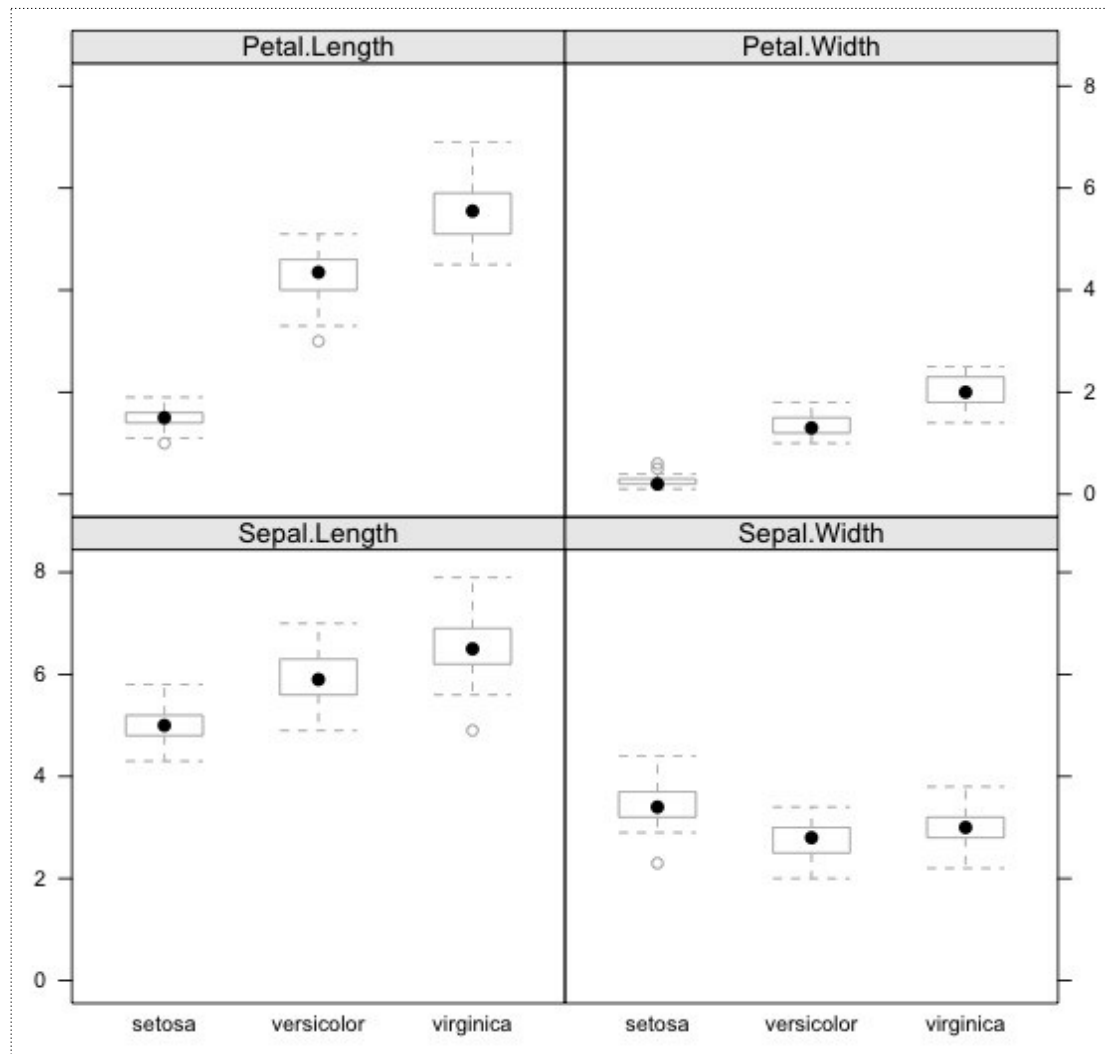
Die vier Merkmalsvariablen zur Breite und Länge von Kelch- und Kronblatt unterscheiden sich jeweils in ihrem Mittelwert als auch in ihrer minimalen und maximalen Merkmalsausprägung deutlich voneinander. Diese Unterschiede können mit dem Befehl **boxplot(iris[1:4])** in der R Konsole visualisiert werden. Dieser Befehl visualisiert nicht alle Variablen des IRIS Datensatzes, sondern lediglich die ersten vier Merkmalsvariablen für alle Schwertlilienarten zusammen:



**Hinweis:** Dass die Merkmalsvariablen unterschiedlich ausgeprägt sind, ist für eine korrekte Identifikation der verschiedenen Schwertlilienarten ausschlaggebend. Ein detaillierter Einblick in die unterschiedlichen Schwertlilienarten ist über diese Darstellung jedoch noch nicht möglich.

