

Schliessende Statistik für Umweltingenieure

Dozentin: Sabine Schilling
Skript: Sabine Schilling, Urs Mürset, Thomas Ott
Institut: Institute of Applied Simulation, ZHAW
Programm: UI16, FS 17

Inhaltsverzeichnis

1	R-Wiederholung	1
1.1	Das Einlesen von Dateien: <code>read.csv</code> , <code>read.csv2</code> , <code>read.table</code>	1
1.2	Wo sind meine EXCEL-files? <code>setwd</code>	2
1.3	Wie sieht meine Datenstruktur aus? Die Befehle <code>View</code> und <code>str</code>	2
1.4	Datentypen in R: Nominale, dichotome und ordinale Daten	2
1.4.1	Nominale und dichotome Daten sind in R vom Datentyp <code>factor</code>	2
1.4.2	Ordinale Daten: <code>factor</code> mit Reihenfolge <code>levels</code>	3
1.5	Zugriff auf die Spalten eines <code>data.frame</code> mit <code>\$</code>	3
1.6	<code>attach</code>	4
2	Regressions-Analyse	5
2.1	Lineare Regression	6
2.1.1	Lineare Regression in R: <code>lm</code> , <code>abline</code> und <code>residuals</code>	9
2.1.2	Korrelations-Koeffizient	10
2.1.2.1	Metrische Daten: Pearson – Korrelationskoeffizient	12
2.1.2.2	Ordinale Daten: Spearman – Korrelationskoeffizient	13
2.1.2.2.1	Berechnung des korrigierten Rangs mit <code>rank</code>	14
2.1.3	Bestimmtheitsmass r^2	16
2.2	Nichtlineare Regression	16
2.2.1	Nullhypothese, p-Wert und <code>summary</code> Befehl	19
2.3	Multivariate Regression	20
3	Schliessende Statistik	25
3.1	Das Problem der Stichproben	25
3.2	Binomialverteilung	28

3.2.1	Kuglbrett	29
3.2.2	Binomialverteilung in R	32
3.2.3	Lage und Streuung einer Binomialverteilung	33
3.3	Zusammenhang zwischen Binomial- und Normalverteilung	34
4	Schätzungen aus Stichproben	39
4.1	Konfidenz und Signifikanz	39
4.2	Anteilsschätzungen	41
4.2.1	Dichotome Daten	41
4.2.2	Nominale und ordinale Daten	44
4.3	Schätzung eines Durchschnitts (metrische Daten)	46
4.4	Schätzung einer Anzahldichte	48
4.5	Standardfehler	50
4.6	Konfidenzintervall der univariaten Regression	51
5	Statistische Tests -Überblick	53
5.1	Die Grundfragen der schliessenden Statistik	53
5.2	Testen von Hypothesen	54
5.3	Parametrische und parameterfreie Tests	54
5.4	Fehler 1. und 2. Art	55
6	Parametrische Tests	57
6.1	Der t -Test	57
6.1.1	Der t -Test im 1-Stichproben-Fall	57
6.1.1.1	Voraussetzungen für den t -Test	59
6.1.1.2	Das t -Test-Rezept	60
6.1.2	Allgemeines und Varianten zur Testerei	62
6.1.2.1	Freiheitsgrad	64
6.1.3	Der t -Test für zwei unabhängige Stichproben	64
6.1.3.1	Voraussetzungen für den t -Test im 2-Stichprobenfall	65
6.1.3.1.1	Normalverteilung in beiden Stichproben	65
6.1.3.1.2	Varianzhomogenität: Student's t -Test und Varianz- heterogenität: Welch- t -Test	65
6.1.4	Visualisierung der Daten vor dem Testen	68
6.1.5	t -Test für zwei abhängige Stichproben	71

6.1.5.1	Gekoppelte Stichproben als 1-Stichproben-Fall	74
6.2	Überprüfen der Normalverteilungsannahme	75
6.2.1	Normal-Quantil-Diagramm oder QQ-Plot	75
6.2.1.1	QQ-Plot in R	75
6.2.2	Statistische Tests zur Überprüfung der Normalverteilungsannahme . .	76
6.3	Überprüfen der Varianzhomogenität: Der F -Test	77
7	Parameterfreie Tests	81
7.1	Anteilstest	83
7.2	Wilcoxon-Test	86
7.3	Der χ^2 -Test	90
7.3.1	Vergleich zweier oder mehrerer Verteilungen in ihrer Form: χ^2 -Test, 2-oder mehr Stichproben-Fall	91
7.3.1.1	χ^2 -Test in R	94
7.3.1.1.1	Voraussetzungen des χ^2 -Tests	94
7.3.2	Vergleich mit einer vorgegebenen Verteilung: χ^2 -Test, 1-Stichproben-Fall	97
8	Rückblick auf die verschiedenen statistischen Tests	101
9	Überblick multivariate Methoden	103
9.1	Einfaktorielle Verfahren zum Vergleich von Mittelwerten und Medianen . . .	104
10	Parametrische einfaktorielle Varianz-Analyse	107
10.1	Einfaktorielle ANOVA	107
10.1.1	Voraussetzungen	107
10.1.2	Vorüberlegungen	107
10.1.2.1	Das Problem multipler Paarvergleiche	109
10.1.3	Die Grundidee der ANOVA	109
10.1.4	Einfaktorielle Varianzanalyse mit R	110
10.1.4.1	Datenvorbereitung	110
10.1.4.1.1	In Excel	110
10.1.4.1.2	In R	111
10.1.4.2	Testrezept der einfaktoriellen ANOVA	112
10.1.5	Post-hoc-Test	114

10.1.6 Varianzanalyse in R: Die unabhängigen Variablen müssen vom Daten- typ <code>factor</code> sein	115
10.2 <code>oneway.test</code>	116
11 Parameterfreie einfaktorielle Varianz–Analyse	119
11.1 Kruskal–Wallis–Test (H–Test) bei metrischen Daten	119
11.2 Kruskal–Wallis–Test (H–Test) bei ordinalen Daten	120
12 Zweifaktorielle Varianzanalyse	123
12.1 Wechselwirkung der Faktoren	123
12.2 Voraussetzungen für die zweifaktorielle ANOVA	127
12.3 Zweifaktorielle ANOVA ohne Wiederholung	128
A Lösungen zu den Fragen im Text	L.1
A.1 Lösungen zu den Fragen in Lektion 2	L.1
A.2 Lösungen zu den Fragen in Lektion 3	L.1
A.3 Lösungen zu den Fragen in Lektion 4	L.2
A.4 Lösungen zu den Fragen in Lektion 5	L.8
A.5 Lösungen zu den Fragen in Lektion 6	L.10
A.6 Lösungen zu den Fragen in Lektion 7	L.11
A.7 Lösungen zu den Fragen in Lektion 8	L.14
A.8 Lösungen zu den Fragen in Lektion 9	L.14
A.9 Lösungen zu den Fragen in Lektion 10	L.16
A.10 Lösungen zu den Fragen in Lektion 11	L.16
B Ergänzungen ANOVA (fakultativ)	A.1
B.1 Grundidee	A.1
B.2 Quadratsummen	A.3
B.3 Berechnung der Teststatistik: die Prüfgrösse F	A.3
B.4 Rezept für einfaktorielle ANOVA ohne R	A.4
B.5 Beispiel: Lebenserwartung	A.4

Kapitel 1

R-Wiederholung

1.1 Das Einlesen von Dateien: `read.csv`, `read.csv2`, `read.table`

In der Mathematik ist das Statistikprogramm **R** erstmals eingeführt worden, im Rahmen des Statistikunterrichts wird **R** eine ganz zentrale Rolle spielen. Eine Hauptschwierigkeit ist oftmals, die Daten, die Du während Deiner Semester-, Bachelor- oder Masterarbeit oft in einem EXCEL-Sheet abgespeichert hast, in **R** einzulesen. Die in EXCEL erfassten Daten werden in **R** übertragen, indem du sie

- zuerst (noch in EXCEL) als csv-Datei abspeicherst („speichern unter ...“),
- und dann in **R** einliest.

Je nachdem, ob Du mit einem Windows-Rechner oder einem Mac arbeitest, trennt Excel beim Abspeichern als csv-Datei die einzelnen Zellen des EXCEL-Sheets standardmässig mit einem Strichpunkt (Windows) oder mit einem Komma (Mac) ab. Für beide Varianten hat **R** spezielle Befehle, die dir das Einlesen erleichtern sollen. In folgendem Beispiel nehmen wir an, dass wir eine csv-Datei namens `so_heisst_es_bisher.csv` einlesen wollen:

- csv-Files, deren Zellen mit einem Strichpunkt getrennt sind (typischerweise Windows generiert):

```
so_soll_das_von_nun_an_heissen=read.csv2("so_heisst_es_bisher.csv")
```

- csv-Files, deren Zellen mit einem Komma getrennt sind (typischerweise Mac generiert):

```
so_soll_das_von_nun_an_heissen=read.csv("so_heisst_es_bisher.csv")
```

`read.csv` und `read.csv2` sind Spezialfälle des allgemeineren Befehls `read.table`. Mit `?read.table` kannst Du Dir zahlreichen Optionen dieses Befehls ansehen.

1.2 Wo sind meine EXCEL-files? `setwd`

Am besten legst Du einen Ordner aller relevanten Dateien an, kopierst alle betreffenden Dateien dort hinein und teilst **R** mit einem Befehl der Art

```
setwd(/Pfad/zu/deinen-Dateien")
```

mit, wo es nach diesen Dateien suchen soll. Mit dem Befehl `getwd()` kannst Du überprüfen, in welchem Ordner **R** gerade nach Dateien sucht.

1.3 Wie sieht meine Datenstruktur aus? Die Befehle `View` und `str`

Um zu überprüfen, ob das Einlesen geklappt hast, kannst Du Dir die eingeleseene Datei mit dem Befehl

```
View(so_soll_das_von_nun_an_heissen)
```

ansehen.

Um zu verstehen, welche Datenstruktur Deine eingeleseene Tabelle besitzt und von welcher Art die Daten sind, benutze den Befehl

```
str(so_soll_das_von_nun_an_heissen)
```

1.4 Datentypen in R: Nominale, dichotome und ordinale Daten

Oftmals stehen metrische Datensätze im Zentrum unserer Fragestellungen, die wir in **R** typischerweise mit entsprechenden Datenvektoren (beispielsweise `x=c(17,31,5,9,...)`) eingeben.

1.4.1 Nominale und dichotome Daten sind in R vom Datentyp `factor`

Ganz ähnlich können wir mit nominalen Daten verfahren. Der entsprechende Datentyp in **R** heisst `factor`. Das sieht dann z.B. so aus:

```
x=factor(c("blau","rot","gruen","rot","blau"))
```

Auch dichotome Daten kann man auf dieser Weise darstellen:

```
x=factor(c("ja","ja","nein","ja","nein"))
```

(Obwohl es dafür eigentlich einen spezialisierten Datentyp (`logical`) gibt.)

1.4.2 Ordinale Daten: `factor` mit Reihenfolge `levels`

Zwischen den Worten „nass“, „feucht“ und „trocken“ besteht für uns eine Reihenfolge. Doch wie bringen wir die `R` bei? Hierbei hilft uns wieder der `factor`-Befehl, diesmal jedoch mit dem Zusatz `levels`. In einem ersten Schritt implementiert man ordinale Daten als Datenvektor:

```
daten=c("feucht", "nass", "feucht", "nass", "trocken")
```

Um den `factor`-Befehl auf ordinale Daten anwenden zu können, müssen wir `R` zusätzlich mitteilen, in welcher Reihenfolge die ordinalen Begriffe (z. B. „nass“ < „feucht“ < „trocken“) zu ordnen sind. Dies macht das `levels`-Argument des `factor` - Befehls. Beispiel:

1. Dateneingabe als Datenvektor:

```
daten=c("feucht", "nass", "feucht", "nass", "trocken")
```

2. Festlegen der Reihenfolge:

```
reihenfolge=c("nass", "feucht", "trocken")
```

3. Wende `factor` - Befehl mit `levels` -Option an:

```
x=factor(daten, levels=reihenfolge)
```

Wenn du die `levels=...`-Option weglässt, geht `R` von alphabetischer Reihenfolge aus (probier es aus!).

R-Frage 1 Führe den `factor`-Befehl auf `daten` ohne die `levels`-Option aus:
`y=factor(daten)`
 Vergleiche mit dem Resultat mit `levels`-Option. Was fällt Dir auf? Hinweis: Benutze den `str`-Befehl.

R-Frage 2 a) Lade von Moodle die Datei `Kaese.xlsx` herunter, verwandle die Datei in ein csv-file und lies die Dateien in ein `data.frame` namens `Kaese` ein.
 b) Sieh Dir den `data.frame` mit dem Befehl `View` an und erkenne, was für ein Datentyp in den Spalten steht.

1.5 Zugriff auf die Spalten eines `data.frame` mit `$`

Ein `data.frame` ist in `R` eine Tabelle, in deren Spalten unterschiedlichen Datentypen stehen können. Auf die einzelnen Spalten eines `data.frame` greifst Du mit der `$`-Notation zu: Im Kaese-Beispiel kannst Du Dir die Spalte „Sorte“ mit der Syntax `Kaese$Sorte` ansehen.

R-Frage 3 Sieh Dir in `R` die Spalten „Sorte“ und „Naehrwert“ des `data.frame` `Kaese` aus der vorigen Aufgabe an.

1.6 `attach`

Manchmal möchte man sich Schreibarbeit sparen und die `$`-Notation vermeiden. Nach Ausführen von `attach(Name_data.frame)` kann man direkt auf die Spalten des `data.frame` zugreifen, in dem man nur den Spaltennamen benutzt.

R-Frage 4 *Führe den Befehl `attach(Kaese)` aus. Vergleiche die Ausgabe der Befehle `Kaese$Naehrwert` und `Naehrwert`.*

CHECKLISTE

Kannst du jetzt:

- *R mitteilen, in welchem Verzeichnis sich Deine Daten befinden?*
- *ein EXCEL-File von Moodle herunterladen, auf Deinem Computer in ein csv-File umwandeln, in ein Dir genehmes Verzeichnis verschieben und dann in R einlesen?*
- *R die Reihenfolge Deiner ordinalen Daten beibringen?*
- *auf die Spalten eines `data.frame` zugreifen?*
- *die R-Fragen dieses Kapitels selbständig lösen?*

Kapitel 2

Regressions–Analyse

Das Ziel der Regressionsanalyse ist es, Beziehungen zwischen einer abhängigen Variable y von einer oder mehreren unabhängigen Variablen x_1, x_2, \dots zu modellieren. Gesucht ist also eine Funktion $y = f(x_1, x_2, \dots)$ in analytischer Form, was wiederum bedingt, dass alle unabhängigen und abhängigen Variablen metrisch sein müssen.

Wir gehen zuerst den einfachsten Fall mit nur einer unabhängigen Variablen x an. Dazu kommt natürlich eine abhängige Variable ($y = f(x)$). D.h. die Daten bestehen aus (x, y) –Wertepaaren. Als Beispiel soll [baum.xlsx](#) dienen, das du auf MOODLE findest. Es enthält Messwerte von 10 Kirschbaumstämmen. Es wurde sohl die Stammdicke (in cm) als auch die Höhe (in m) erfasst

R–Frage 5 Lade den Datensatz [baum.xlsx](#) von MOODLE herunter. Wandle ihn in EXCEL in ein [cvs](#)–file um und lese dieses in **R** ein. Betrachte den Datensatz mit [View](#) and analysiere die Datenstruktur mit [str](#). Führe den [attach\(baum\)](#)–Befehl aus und erstelle sodann einen Scatterplot mit

```
plot(baum$Dicke,baum$Hoehe)
```

Überzeuge Dich vom Funktionieren des [attach](#)–Befehls, in dem Du zeigst, dass

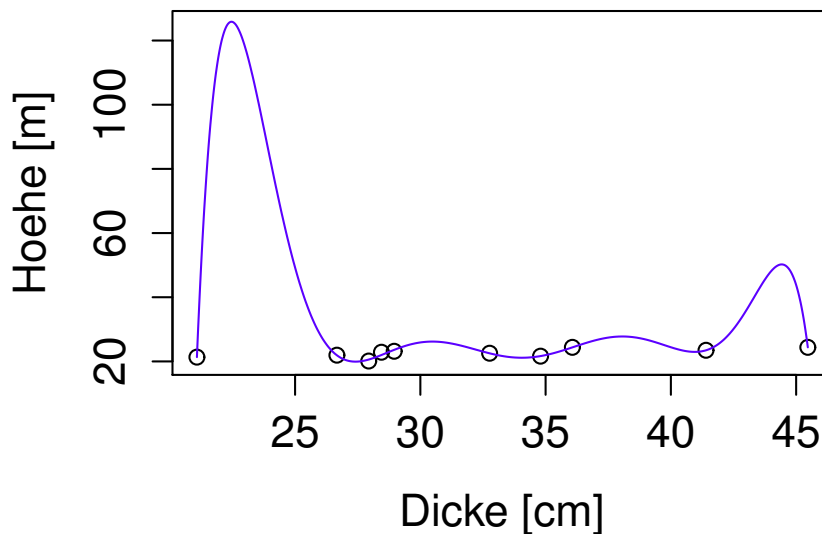
```
plot(Dicke,Hoehe)
```

zum selben Plot führt.

Bei zwei Variablen drängt sich meist diese Frage zuerst auf: Haben die beiden Variablen etwas miteinander zu tun? Oder anders ausgedrückt: Bedeuten verschiedene x –Werte auch verschiedene y –Werte?

Wie beurteilst du das Diagramm mit den Baum–Daten? Du wirst vielleicht sagen: Es gibt eine gewisse Tendenz, dass dickere Stämme höher gewachsen sind – scheint ja auch logisch. Aber ganz so perfekt ist die Sache nicht, die Punkte ordnen sich nicht wirklich auf einer klaren und einfachen Linie an.

Trotzdem können wir jetzt versuchen, die wahrgenommene Tendenz durch eine Funktion $f(x)$ zu beschreiben. Das kann grundsätzlich *irgend* ein Typ von Funktion sein, etwa ein Polynom oder eine Exponentialfunktion. Es ist aber wichtig, dass man einen möglichst einfachen Typ von Funktion wählt, am besten eine lineare Funktion. Wählte man zu den Baumdaten z.B. ein Polynom 9. Grades, käme es so heraus:



Die Kurve trifft die Punkte jetzt zwar perfekt; denn ein Polynom 9. Grades hat 10 Koeffizienten, die man immer so wählen kann, dass 10 Punkte getroffen werden (vorausgesetzt sie unterscheiden sich alle in der x -Koordinate). Aber das Polynom gibt wohl nicht wirklich die Tendenz der Daten wieder, oder? 120 m hohe Kirschbäume mit Stammdurchmessern zwischen 20 cm und 25 cm, aber bei grösserem Durchmesser dann wieder nicht so hohe? Und käme ein 11. Punkt hinzu, könnte er durchaus sehr weit von der Kurve entfernt sein, bzw. zu einem völlig anderen Regressionsresultat führen.

2.1 Lineare Regression

Tatsächlich sollte man von einer linearen Funktion (einem Polynom 1. Grades) nicht abweichen, wenn nicht entweder eine Krümmung nach oben oder unten deutlich erkennbar ist oder vom Verständnis der Situation her ein bestimmter Funktionstyp angezeigt ist. Ein Beispiel hierfür ist der freie Fall eines Gegenstandes; die Abhängigkeit der zurückgelegten Strecke s von der unabhängigen Variablen Zeit t wird hier durch ein Polynom 2. Grades beschrieben: $s = f(t) = \frac{1}{2} \cdot g \cdot t^2$. Kompliziertere Funktionen sind nur dann sinnvoll, wenn die Daten sehr gut sind (so dass man den Funktionstyp tatsächlich festlegen kann) oder der Funktionstyp

vorgegeben ist (wie im Beispiel des freien Falls). Eine Regression mit einer linearen Funktion nennt man verkürzt lineare Regression. Das Resultat nennt man auch Trendgerade. Doch wie findet man diese Trendgerade? „Von Hand“ legt man eine Gerade ins Diagramm, die möglichst gut zu den Datenpunkten passen soll. Angenommen, du hast eine Anzahl Messungen gemacht, sie fein säuberlich in einem Diagramm dargestellt, und nun solltest du in deine Messdaten eine Kurve hineinlegen. Kein Problem, wenn die Messdaten so aussehen wie im Diagramm links in der Abbildung 4.3.

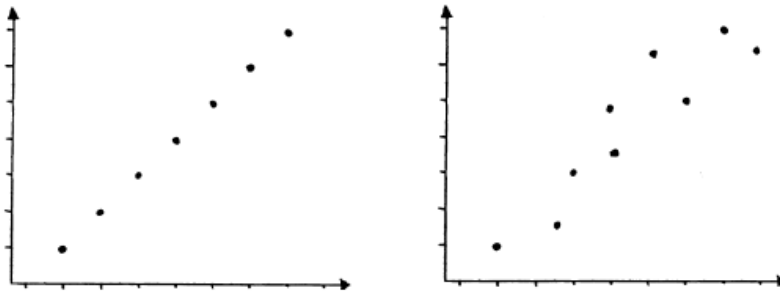


Abbildung 2.1: Messdaten

Offensichtlich zeigen die Messungen einen linearen Zusammenhang zwischen x und y , und es ist auch nicht schwierig, eine Gerade hineinzulegen und Achsenabschnitt und Steigung daraus zu bestimmen. Was machst du, wenn deine Messergebnisse eher denen im Diagramm rechts entsprechen? Wenn also wegen Messfehlern oder sonstigen Gründen nicht alle Daten perfekt zusammenpassen?

Die Daten deuten noch immer auf einen linearen Zusammenhang zwischen x und y hin; die Punkte ordnen sich aber nicht exakt auf einer Geraden an, z.B. weil sie Ungenauigkeiten enthalten; ich lege eine Gerade durch die Punkte, so dass diese *so gut wie möglich* durch die Gerade repräsentiert werden. Man nennt dies eine „Regressions-“ oder „Trend-Gerade“. Du kannst eine solche Trend-Gerade nach Augenmass hineinlegen. Reichst du die ursprünglichen Daten an eine Kommilitonin weiter, wird diese eine ähnliche Gerade hineinlegen – aber wahrscheinlich doch nicht exakt die gleiche.

Und jetzt? Wer hat recht, welches ist die *beste* Gerade? Diese Frage lässt sich im Prinzip nicht vollkommen objektiv beantworten; aber es gibt eine Methode, die mit Abstand am weitesten verbreitet ist, die „Methode der kleinsten Quadrate“ (engl.: „least squares fit“): Am besten stellst du dir vor, dass die Punkte Ortschaften sind, und dass du eine gerade Autobahn so legen sollst, dass die Summe aller Zufahrtsstrassen möglichst kurz ist.

- Aus rechnerischen Gründen minimieren wir nicht die Länge der Zufahrtsstrassen direkt, sondern die Quadrate der Längen (daher der Name der Methode).
- Die Zufahrtsstrassen (Residuen genannt) werden vertikal (d.h. in y -Richtung) gezogen. Residuen sind somit die Distanzen, um die die Messwerte von der Regressionsfunktion in y -Richtung abweichen. Also sozusagen wie falsch die Messwerte aus Sicht

der Regression sind. Im Baumstamm-Beispiel sind die Residuen als dicke, vertikale Linien hervorgehoben (Abbildung 2.2a). Oftmals zeigt man nur die Residuen, ohne die Trendlinie (Abbildung 2.2b) oder man lässt die „Zufahrtstrassen“ ganz weg und zeigt nur die Punkte (Abbildung 2.2c).