## Univariate Statistik (Teil 3)

Weil für Machine Learning Anwendungen ein gutes Verständnis des zugrundeliegenden Datensatzes unerlässlich ist, folgt noch eine weitere deskriptive Analyse. Hierfür wird zunächst über die Befehle **install.packages("caret", dependencies=TRUE)** und **library(caret)** der Funktionsumfang von R erweitert. Diese Erweiterung soll verdeutlichen, dass die verschiedenen Schwertlilienarten jeweils eine unterschiedliche Breite und Länge von Kelch- und Kronblatt aufweisen. Hierzu müssen mit den Befehlen **x <- iris[1:4]** und **y <- iris[,5]** die vier Merkmalsvariablen zur Breite und Länge von Kelch- und Kronblatt von der Variablen getrennt werden, welche die verschiedenen Schwertlilienarten auflistet. Somit können die vier Merkmalsvariablen (x) für jede der drei Schwertlilienarten (y) über den Befehl **featurePlot(x=x, y=y, plot="box")** in der R Konsole visualisiert werden:



(3) Welche Schwertlilienart hat die längsten Kelch- und Kronblätter?



(4) Welche Schwertlilienart hat das schmalste Kronblatt?



### Machine Learning (Teil 1)

Machine Learning Algorithmen beziehen sich in der Regel nicht auf den gesamten Datensatz, sondern werden einem Teil des Datensatzes modelliert und über einen weiteren Teil des Datensatzes validiert. Die zufällige Aufteilung des Datensatzes in bspw. 80% und 20% erfolgt über den Befehl **`validation_index <- createDataPartition(iris$Species, p=0.80, list=FALSE)`** in der R Konsole. Das Besondere an diesem Befehl ist, dass eine zufällige Auswahl an Fällen erfolgt, die relativen Anteile an Schwertlilienarten aber sowohl im 80% Datensatz als auch im 20% Datensatz identisch sind. Diese Vergleichbarkeit erlaubt die Validierung. Den 20% Validierungsdatensatz generiert man



über den Befehl **`validation <- iris[-validation_index, ]`** und den 80% Modellierungsdatensatz über den Befehl **`model <- iris[validation_index, ]`**. In der R Konsole kann das Ergebnis mittels der Befehle **`summary(validation)`** und **`summary(model)`** anschließend kontrolliert werden:

#### **Validierungsdatensatz**



```
Species
setosa   :10
versicolor:10
virginica :10
```

#### **Modellierungsdatensatz**



```
Species
setosa   :40
versicolor:40
virginica :40
```

(5) Wie unterscheiden sich die Mittelwerte als auch die minimalen und maximalen Merkmalsausprägungen im Validierungsdatensatz vom Modellierungsdatensatz?



### **Machine Learning (Teil 2)**

Die zu programmierende Machine Learning Anwendung soll auf ein 10-faches Kreuzvalidierungsverfahren zurückgreifen können. Details hierzu folgen im Seminar / in der Vorlesung. Ein 10-faches Kreuzvalidierungsverfahren wird in der R Konsole über den Befehl **`control <- trainControl(method = "cv", number = 10)`** in Verbindung mit **`metric <- "Accuracy"`** generiert. Nun folgt die Entscheidung für einen oder mehrere Machine Learning Algorithmen. Hier gibt es verschiedene Verfahren, bspw. lineare Modelle, nicht-lineare Modelle sowie die sogenannten komplexen Modelle. Je nach Datensatz und Ausprägung der Merkmalsvariablen führen diese Modelle zu unterschiedlichen Resultaten hinsichtlich ihrer Genauigkeit. Die verschiedenen Machine Learning Algorithmen werden wie folgt in der R Konsole angelegt:

(A) Linearer Algorithmus, hier: Linear Discriminant Analysis (LDA)

```
fit.lda <- train(Species~., data=model, method="lda", metric=metric,
trControl=control)
```

(B) Nicht-linearer Algorithmus, hier: K-Nearest Neighbors (KNN)

```
fit.knn <- train(Species~., data=model, method="knn", metric=metric,
trControl=control)
```

(C) Komplexer Algorithmus, hier: Random Forest (RF)

```
fit.rf <- train(Species~., data=model, method="rf", metric=metric,
trControl=control)
```