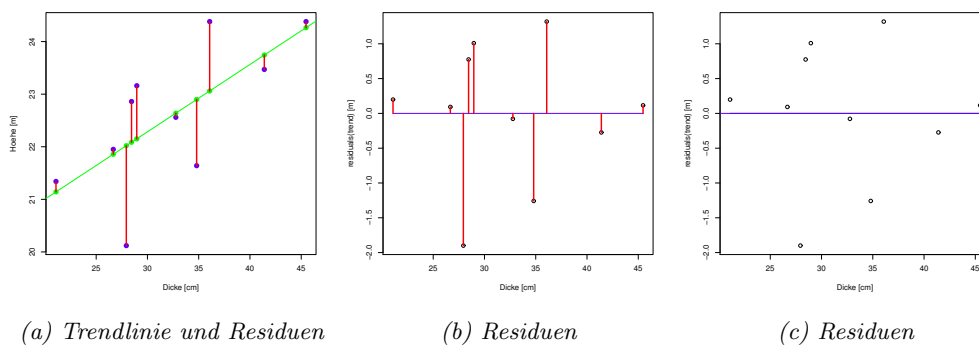


Abbildung 2.2: Trendlinie und Residuen

- Die Messwerte (blaue Punkte in Abb. 2.2a) bezeichnen wir mit $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, oder allgemein (x_i, y_i) mit $i \in [1, 2, \dots, n]$ ¹, wobei n die Anzahl der Messpunkte ist.
- Die zugehörigen Wertepaare auf der „optimalen“ Regressionsgeraden (grüne Punkte in Abb. 2.2a) bezeichnen wir mit (\hat{x}_i, \hat{y}_i) .
- Die Residuen (die Länge der Zufahrtsstrassen) sind gegeben als

$$e_i = y_i - \hat{y}_i.$$

- Aus Abbildung Fig. 2.2a sehen wir: $\hat{x}_i = x_i$ für alle i .
- Die Gleichung der linearen Regressionsgeraden setzen wir an als $\hat{y} = f(x) = a \cdot x + b$.
- Ziel ist die Bestimmung von a und b .
- Mathematisch löst man das Problem, dass **die Summe der Quadrate der Residuen möglichst klein** ist:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - f(x_i))^2 = \sum_{i=1}^n (y_i - (a \cdot x_i + b))^2. \quad (2.1)$$

Das ist ein Optimierungsproblem für die zwei Variablen a und b . Um a und b zu bestimmen, bildet man von Gleichung 2.1 die partiellen Ableitungen nach a bzw. b , setzt das jeweilige Ergebnis 0 und löst nach a und b auf².

¹ $i \in [1, 2, \dots, n]$ bedeutet, dass i ein Teil der Menge der natürlichen Zahlen zwischen 1 und n ist. „ \in “ ist die mathematische Kurznotation für „ist ein Element von“.

² Für Interessierte und Neugierige: Auf Moodle findest Du eine ausführliche Herleitung.

2.1.1 Lineare Regression in R: `lm`, `abline` und `residuals`

In diesem Abschnitt zeigen wir, wie man mit `R` die Koeffizienten a und b der linearen Regression berechnet und auch gleich die entsprechende „optimale“ Trendgerade durch die Datenpunkte legt. Die lineare Regression wird in `R` mit der Funktion `lm` ausgeführt. `lm` steht für lineares Modell – eine abgekürzte Formulierung dafür, dass eine lineare Funktion gesucht wird, die als Modell für die Daten dienen kann. Mit `abline` kann man die Trendgerade anschliessend zeichnen. Es muss zuerst die abhängige Variable, dann die unabhängige Variable genannt werden: `lm(abhängige Variable ~ unabhaengige Variable)`.

R-Frage 6 *Probiere es aus:*

```
trend=lm(baum$Hoehe~baum$Dicke)
```

Du kannst die Trendgerade mit

```
abline(trend,col="blue")
```

einzeichnen.

R-Frage 7 *Die Koeffizienten a und b der Trendgerade $y(x) = ax + b$ kannst du mit*

```
coefficients(trend)
```

anschauen. `Intercept` bezeichnet den y -Achsenschnittpunkt, also b , `baum$Dicke` die Steigung der Geraden $\Delta y/\Delta x$, das wäre dann a .

Einzeln herausholen (zum Weiterrechnen) kannst du die Parameter der Geradengleichung mit

```
b=trend$coefficients[1]
```

```
a=trend$coefficients[2]
```

Damit kannst du nun selber Berechnungen anstellen, z.B.:

```
x=0:50
```

```
y=a*x+b
```

```
plot(x,y)
```

Die Gleichung der Trendgeraden lautet somit

$$y(x) = 0.13 x + 18.43$$

Beachte, dass die Dicke x in cm, die Höhe jedoch in Metern in unsere Formel eingeht. Wenn ein Baumstamm also um 1 cm dicker ist als ein anderer, dann ist er um 0.13 m höher. Und ein Baumstamm mit Dicke 0 ist 18.43 m hoch. Findest du die letzte Aussage logisch? Das bringt uns auf ein generelles Problem mit Regressionen: *Die Extrapolation weit über den Messbereich hinaus ist höchst gefährlich.* Vermutlich ist das lineare Modell in einem gewissen Bereich gut, eben jenem der Werte in `baum.csv` – aber nicht für alle Bäume (z.B. nicht für ganz junge). Es brauchte für die Beschreibung *aller* Kirschbäume ein komplizierteres Modell. Dies aber kann man jedoch nur auf entsprechende Daten aufbauen.

R-Frage 8 Du kannst auch erzwingen, dass die Regressionsgerade durch den Ursprung geht. Das macht die Regression mit dem Modell

```
trendu=lm(baum$Hoehe~baum$Dicke+0)
```

Betrachte den Output zusammen mit den Daten, und du wirst erkennen, dass dieses Modell nicht überzeugt.

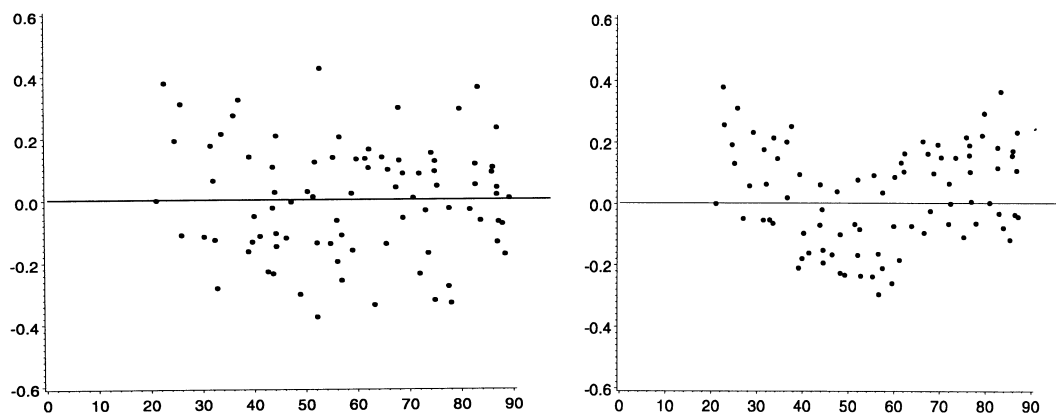
R-Frage 9 Du kannst einen „Residuen-Plot“ ganz einfach produzieren:

```
plot(baum$Dicke,residuals(trend))
```

```
plot(baum$Dicke,baum$Hoehe-predict(trend))
```

liefert den selben Plot. Überzeuge Dich!

Die Residuen sollten gleichmässig verstreut sein, so wie in der folgenden Abbildung links:



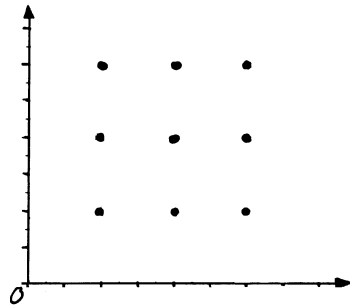
Ist dagegen, wie rechts, auch in den Residuen wieder ein Muster zu erkennen (hier so ein nach unten durchhängendes Band, in welchem alle Punkte liegen), deutet dies darauf hin, dass eine Gerade nicht das richtige Modell für eine (lineare!) Regression ist.

Frage 1 Wie sehen die ursprünglichen Daten aus, wenn der Residuen-Plot herauskommt wie im Fall rechts?

2.1.2 Korrelations-Koeffizient

Damit wissen wir, welches die *beste* Gerade ist; aber wir wissen noch nicht, ob sie *gut* ist. Liegen die Daten nahe bei der Geraden oder streuen sie weit umher? Um der Sache etwas näher zu kommen probieren wir uns an einem weiteren Datensatz:

Frage 2 *Versuche zu schätzen (ohne Rechnung), wie die Regressionsgerade zu folgenden Daten aussieht:*



Noch ein Gedanke zu den Daten dieser letzten Frage: *gibt* es denn wirklich einen Zusammenhang zwischen den Variablen y und x ? Es kommen ja alle Kombinationen

kleines x / kleines y	kleines x / grosses y
grosses x / kleines y	grosses x / grosses y

in gleichem Masse vor. Es ist nicht so, dass beispielsweise ein grösseres x auch ein grösseres y implizieren würde und umgekehrt. Eine Regression ist schlicht unsinnig, weil es gar keine zu beschreibende Tendenz gibt.

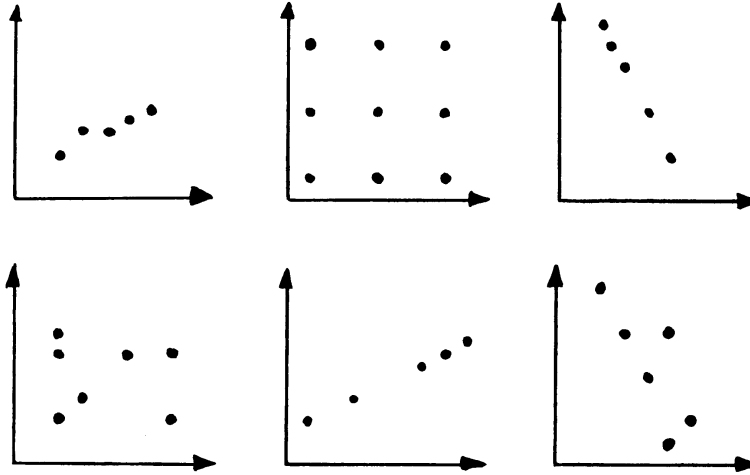
Reihen sich die Daten wirklich auf einer Geraden auf? Oder sieht es eher so aus wie in der letzten Frage, oder irgendwie dazwischen? Es braucht unübersehbar ein Mass, das uns sagt, wie klar der Zusammenhang ist, wie gut die Trend-Gerade passt, wie sinnvoll die Regression ist. Ein „Korrelations-Koeffizient“ ist ein Mass für die Güte einer Regression. Üblicherweise wird ein Korrelations-Koeffizient mit der Formel-Zeichen r abgekürzt. Dabei bedeutet

- positives r : steigende Trend-Gerade („positive Korrelation“)
- negatives r : fallende Trend-Gerade („negative Korrelation“, „Anti-Korrelation“)
- $r = 0$: keinerlei Korrelation (man kann überhaupt keine sinnvolle Gerade legen)
oder die Trendgerade wäre exakt horizontal
- $|r| = 1$: perfekte Korrelation, alle Punkte liegen exakt auf einer steigenden ($r = +1$)
oder fallenden ($r = -1$) Geraden)
- $0 \leq |r| \leq 1$: je näher $|r|$ bei 1, desto klarer die Korrelation

Steigung 0 (horizontale Trendgerade) bedeutet, dass kein Zusammenhang zwischen x und y besteht, denn zu jedem x würde so ja das gleiche y gehören, bzw. zu diesem y würden *alle* x -Werte gehören, während alle anderen y -Werte ohne ein ihnen entsprechendes x auskommen müssten. Daher ist auch dann $r = 0$, selbst wenn die Punkte sich auf einer horizontalen Geraden perfekt aufreihen. Hier gibt es einen Unterschied zwischen der „normalen“ mathematischen Sichtweise und der statistischen: exakt horizontal liegende Punkte können mathematisch durch eine lineare Funktion perfekt beschrieben werden, aber sie stellen kei-

nen statistischen Zusammenhang zwischen den involvierten Variablen dar.

Frage 3 In der folgenden Galerie von Diagrammen sind die Korrelations-Koeffizienten $-1.00, -0.79, -0.13, 0.00, +0.96, +1.00$ vertreten. Welcher Wert gehört zu welchem Diagramm?



Werte mit negativem r bezeichnet man auch als „*anti*korreliert“, um das Vorzeichen herauszustreichen.

Im Bereich $|r| = 0.7 \dots 1.0$ spricht man von einer „starken Korrelation“, im Bereich $|r| = 0.4 \dots 0.7$ von einer „schwachen Korrelation“. Unter $|r| = 0.4$ ist die Korrelation normalerweise nicht befriedigend. Diese Grenzen sollte man aber nicht allzu absolut auffassen, es kommt eben auf die Ansprüche und auf sonstige Gegebenheiten an. Ausserdem ist bei sehr kleinen Datenmengen (sagen wir bis 5 Datenpunkte) die Korrelation fast immer gut und daher oft nur beschränkt aussagekräftig. Um es prägnant auszudrücken: bei nur 2 Datenpunkten gibt es natürlich immer eine Gerade, die perfekt passt.

Es gibt mehrere Definitionen, mit denen man einen Korrelations-Koeffizienten berechnen kann, doch am meisten verwendete ist der Korrelationskoeffizient metrischer Daten, der Pearson – Korrelationskoeffizient:

2.1.2.1 Metrische Daten: Pearson – Korrelationskoeffizient

Der mit Abstand bekannteste ist der „**Pearson– Korrelationskoeffizient**“ r (Produkt-Moment-Korrelation), der für metrische Daten Anwendung findet (wenn nichts anderes betont wird, ist meist wie selbstverständlich immer das Pearsonsche r gemeint). Es ist folgendermassen definiert:

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{s_{xy}}{s_x \cdot s_y} \end{aligned} \quad (2.2)$$

Hier sind n gleich der Anzahl Datenpaare (x_i, y_i) und

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}, \quad s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}$$

die Standardabweichungen s_x und s_y sowie s_{xy} die *Kovarianz*

$$s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}). \quad (2.3)$$

- Die Kovarianz ist positiv, wenn x und y einen monotonen Zusammenhang besitzen, d. h. hohe (niedrige) Werte von x gehen mit hohen (niedrigen) Werten von y einher.
- Die Kovarianz ist hingegen negativ, wenn x und y einen gegensinnigen monotonen Zusammenhang aufweisen, d. h. hohe Werte der einen Zufallsvariablen gehen mit niedrigen Werten der anderen Zufallsvariablen einher und umgekehrt.
- Ist das Ergebnis null, so besteht kein monotoner Zusammenhang zwischen x und y (nichtmonotone Beziehungen sind aber möglich).

Die Kovarianz gibt zwar die Richtung einer Beziehung zwischen zwei Variablen an, über die Stärke des Zusammenhangs wird aber keine Aussage getroffen. Um einen Zusammenhang vergleichbar zu machen, muss die Kovarianz normiert werden. Dies geschieht beim Pearson-schen Korrelationskoeffizienten durch Division durch die Standardabweichungen s_x und s_y . Pearsons r ist somit ein Mass für die Stärke der linearen Abhängigkeit zweier statistischer Variablen. In **R** kannst Du r mittels

```
cor(x,y,method="pearson")
```

oder einfacher

```
cor(x,y)
```

berechnen, wo **x** und **y** die (metrischen) Datenvektoren der unabhängigen und abhängigen Variablen sind

2.1.2.2 Ordinale Daten: Spearman – Korrelationskoeffizient

Der Spearman-Korrelationskoeffizient berechnet die Korrelation gemäss Gleichung (2.2) nicht zwischen den Daten selbst, sondern zwischen ihren *korrigierten Rängen*. Dies ist insbesondere bei ordinalen Daten wichtig, da man keine metrischen Grössenangaben ausnutzen kann, alles, was zur Verfügung steht, sind die **Ränge** der Daten.

Doch was ist ein Rang? Und was ein korrigierter Rang?

Aus dem Sport sind wir mit dem Begriff „Rang“ vertraut: 100 m - Lauf, Leichtathletik - Weltmeisterschaft 1991: Carl Lewis braucht 9.86 s, Rang 1; Leory Burrell: 9.88 s, Rang 2; Dennis Mitchel, 9.91 s, Rang 3.

Was wäre aber nun passiert, wenn Lewis und Burrell gleichzeitig in 9.86 s eingelaufen wären? Sie hätten sich den ersten Rang und somit die Goldmedaille geteilt. Und Mitchel hätte Bronze erhalten, immer noch Rang 3. In der Statistik spricht man von einer *Rangbindung*, wenn mehrere Werte sich den gleichen Rang teilen. Liegt eine solche Rangbindung, können wir den korrigierten Rang berechnen:

2.1.2.2.1 Berechnung des korrigierten Rangs mit `rank`

Hierzu bringt man die Daten zuerst in eine Rangfolge (beim 100 m - Lauf beim hypothetischen gleichzeitigen Einlauf von Lewis und Burrell: Ränge: 1,1,3). Liegen Rangbindungen vor, bestimmen wir den korrigierten Rang im Allgemeinen folgendermassen:

Wenn n Werte auf dem gleichen x . Rang landen, dann wird ihnen ein **korrigierter Rang** (auch Durchschnittsrang genannt) zugeteilt. Der korrigierte Rang wäre also

$$\frac{x + (x + n - 1)}{2}. \quad (2.4)$$

Im Sportbeispiel beim hypothetischen gleichzeitigen Zieleinlauf von Lewis und Burrell: $x = 1$, $n = 2$ und Lewis und Burrell erhielten beide jeweils den korrigierten Rang $(1 + (1 + 2 - 1))/2 = 1.5$.

Wert	9.86	9.86	9.91
Rang	1	1	3
korr. Rang	1.5	1.5	3

Wir wollen die Berechnung des korrigierten Rangs an einem weiteren Beispiel erläutern: Gegeben seien in einem Betrieb mit 19 Mitarbeitenden die Anzahl Krankheitstage der Mitarbeitenden im letzten Jahr:

0	0	8	7	0	14	2	0	18	3	12	8	0	20	3	9	1	14	50
---	---	---	---	---	----	---	---	----	---	----	---	---	----	---	---	---	----	----

Mit Gleichung (2.4) erhalten wir für die korrigierten Ränge:

Wert	0	0	8	7	0	14	2	0	18	3	12	8	0	20				
Rang	1	1	11	10	1	15	7	1	17	8	14	11	1	18				
korr. Rang	3	3	11.5	10	3	15.5	7	3	17	8.5	14	11.5	3	18				

Wert	3	9	1	14	50
Rang	8	13	6	15	19
korr. Rang	8.5	13	6	15.5	19

Mit dem R-Befehl `rank` können wir die korrigierten Ränge automatisch zu berechnen. In unserem Beispiel:

```
krankheitstage=c(0,0,8,7,0,14,2,0,18,3,12,8,0,20,3,9,1,14,50)
```

`rank(krankheitstage)` berechnet nun die korrigierten Ränge und liefert folgenden Output:

```
[1] 3.0 3.0 11.5 10.0 3.0 15.5 7.0 3.0 17.0 8.5 14.0 11.5 3.0 18.0 8.5 13.0 6.0 15.5 19.0.
```

Falls keine Rangbindungen vorliegen, können wir den Spearman - Koeffizienten direkt mit der Formel

$$r_s = 1 - \frac{6}{n \cdot (n^2 - 1)} \cdot \sum_{i=1}^n d_i^2 \quad (2.5)$$

berechnen. Hier sind n die Anzahl Datenpaare und d_i ist die Differenz der Ränge des i ten Daten-Paares (x_i, y_i) .

Falls Rangbindungen vorliegen, ist obige Formel jedoch nur eine Abschätzung und wir müssen die Korrelation der korrigierten Ränge mittels Gleichung (2.2) berechnen. Als Datenpaare müssen wir nun nicht die gemessenen Werte selbst, sondern die korrigierten Ränge einsetzen. Auch bei den Mittelwerten müssen wir nun die Mittelwerte der korrigierten Ränge einsetzen.

Viel einfacher geht das aber mit **R**: **R** berechnet den Spearmanschen - Korrelationskoeffizienten mittels

```
cor(x,y, method="spearman")
```

wo **x** und **y** die ursprünglichen Datenvektoren der gemessenen Werte sind. Die ganze (umständliche) Berechnung der korrigierten Ränge macht **R** im Hintergrund und wir müssen uns um nichts kümmern!

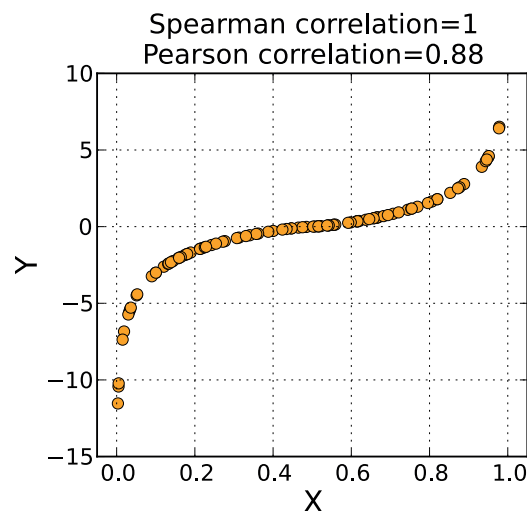


Abbildung 2.3: Wir erhalten eine Spearman-Korrelation von 1, wenn die Daten monoton steigend sind, auch wenn kein perfekter linearer Zusammenhang besteht.

Der Spearmansche Koeffizient muss bei ordinalen Daten, kann aber auch auf metrische Daten angewendet werden. Dazu ist zu bemerken, dass die Reduktion auf Ränge einen Informationsverlust bedeutet. Man weiss z.B. nicht mehr, wie weit voraus der erste Rang vor dem zweiten liegt (vgl. Abbildung (2.3)). In diesem Sinn ist der Spearmansche Koeffizient weniger aussagekräftig. Es gibt aber auch einen Vorteil, der es in manchen Fällen sogar empfehlenswert macht, den „normalen“ Pearsonschen Korrelationskoeffizienten durch den Spearmanschen - Rangkorrelationskoeffizienten zu ersetzen: Der Rangkorrelationskoeffizient ist robuster, reagiert also z.B. weniger stark auf Ausreisser.

R-Frage 10 *Wie gross ist der Pearsonsche Korrelations-Koeffizient der Baumdaten? Rechne es aus:*

```
cor(baum$Dicke,baum$Hoehe,method="pearson")
```

Das Resultat besagt, dass die Werte eine klare Tendenz zeigen, sich aber nicht gerade gut auf einer Geraden anordnen.

Berechne nun die Korrelation mit Spearmans-Korrelations-Koeffizient:

```
cor(baum$Dicke,baum$Hoehe,method="spearman")
```

Was fällt Dir auf?

R-Frage 11 *Lade von Moodle das Datenfile NOxEmissions.csv herunter (kein vorheriges Öffnen in EXCEL nötig). Lies die Webdokumentation zu diesem an der ZHAW generierten Datenfile unter <https://vincentarelbundock.github.io/Rdatasets/doc/robustbase/NOxEmissions.html> durch.*

Lies die Daten anschliessend in ein data.frame namens NOx ein. Berechne nun die Korrelation zwischen dem Logarithmus des stündlichen Mittelwertes der NOx Konzentration LNOx und der Summe der NOx Emissionen entlang der Autobahn LNOxEm in einer Stunde. Handelt es sich um eine starke oder schwache Korrelation?