### Machine Learning (Teil 3)

Nachdem die Machine Learning Algorithmen in R angelegt wurden, kann über den Befehl **`results <- resamples(list(lda=fit.lda, knn=fit.knn, rf=fit.rf))`** ein Vergleich der Genauigkeit der verschiedenen Algorithmen aufgerufen werden. Zur Erinnerung: Gesucht ist ein Algorithmus, der ausgehend von den vier Merkmalsvariablen im IRIS Datensatz automatisch die richtige Schwertlilienart identifizieren kann. Der Grad der Genauigkeit kann dabei in Prozent angegeben werden. Mit dem Befehl **`summary(results)`** kann der Grad der Genauigkeit tabellarisch und über **`dotplot(results)`** grafisch aufgerufen werden.

Die Accuracy wird in Prozent ausgewiesen, hier bspw. mit 91,66 % min. für den LDA-Algorithmus



Accuracy								
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	
lda	0.9166667	1.0000000	1	0.9833333	1	1	0	
knn	0.8333333	0.9375000	1	0.9666667	1	1	0	
rf	0.7500000	0.9166667	1	0.9416667	1	1	0	

Kappa								
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	
lda	0.875	1.00000	1	0.9750	1	1	0	
knn	0.750	0.90625	1	0.9500	1	1	0	
rf	0.625	0.87500	1	0.9125	1	1	0	



Kappa vergleicht die tatsächliche Genauigkeit mit einer erwarteten Genauigkeit: Zwischen 0.60 und 1.00 liegen akzeptable Werte vor

### Machine Learning (Teil 4)

Von den drei zur Anwendung gebrachten Algorithmen hat die Linear Discriminant Analysis (LDA) als lineares Modell die größte Accuracy und den größten Kappa Wert aufgewiesen. Einen Detailblick in dieses Modell für den Modellierungsdatensatz erhält man in der R Konsole über den Befehl **`print(fit.lda)`**. Man erhält folgenden Output:

```
Linear Discriminant Analysis

120 samples
 4 predictor
 3 classes: 'setosa', 'versicolor', 'virginica'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 108, 108, 108, 108, 108, 108, ...
Resampling results:

Accuracy   Kappa
0.9833333  0.975
```

Hinweis: Zuvor lag der Fokus auf den minimalen Werten für Accuracy und Kappa. Nachdem die Entscheidung für einen Algorithmus getroffen wurde, werden üblicherweise deren Mittelwerte für Accuracy und Kappa ausgewiesen.

(6) Woran erkennen Sie, dass es sich um den Modellierungsdatensatz handelt?



### Machine Learning (Teil 5)

Abschließend greifen wir auf den Validierungsdatensatz zurück und prüfen die Identifikation der Schwertlilienarten anhand der vier Merkmalsvariablen. Dazu gibt man in der R Konsole folgenden Befehl ein:

```
predictions<-predict(fit.Ida, validation)  
confusionMatrix(predictions, validation$Species)
```

Für den Validierungsdatensatz ergibt sich bspw. folgender Output, der anzeigt, wie viele der jeweils 10 verschiedenen Schwertlilienarten aufgrund der vier Merkmalsvariablen richtig identifiziert werden können. Bei einer Genauigkeit von 93,33% werden 10 von 10 Setosa richtig erkannt, 9 von 10 Versicolor und 9 von 10 Virginica:

Prediction	Reference		
	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	9	1
virginica	0	1	9
Overall Statistics			
Accuracy : 0.9333			

(7) Rufen Sie ein Vorhersagemodell auf Basis des gesamten IRIS Datensatzes auf? Wie lauten die Werte für Accuracy und Kappa? Beachten Sie dabei, dass Ihre Werte leicht unterschiedlich ausfallen können, da der Modellierungsdatensatz jeweils nach dem Zufallsprinzip erstellt wird.



## Lösungen

- (1) Das Kelchblatt ist stets (min., max. und mean) länger als das Kronblatt.
- (2) Im gesamten Datensatz ist jede Schwerlilienart 50-mal vertreten.
- (3) Virginica.
- (4) Setosa.
- (5) Annähernd identisch mit  $\pm 0.200$  Unterschied.
- (6) Der Modellierungsdatensatz mit 80 % weist insgesamt 120 Fälle auf.
- (7) Accuracy: 98 % und Kappa: 97 % (dies sind Beispielwerte!).

*The End*