Bestimme nun Steigung und y-Achsenabschnitt der Trendgeraden mit

```
trend=lm(NOx$LNOx~NOx$LNOxEm)
```

2.1.3 Bestimmtheitsmass r^2

Das Quadrat r^2 des Pearson-Korrelations-Koeffizienten (häufig auch als R^2 geschrieben) hat einen eigenen Namen: Das Bestimmtheits-Mass. Es ist ein weiteres, häufig verwendetes, Gütemass der linearen Regression. r^2 ist der Anteil der Variation der abhängigen Variablen y , der durch die lineare Regression erklärt wird, und liegt daher zwischen

0 (oder 0 %): kein linearer Zusammenhang und

1 (oder 100 %): perfekter linearer Zusammenhang.

Im Baum-Beispiel: $r = 0.6891873$ ergibt ein Bestimmtheitsmass von $r^2 = 0.47$. D.h. man kann sagen, dass die Baumhöhe (abhängige Variable) zu 47% durch die Dicke des Stammes festgelegt ist. Die anderen 53% sind auf andere Ursachen zurückzuführen, vielleicht auf den Schattenwurf durch andere Bäume oder die Qualität des Bodens oder wie sie geschnitten wurden oder ... oder die nicht ergründbaren Charaktere der einzelnen Bäume.

2.2 Nichtlineare Regression

Im folgenden benutzen wir die weltweit längste Messreihe von CO₂-Konzentrationen in der Erdatmosphäre als Beispiel. Quelle und Einzelheiten:

<http://co2now.org/Current-CO2/CO2-Now/noaa-mauna-loa-co2-data.html>.

R-Frage 12 Lade `co2.csv` von MOODLE herunter und schau dir den Datensatz in `R` an. U.a. indem du dir von `R` die zeitliche Entwicklung der CO_2 -Konzentration aufzeichnen lässt. Berechne anschliessend die lineare Regression, zeichne sie ins Diagramm hinzu, erzeuge dann den Residuen-Plot, und berechne schliesslich auch noch den Korrelations-Koeffizienten.

Wie gut ist die Regression gelungen? Die Gerade gibt eindeutig eine Tendenz gut wieder. Aber erstens scheinen die Daten eine Aufwärts-Krümmung zu besitzen, die eine Gerade natürlich nicht mitmachen kann. Zusätzlich ist in den Daten eine Schwingung überlagert.

Was würde besser passen? Versuchen wir die allgemeine Tendenz zu beschreiben, weiterhin ohne auf die Schwingungen Rücksicht zu nehmen: Wir könnten es mit einer quadratischen Funktion versuchen.

R-Frage 13 Damit sich die Befehle im Folgenden etwas kürzer und übersichtlicher gestalten, produzierst du vorerst Variablen mit kurzen Namen:

```
CO2=co2$concentration
t=co2$time
```

Funktion nochmal neu malen:

```
plot(t,CO2,type="l")
```

Dann erfinden wir eine Variable, die dem Quadrat von `t` entspricht:

```
t2=t^2
```

Nun wird dieses `t2` ins Modell eingebaut, als wäre es eine weitere Variable, die mit `t` nichts zu tun habe:

```
trend2=lm(CO2~t+t2)
```

Resultat Anschauen:

```
points(t,predict(trend2),type="l",col="blue")
```

Wie du siehst, wird die Tendenz durch die Parabel nun sehr gut repräsentiert.

Fehlt noch die auffällige saisonale Komponente. Der naheliegendste Versuch wäre, noch eine Sinus-Schwingung hinzuzunehmen. Diesmal können wir aber nicht einfach noch etwas zum linearen Modell hinzuaddieren, denn eine Sinus-Funktion enthält selber schon interne Parameter, die erst noch gefunden werden müssen. Die allgemeinste Form einer sinusförmigen Funktion lautet:

$$f(t) = d \cdot \sin(p \cdot t + e)$$

p (für Periode p) können (und sollten) wir so einstellen, dass die Schwingungsperiode einem Jahr entspricht. Dies gelingt mit $p = 2\pi/\text{Jahr}$. d und e hingegen bleiben Parameter, die noch zu finden sind. Während wir mit dem Trick t^2 zu einer Variablen `t2` zu machen, zu einem „pseudo-linearen Modell“ kamen, müssen wir zur zusätzlichen Beschreibung der Schwingung ein nicht-lineares Modell ins Auge fassen. Die Funktion `nls` (eine Abkürzung für „nonlinear

least squares“) erlaubt uns dies. Die gesamte Gleichung lautet nun:

$$CO_2(t) = b + a \cdot t + c \cdot t^2 + d \cdot \sin\left(\frac{2\pi}{\text{Jahr}} t + e\right),$$

wo die Zeit t in Jahren und die CO_2 -Konzentration in „parts per million“, abgekürzt ppm, gemessen wird. `nls` bestimmt die Koeffizienten a, b, c, d, e . Hierzu braucht `nls` Startwerte, von wo aus sie die Suche nach der Lösung beginnen kann. In unserem Fall reichen mehr oder weniger symbolische Werte:

```
guess=list(a=1,b=1,c=1,d=1,e=1)
```

R-Frage 14 *Plot der Daten:*

```
plot(t,CO2,type="l")
```

Und jetzt kommt die eigentliche (nicht-lineare) Regression:

```
trend3=nls(CO2~b+a*t+c*t^2+d*sin(t*2*pi+e),start=guess)
```

Plot der nichtlinearen Regression:

```
points(t,predict(trend3),type="l",col="blue")
```

Plotte im Anschluss auch noch das Resultat der linearen, quadratischen Regression in diese Figur. Was fällt Dir auf?

Hinweis: Der Befehl `points` zeichnet die Graphen über eine durch `plot` angefangene Figur. So können wir auf einfache Art und Weise mehrere Graphen in einer Abbildung darstellen.

Das sieht doch recht überzeugend aus, oder? Mit

```
summary(trend3)
```

wird (in der ersten Zahlenkolonne) sichtbar, wie gross die Parameter geschätzt werden. Wir erhalten für die CO_2 -Konzentration das Modell

$$CO_2(t) = 4.50 \cdot 10^4 \text{ ppm} - 46.43 \frac{\text{ppm}}{\text{Jahr}} \cdot t + 1.21 \cdot 10^{-2} \frac{\text{ppm}}{\text{Jahr}^2} \cdot t^2 + 2.81 \text{ ppm} \cdot \sin\left(\frac{2\pi}{\text{Jahr}} \cdot t - 0.38\right) \quad (2.6)$$

Gleichung (2.6) erlaubt die Berechnung der CO_2 -Konzentration im Jahr *soundso* unter Berücksichtigung der saisonalen Schwankungen (die Jahreszeit wird mit Kommastellen in t angegeben).

Aber Achtung: Man soll nicht zuweit hinaus extrapolieren. Es könnte ja z.B. sein, dass sich in der irdischen Politik noch etwas tut ...

2.2.1 Nullhypothese, p-Wert und `summary` Befehl

Oft interessiert uns bei einer Regression, ob die berechneten Koeffizienten signifikant von 0 verschieden sind. Doch was bedeutet signifikant? Ein Mass dafür ist der sogenannte p-Wert (p für „probability“). Je kleiner der p-Wert ist, desto wahrscheinlicher ist es, dass der Koeffizient nicht nur zufällig bei unserem Versuch so herausgekommen ist, sondern tatsächlich von 0 verschieden ist. **Der p-Wert ist die Wahrscheinlichkeit der Nullhypothese.** Doch was ist die Nullhypothese? Bei einer linearen Regression lautet die Nullhypothese: Es gibt **keinen** linearen Zusammenhang zwischen Variablen abhängigen Variablen y und der unabhängigen Variablen x . Bei der linearen Regression ist die Nullhypothese somit genau das Gegenteil von dem, was wir eigentlich zeigen wollen, nämlich einen linearen Zusammenhang. Je kleiner die Wahrscheinlichkeit der Nullhypothese (= der p-Wert) ist, umso wahrscheinlicher ist es, dass es einen linearen Zusammenhang gibt. Folgende Nomenklatur hat sich eingebürgert:

$$\begin{aligned} p \leq 0.05 &= 5\% && * \text{ signifikantes,} \\ p \leq 0.01 &= 1\% && ** \text{ sehr signifikantes,} \\ p \leq 0.001 &= 0.1\% && *** \text{ hoch signifikantes Ergebnis} \end{aligned}$$

Ist z.B. der p-Wert der Steigung a kleiner als 0.05, so besteht eine Abhängigkeit zwischen abhängiger Variable y und unabhängiger Variable x . Der `summary(name_des_modells)` zeigt uns die p-Werte an:

R-Frage 15 *Wir nehmen wieder den `baum.csv`-Datensatz als Beispiel. Lade den Datensatz nochmal, falls du ihn nicht mehr im Workspace hast.*

Wiederhole dann die lineare Regression:

```
plot(baum$Dicke,baum$Hoehe)
trend=lm(baum$Hoehe~baum$Dicke)
abline(trend,col="blue")
coefficients(trend)
```

Soweit alles wie gehabt. Wenn du nun den Befehl

```
summary(trend)
```

*eingibst, kommt einiges an weiteren Informationen zum Vorschein. U.a. steht etwa auf halber Höhe eine Tabelle, deren rechte Spalte (`Pr(>|t|)`), was t bedeutet, werden wir in den kommenden Lektionen lernen) die p-Werte enthält. Der p-Wert $2.59 \cdot 10^{-6}$ bedeutet, dass der y -Achsen Schnittpunkt b hoch signifikant von 0 verschieden ist. Auch a ist signifikant von 0 verschieden, aber nur mit $P = 0.0275$. Ausserdem findest Du **Multiple R-squared** von 0.475, hier handelt es sich um das Bestimmtheitsmass r^2 . Probiere es mir dem `cor`-Befehl aus.*

Übrigens: Die quadratische Regression ist nichts weiter als ein Spezialfall von „non-linear-least squares“. Probiere es aus!

R-Frage 16 Führe die quadratische Regression am CO_2 Beispiel (R-Frage 13) mit den Startwerten

```
guess2=list(a=1,b=1,c=1)
und
nls(CO2~b+a*t+c*t^2,start=guess2)
```

durch. Vergleiche das Ergebnis mit dem Resultat aus R-Frage 13.

2.3 Multivariate Regression

Oftmals hängt eine Messgröße nicht nur von einer Variablen, sondern von mehreren Variablen ab. Kommen wir zurück zum Baumbeispiel. Jetzt benutzen wir den vollständigen Baumdatensatz `trees`, der in R bereits als Anschauungsmaterial integriert ist.

R-Frage 17 Mit

```
View(trees)
```

kannst du dir den Datensatz ansehen, und mit

```
?trees
```

bekommst du weitere Informationen darüber. `Girth` ist der Baum-Durchmesser in Zoll (Englisch: *inch*). Es gilt: 1 *inch* = 2.54 cm.

Diesmal haben wir es mit drei Variablen zu tun: Stammdurchmesser (`Girth`), Stammhöhe (`Height`) und Stammvolumen (`Volume`). Es ist zu erwarten, dass das Volumen sowohl mit der Höhe als auch mit dem Durchmesser zunimmt. Wir setzen die Sache also so auf:

- Durchmesser und Höhe sind unabhängige Variablen (obwohl sie natürlich untereinander miteinander verknüpft sind).
- Das Volumen ist die abhängige Variable.

Es gilt somit eine Funktion zu finden, die das Volumen als eine Funktion des Durchmessers und der Höhe beschreibt: $\text{Volumen} = f(\text{Durchmesser}, \text{Höhe})$ (oder Englisch, wie in unserem Datensatz: $\text{Volume} = f(\text{Girth}, \text{Height})$). `plot` bringt dir ganz einfach und schnell einen Überblick über die paarweisen Abhängigkeiten:

R-Frage 18 Mit

```
plot(trees)
```

bekommst du auf einen Schlag alle Plots, die gemäss Kombinatorik möglich sind.

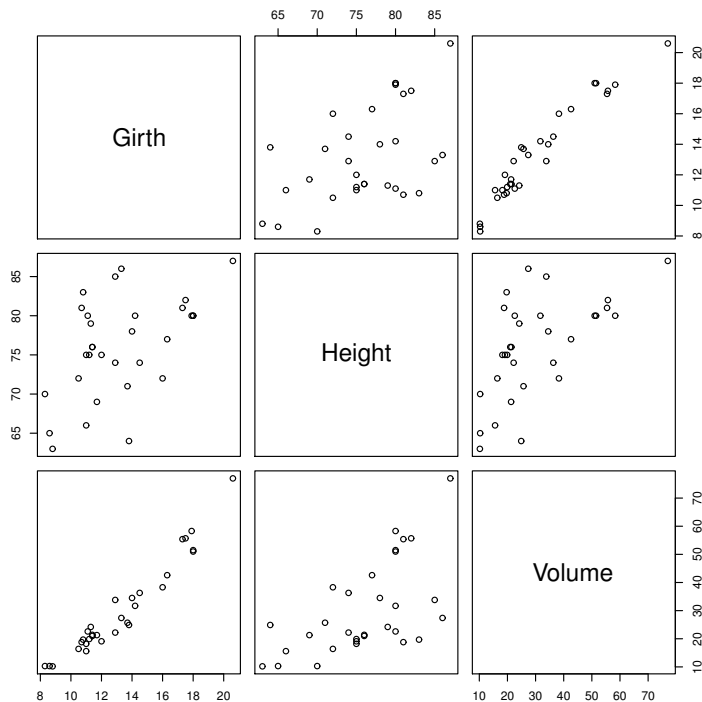


Abbildung 2.4: Scatterplots.

Lesebeispiel: Diagramm oben Mitte (1. Zeile, 2. Spalte): Auf der y-Achse ist der Baumdurchmesser (Girth in inch), auf der x-Achse die Höhe in Fuss aufgetragen. Das Gleiche mit vertauschten Achsen finden wir in der zweiten Zeile, links (2. Zeile, 1. Spalte).

Wie ist diese Anordnung von Abbildungen zu verstehen? In den Diagonalelementen ist jeweils eine der drei Variablen angeschrieben. Das bezieht sich jeweils sowohl auf die entsprechende Zeile als auch auf die entsprechende Spalte. Die Skalen machen klar, was nach oben und was nach rechts aufgetragen ist: Der Stammdurchmesser (Skala der x-Achse aller Diagramme der 1. Spalte und y-Achse aller Diagramme der 1. Zeile) liegt zwischen 8.3 inches und 20.6 inches, die Stammhöhe (Skala der x-Achse aller Diagramme der 2. Spalte und y-Achse aller Diagramme der 2. Zeile) zwischen 63 Fuss und 87 Fuss (Englisch: feet, abgekürzt ft, 1 ft = 0.3048 m) und das Stammvolumen (Skala der x-Achse aller Diagramme der 3. Spalte und y-Achse aller Diagramme der 3. Zeile) zwischen 10.2 ft^3 und 77 ft^3 .

Nehmen wir nun exemplarisch das Diagramm links unten (3. Zeile, 1. Spalte). Es gehört einerseits zur **Volume**-Zeile andererseits zur **Girth**-Spalte. Im Diagramm rechts oben sieht man das Gleiche (1. Zeile, 3. Spalte), aber mit vertauschten Achsen.

Von Auge erkennt man einen deutlichen, linearen Zusammenhang zwischen Volumen und Durchmesser (Girth). Zwischen Volumen und Höhe sieht es auch danach aus, aber wesentlich weniger deutlich. Noch weniger deutlich ist der Zusammenhang zwischen Durchmesser und Höhe.

Genau so einfach bringt `cor` alle paarweisen Korrelations-Koeffizienten auf den Bildschirm:

R-Frage 19 *Probiere aus:*

```
cor(trees)
```

Passt das Resultat zum optischen Eindruck?

Wahrscheinlich wirst du interessante Befunde etwas genauer anschauen wollen, als oben mit `plot(trees)`. `cor(trees)` zeigte uns, dass der grösste Korrelationskoeffizient zwischen den Variablen `Volume` und `Girth` besteht. Deswegen führen wir in einem ersten Schritt eine lineare Regression zwischen den Variablen `Volume` und `Girth` durch:

R-Frage 20 *Führe eine lineare Regression zwischen `Volume` und `Girth` aus. Nimm an, dass `Volume` die abhängige Variable, `Girth` die unabhängige Variable ist. Plote zuerst die Daten*

```
plot(trees$Girth,trees$Volume)
```

und bestimme anschliessend y-Achsenabschnitt und Steigung der zugehörigen Geradengleichung mit

```
univarlm=lm(trees$Volume~trees$Girth)
```

Zeichne anschliessend die Trendlinie mit

```
abline(univarlm,col="blue")
```

ein.

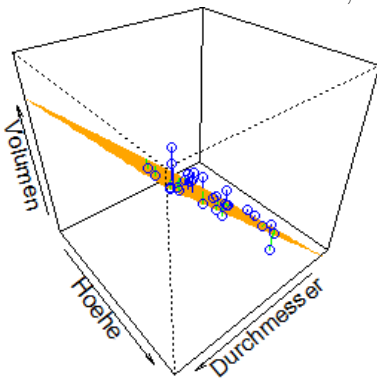
Aber eigentlich wollten wir ja das Volumen als Funktion f zweier unabhängige Variablen (`Girth` und `Height`) modellieren

$$\text{Volume} = f(\text{Girth}, \text{Height})$$

an die Daten anzupassen versuchen. Im einfachsten Fall ist dies wieder eine lineare Funktion, d.h. etwas von der Form

$$\text{Volume} = a \cdot \text{Girth} + b \cdot \text{Height} + c .$$

So wie bei einer einzigen unabhängigen Variablen eine Gerade an die Datenpunkte angepasst wird, ist es diesmal eine Ebene. Das ist nicht mehr so leicht darzustellen – der Versuch, den ich trotzdem unternommen habe, hat zu diesem Resultat geführt:



Die eigentliche Analyse führen wir wieder mit `lm` aus, da wir ja einen linearen Zusammenhang zwischen Durchmesser und Volumen und Höhe und Volumen annehmen.

R-Frage 21 *Führe eine multivariate, lineare Regression am das trees-Beispiel aus:*

```
multivarlm=lm(trees$Volume~trees$Girth+trees$Height)
multivarlm$coefficients
```

Jetzt kannst du die Funktionsgleichung ablesen:

$$\text{Volume} = 4.7 \cdot \text{Girth} \frac{\text{ft}^3}{\text{inch}} + 0.3 \cdot \text{Height} \text{ ft}^2 - 58.0 \text{ ft}^3 .$$

Beachte, dass Girth in inch, Height in ft und Volume in ft^3 im unserem Datensatz angegeben waren.

Weitere Informationen fördert `summary` zu Tage:

R-Frage 22 *Interpretiere die p-Werte, die*

```
summary(multivarlm)
```

ausspuckt. Unterscheiden sich die Koeffizienten signifikant von 0?

R-Frage 23 *Führe nun eine multivariate, lineare Regression mit den NOx - Daten aus R-Aufgabe 11 durch. Nimm an, dass LNOx die abhängige Variable, LNOxEm und die Wurzel aus der Windgeschwindigkeit sqrtWS die unabhängigen Variablen sind. Schau Dir das Resultat mit dem `summary`-Befehl an. Vergleiche das Resultat mit dem Ergebnis aus R-Aufgabe 11. Welches Modell beschreibt die Daten besser? Begründe Deine Aussage.*

CHECKLISTE

Kannst du jetzt:

- mit R y -Achsenabschnitt und Steigung einer linearen Regression bestimmen?
- mit R den korrigierten Rang bei Rangbindungen bestimmen?
- die Güte einer linearen Regression bestimmen?
- den Unterschied zwischen Pearsonschem und Spearmanschem Korrelationskoeffizienten erklären?
- mit R den Pearsonschen und Spearmanschen Korrelationskoeffizienten berechnen?
- definieren, was eine starke und schwache Korrelation ist?
- eine nichtlineare Regression am Beispiel der CO₂-Daten ausführen?
- eine multivariate, lineare Regression ausführen?

Kapitel 3

Schliessende Statistik

Daten sammeln und beschreiben ist wichtig und führt oftmals schon ohne weitere weitere mathematische „Verarbeitung“ zum Erkenntnisgewinn. So ist die Zusammenstellung der Studierenden-Zahlen der *aw* für deren Leitung essenziell (z.B. für die Berechnung von Einnahmen und Ausgaben, um einen Eindruck der Entwicklung zu bekommen, um diese gegenüber Behörden und Öffentlichkeit auszuweisen usw.). Soweit ist das beschreibende Statistik.

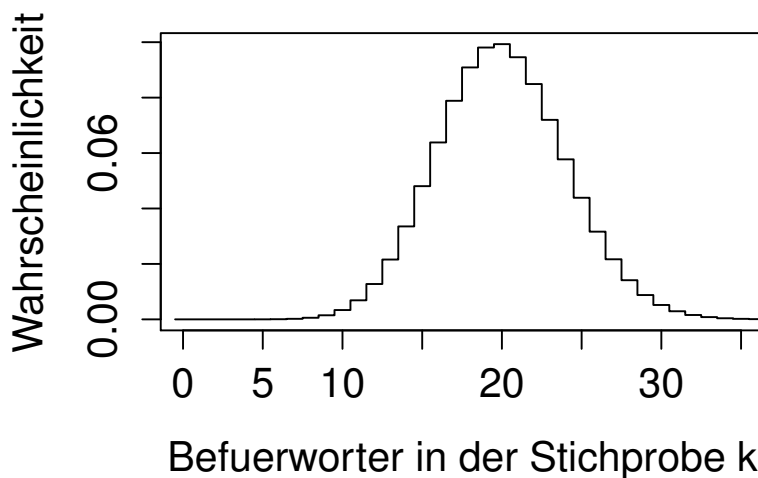
Aber die Daten inspirieren meist auch zu *Schlussfolgerungen und Prognosen, die über den zu Grunde liegenden Datensatz hinaus gehen*. Beispielsweise könnte die Leitung der *aw* an der zukünftigen Entwicklung der Studierenden-Zahlen interessiert sein. Oder an der Frage, ob der Anteil der Frauen unter den Studierenden definitiv zunimmt. Andere mögliche Fragestellungen könnten lauten: Gibt es eine Entwicklung zu immer jüngeren Studierenden? Hat die Zahl der Studien-Anfänger etwas mit der Konjunktur zu tun? Hier setzt die „schliessende Statistik“ (auch „beurteilende“, „deduktive“ oder „Inferenz –Statistik“ genannt) ein: welche Daten lassen welche Schlüsse zu? Insbesondere: wie sicher sind die gezogenen Schlüsse? Schliessende Statistik wäre nämlich nicht schliessende Statistik, wenn die Aussagen 100%ige Sicherheit hätten.

Schliessende Statistik braucht es weniger, wenn man eine Vollerhebung zur Verfügung hat – dann hat man ja eben *alle* Information bereits in Händen. Aber Vollerhebungen sind in der Praxis eher die Ausnahme. Wenn man eine Umfrage über die Zufriedenheit der Kunden macht oder wenn man das Wachstum einer neuen Tomatensorte testet, erreicht man nie alle Kunden (schon gar nicht die zukünftigen), geschweige denn alle Tomaten dieser Welt. Trotzdem leitet man eine Idee ab, die sich auf *alle* Kunden, bzw. *alle* Tomaten bezieht. Das Stichprobenwesen ist der zentrale Ausgangspunkt der schliessenden Statistik.

3.1 Das Problem der Stichproben

Dieser Abschnitt soll die Problematik des Stichprobenwesens an einem Beispiel erläutern. Angenommen, du willst im Auftrag des „Sonntagsanzeigers“ herausfinden, ob die Schweizer Bevölkerung für oder gegen die Initiative „Strom ohne Atom“ ist. Nehmen wir weiter an, dass 20% dafür und 80% dagegen sind (was du aber „eigentlich“ noch nicht weisst, son-

dern eben gerade herausfinden sollst). Um den Auftrag des „Sonntagsanzeigers“ zu erfüllen, befragst Du 100 Leute auf der Strasse. Wie viele „für“ und wie viele „dagegen“-Stimmen wirst du einsammeln? Das wahrscheinlichste Resultat ist tatsächlich 20 für und 80 Gegenstimmen. Das ist von allen möglichen Resultaten dasjenige, das deine Stichprobe mit der grössten Wahrscheinlichkeit ergeben wird. Aber *nur das wahrscheinlichste, nicht das einzige mögliche*. Je nach dem, wer dir gerade so entgegen schlendert, kann das Resultat gut auch z.B. 15/85 sein. Die Wahrscheinlichkeitsrechnung sagt uns, welches Resultat wie wahrscheinlich – oder wie unwahrscheinlich – ist:



Hier sind also Wahrscheinlichkeiten aufgetragen. Das ist eine „Wahrscheinlichkeits-Verteilung“. Das ist gewissermassen das theoretische Gegenstück zur gemessenen Häufigkeits-Verteilung. Man könnte sagen, dass schliessende Statistik zu einem grossen Teil aus dem Vergleich von Häufigkeits- und Wahrscheinlichkeits-Verteilungen besteht.

Frage 4 Was bedeutet z.B. der Punkt der Kurve bei $k = 15$?

In der Tabelle sind noch einige Zahlen aufgelistet, die der Figur entsprechen:

Befürworter in der Stichprobe	Wahrscheinlichkeit
0	$2.0 \cdot 10^{-8} \%$
5	0.0015%
10	0.34%
15	4.81%
18	9.09%
20	9.93%
22	8.49%
25	4.39%
30	0.52%
\vdots	\vdots
100	$1.3 \cdot 10^{-68} \%$

Es kommen auch extrem „falsche“ Stichproben-Resultate vor, diese aber dafür auch nur extrem selten.

Stichproben zeichnen sich durch folgende Eigenschaften aus:

- Die Stichprobe gibt uns mit der grössten Wahrscheinlichkeit das richtige Resultat.
- Die Stichprobe kann aber auch – mit nicht viel kleinerer Wahrscheinlichkeit – ein leicht falsches Resultat liefern.
- Die Stichprobe kann – mit allerdings nur winziger Wahrscheinlichkeit – ein völlig falsches Resultat liefern.

Dieses Prinzip lässt sich nicht aus der Welt schaffen. Aber die zahlenmässige Ausprägung des Prinzips lässt sich sehr wohl beeinflussen, nämlich mit dem Umfang der Stichprobe. Würdest du nicht nur 100 Personen befragen, sondern deren 10 000, sähen die Wahrscheinlichkeiten so aus:

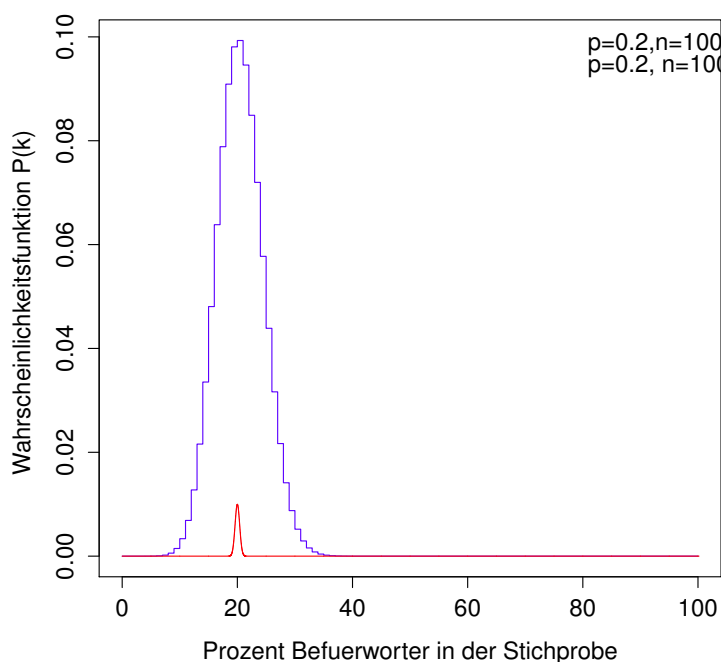


Abbildung 3.1: Wahrscheinlichkeit $P(k)$ als Funktion der prozentualen Befürworter.

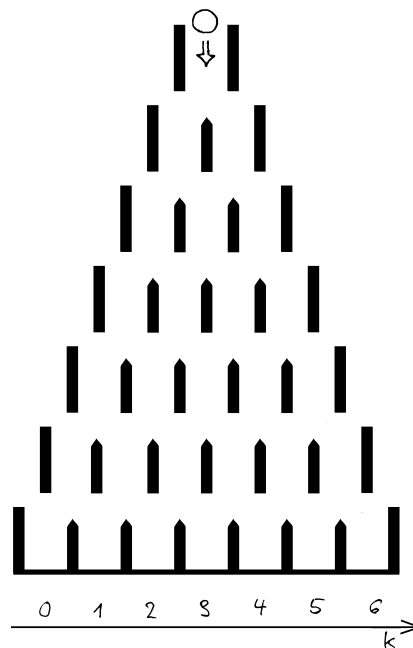
Die blaue Kurve entspricht der Kurve auf S. 26 (100 Befragte) mit veränderter x-Achse - jetzt haben wir den prozentualen Anteil der Befürworter und nicht die Anzahl der Befürworter aufgetragen. Die schmalere Kurve bezieht sich auf eine Umfrage, an der 10 000 Leute teilnahmen.

3.2 Binomialverteilung

Schliessende Statistik beruht letztlich immer auf Wahrscheinlichkeits-Rechnungen. Eben z.B.: mit welcher Wahrscheinlichkeit zieht man bei einer Stichprobe dies oder jenes aus der Gesamtheit heraus? Wahrscheinlichkeitsrechnung ist ein eigenes (recht umfangreiches) Kapitel der Mathematik, das im Rahmen von „Big Data“ immer mehr an Bedeutung gewinnt. Im Rahmen der Veranstaltung „Schliessende Statistik“ ist es aufgrund des begrenzten Zeitbudgets nicht möglich, die beschreibende Statistik von der Wahrscheinlichkeitsrechnung her sauber herzuleiten. Doch anhand der Wahrscheinlichkeits-Verteilung, die schon im Umfrage-Beispiel (S. 26) auftauchte, der sogenannten „Binomialverteilung“, wollen wir exemplarisch den Zusammenhang zwischen zugrunde liegender Verteilung und Wahrscheinlichkeit erläutern. Schon auf S. 26 fiel dir vielleicht die Ähnlichkeit mit der Normalverteilung auf.

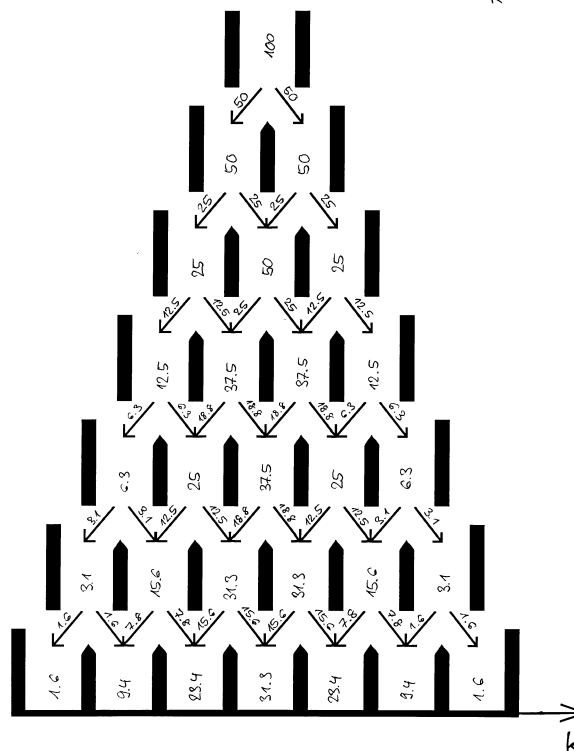
3.2.1 Kugelbrett

Denke dir folgendes Glücksspiel: man lässt eine Kugel eine schiefe Ebene hinunterrollen, auf der sie immer wieder auf Hindernisse trifft, wo sie nach links oder nach rechts ausweichen muss. Bei jedem Hindernis habe die Kugel exakt 50% Chance, nach rechts auszuweichen, und 50% Chance, nach links auszuweichen. Man kann Wetten abschliessen, wo die Kugel zuletzt landet. Du mit deiner höheren Mathematik-Bildung wirst deine Wetten natürlich erst verlieren, nachdem du dir die Wahrscheinlichkeiten sorgfältig überlegt hast.

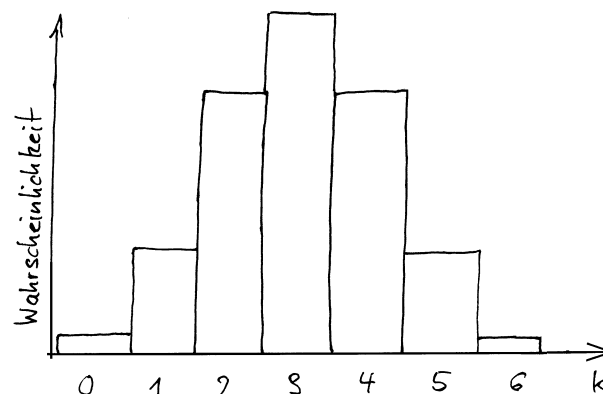


Du könntest die Sache z.B. so analysieren:

Du denkst dir (z.B.) 100 Versuche und überlegst dir die jeweiligen Wege, die sich gemäss den Wahrscheinlichkeiten ergeben. Das führt direkt auf die Wahrscheinlichkeitsverteilung (in Prozent).



Das ist eine Binomialverteilung:



(Eine Demonstration des Kugelbrett-Versuchs mit **R** findest auf MOODLE im Skript [kugelbrett.R](#). Das Skript enthält an zwei Stellen den Befehl `Sys.sleep(...)`. Dieser Befehl baut Pausen in den Programmablauf ein – je grösser das Argument, desto länger die Pause. Die Pausen werden benötigt, weil das Programm sonst so schnell abläuft, dass du gar nichts mehr siehst. Leider muss man die Pausen je nach Rechner anders einstellen. Du musst also etwas experimentieren mit den Zahlenwerten.)

Diese Binomialverteilung lässt sich auch mittels einer Formel darstellen, so dass du nicht immer ein Schema wie oben malen musst. Als Vorbereitung ein paar Begriffe:

- Die Binomialverteilung ist das Ergebnis von Serien („Bernoulli-Ketten“) immer gleicher Zufalls-„Ereignisse“, bei denen eine Entscheidung zwischen zwei Möglichkeiten nach einem Zufallsprinzip gefällt wird. Bei unserem bisherigen Beispiel besteht ein Ereignis aus einer einzelnen Entscheidung der Kugel, ob sie sich nach rechts oder nach links ablenken lässt.
- Eine solche Serie von Ereignissen nennen wir „Kette“. In unserem Beispiel: *eine* hinunterrollende Kugel.
- Von den beiden Möglichkeiten, die bei einem Ereignis zur Auswahl stehen, bezeichnet man eine als „Erfolg“ und die andere als „Misserfolg“. Als Erfolg bezeichnen wir also z.B. eine Ablenkung nach rechts, als Misserfolg eine Ablenkung nach links. (Die Zuteilung Erfolg \longleftrightarrow Misserfolg ist willkürlich. Die umgekehrte Wahl funktioniert genau so gut, wenn man alle folgenden Schritte konsequent anpasst.)
- Bei jedem Ereignis ist die Wahrscheinlichkeit für Erfolg gleich. Die Wahrscheinlichkeit für Erfolg muss nicht gleich der Wahrscheinlichkeit für Misserfolg sein, kann aber, wie im Falle der Kugeln gleich gross sein, nämlich 0.5.

⊗ Notation: Im Zusammenhang mit der Binomialverteilung benutzen wir folgende Notation:

- n immer die Anzahl der Ereignisse in einer Kette,
- k die Anzahl Erfolge,
- p die Erfolgswahrscheinlichkeit im einzelnen Ereignis,
- und P die Gesamtwahrscheinlichkeit einer Kette für k Erfolge aus n Ereignissen. Die Funktion $P(k)$ ist dann die Binomialverteilung.

Frage 5 *Wie gross sind im Beispiel mit den herunterrollenden Kugeln n , k und p ?*

Die Wahrscheinlichkeiten für die möglichen Gesamtergebnisse lassen sich dann folgendermassen berechnen:

DEFINITION DER WAHRSCHEINLICHKEITSFUNKTION DER BINOMIALVERTEILUNG:

SEI DIE WAHRSCHEINLICHKEIT FÜR ERFOLG IM EINZELNEN EREIGNIS (KLEIN!) p ; SEI DIE ANZAHL EREIGNISSE PRO KETTE n UND k DIE ANZAHL ERFOLGE. DANN BERECHNET SICH DIE WAHRSCHEINLICHKEIT (GROSS!) P FÜR k ERFOLGE NACH DER FORMEL,

$$\begin{aligned} B(k|p, n) = P(k) &= \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k} = \\ &= \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} . \end{aligned} \quad (3.1)$$

DAS IST DIE WAHRSCHEINLICHKEITSFUNKTION DER „BINOMIALVERTEILUNG“.

In Gleichung (3.1) kommen Zeichen vor, die Du bereits in der Mathematik kennengelernt hast:

(ξ) Notation: Die „Fakultät“ $n!$ einer natürlichen Zahl n ist das Produkt aller natürlichen Zahlen von 1 bis n :

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 2 \cdot 1 . \quad (3.2)$$

Für $n = 1$ und $n = 0$ ist die Fakultät

$$0! = 1! = 1 . \quad (3.3)$$

(ξ) Notation: $\binom{n}{k}$ ist eine Abkürzung für den Bruch mit den Fakultäten

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} \quad (3.4)$$

und wird als „Binomial-Koeffizient“ bezeichnet und „n über k“ gelesen. Beachte die Spezialfälle

$$\binom{n}{n} = \binom{n}{0} = 1 \quad (3.5)$$

Der Binomalkoeffizient gibt an, auf wie viele verschiedene Arten man k Objekte aus einer Menge von n verschiedenen Objekten auswählen kann (ohne Zurücklegen, ohne Beachtung der Reihenfolge). Der Binomalkoeffizient ist also die Anzahl der k -elementigen Teilmengen einer n -elementigen Menge. In **R** kann er einfach mittels `choose(n,k)` berechnet werden. In unserem Kugelbeispiel ist der Binomalkoeffizient der kombinatorische Faktor, der angibt, auf wievielen verschiedenen Wegen, wir bei n Wiederholungen des Experimentes (Kugel wird nach rechts oder links abgelenkt) k Erfolge (Rechts-Ablenkung) erzielen können. $p^k \cdot (1-p)^{n-k}$ ist die Wahrscheinlichkeit bei einmaligem Durchführen des Experimentes k mal Erfolg zu haben.

Die Herleitung der Formel für die Binomialverteilung Gleichung (3.1) ist nun gar nicht so schwierig zu verstehen. Sagen wir, eine Kugel muss $n = 6$ mal zwischen links und rechts entscheiden. Wie bisher definieren wir als Erfolg, dass die Kugel nach rechts abgelenkt wird. Die Wahrscheinlichkeit für rrrlll (also erst 3 mal rechts (r), dann 3 mal links (l)) ist $p \cdot p \cdot p \cdot (1-p) \cdot (1-p) \cdot (1-p) = p^3 \cdot (1-p)^{6-3}$ wegen des Multiplikationsgesetzes für unabhängige Ereignisse. 3 mal rechts könnte aber auch sein: rrlrl, rrlrl, etc. Insgesamt gibt es $\binom{n=6}{k=3} = 20$ dieser Möglichkeiten. Damit ist die Wahrscheinlichkeit für 3 mal rechts gleich der angegebenen Formel.

Frage 6 *Rechne mit Gleichung 3.1 die Wahrscheinlichkeit nach, die laut der Abbildung auf S. 29 für eine Ankunft der Kugel in der Mitte bestehen soll.*

Für $p = 1/2$ ist die Binomialverteilung symmetrisch (siehe linke Figur in Abbildung 3.2), offenbar wandert der Hauptteil der Verteilung für grössere n immer weiter nach rechts und die Verteilung wird immer breiter. Ist $p \neq 1/2$ ist die Verteilung schief (rechte Figur in Abbildung 3.2).

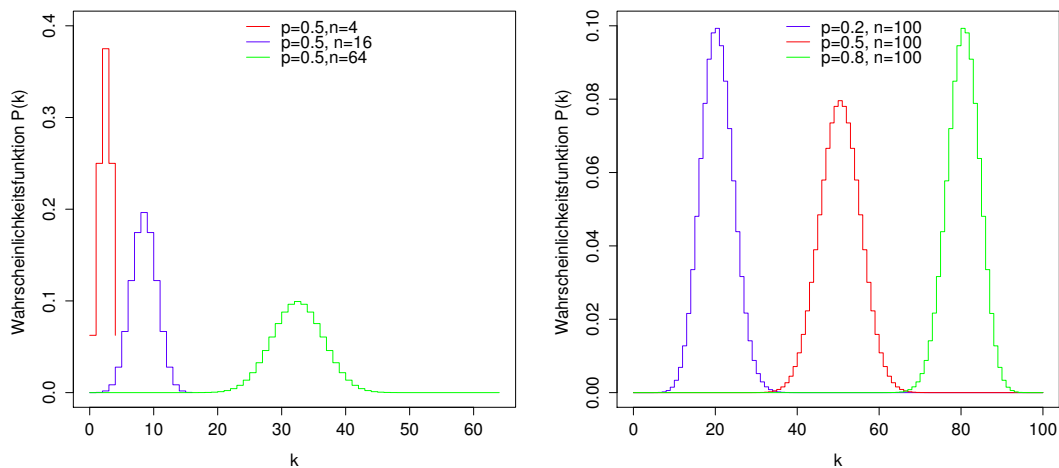


Abbildung 3.2: Links: Binomialverteilung für $p = 0.5$ und variierende n . Rechts: Binomialverteilung für $n = 100$ und variierende p .

3.2.2 Binomialverteilung in R

In **R** können wir Gleichung (3.1) mit dem **R** - Befehl

```
dbinom(k,n,p)
```

auswerten. `dbinom` berechnet die Wahrscheinlichkeitsfunktion $P(k)$ der Binomialverteilung für gegebene k , n und p . Analog berechnet `dnorm(x, mean=mu, sd=sigma)` die Höhe der Wahrscheinlichkeitsdichtefunktion (englisch: **density**) der Normalverteilung mit Mittelwert `mu` und Standardabweichung `sigma` an der Stelle x .

3.2.3 Lage und Streuung einer Binomialverteilung

Wie bei einer (gemessenen) Häufigkeitsverteilung kann man auch bei einer (theoretischen) Wahrscheinlichkeitsverteilung wie der Binomialverteilung Mittelwert und Standardabweichung angeben. Sie folgen im Fall der Binomialverteilung aus n und p :

EINE BINOMIALVERTEILUNG FÜR OBJEKTE MIT n EREIGNISSEN MIT ERFOLGSAHRSCHENLICHKEIT p HAT DEN MITTELWERT

$$\mu = n \cdot p, \quad (3.6)$$

UND DIE STANDARDABWEICHUNG

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)}. \quad (3.7)$$

Beachte, dass die relative Breite

$$\frac{\sigma}{\mu} = \frac{\sqrt{1-p}}{\sqrt{n \cdot p}} \sim \frac{1}{\sqrt{n}} \quad (3.8)$$

mit zunehmenden n immer kleiner wird: das Verhältnis aus Standardabweichung (ein Mass für die Breite der Verteilung) und dem Mittelwert der Stichprobe ist nach Gleichung 3.8 umgekehrt proportional zur Wurzel des Stichprobenumfanges: Je mehr Versuche man macht (grösseres n), desto kleiner wird die Wahrscheinlichkeit für eine grosse prozentuale Abweichung vom Mittelwert. Man nennt dies das Gesetz der grossen Zahlen. Dies erklärt, warum in Abbildung 3.1 die Standardabweichung der Breite der Wahrscheinlichkeitsverteilung des prozentualen Anteils der Befürworter bei 100 Befragten 4% ist, bei 10 000 Befragten aber nur 0.4% ist.

o! *Merke!*

Grundsätzlich ist σ ein universelles Streumass und damit eine Eigenschaft, die man bei jeder Verteilung angeben kann. Du hast im letzten Semester gelernt,

- wie man s (das Pendant zu σ ist, falls die Grundgesamtheit nicht bekannt ist) aus den Einzelwerten einer Stichprobe berechnet;
- wie man σ in der grafischen Darstellung einer Häufigkeitsverteilung ungefähr schätzen kann;
- σ im Zusammenhang mit der Normalverteilung als Masseinheit zu brauchen (\rightarrow z -Wert).

Soeben hast Du wiederholt, wie man

- σ aus den Modellannahmen zu berechnen, falls ein Binomialverteilungs-Modell angebracht ist.

3.3 Zusammenhang zwischen Binomial- und Normalverteilung

Es ist dir vielleicht schon aufgefallen, dass die Binomialverteilung eine gewisse Ähnlichkeit mit der Normalverteilung hat. Voraussetzung ist, dass sie in Lage und Streuung (μ und σ) übereinstimmen. Diese Ähnlichkeit ist umso besser, je grösser das n der Binomialverteilung ist und je näher die Wahrscheinlichkeit des Erfolges p der Binomialverteilung bei 0.5 liegt.

Beispiel: Die Treppenkurve in Abbildung 3.3 stellt die Binomialverteilung zu $n = 20$ und $p = 0.5$ dar:

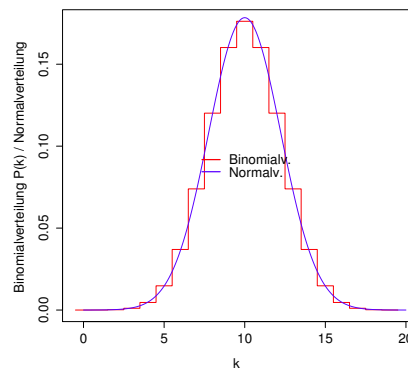


Abbildung 3.3: Rot: Binomialverteilung für $n = 20$ und $p = 0.5$. Blau: Normalverteilung. Für $p = 0.5$ und grosse n geht die Binomialverteilung in die Normalverteilung über. In unserem Beispiel hat die Normalverteilung einen Mittelwert $\mu = n \cdot p$ und eine Standardabweichung $\sigma = \sqrt{n \cdot p \cdot (1 - p)}$.

Sie wird in **R** mittels

```
dbinom(0:20,20,0.5)
```

berechnet. Als nächstes berechnet man Mittelwert und Standardabweichung dieser Verteilung:

```
mu=p*n
sigma=sqrt(n*p*(1-p)).
```

Die zugehörige Normalverteilung (oben eingezeichnet) kann mittels `dnorm(0:20,mean=mu,sd=sigma)` berechnet werden. Manche der statistischen Verfahren, die uns noch begegnen werden, bauen auf dieser Approximation auf.

R-Frage 24 *Rechne die Wahrscheinlichkeit nach, die laut der Abbildung auf S. 29 für eine Ankunft der Kugel in der Mitte bestehen soll.*

In den nachfolgenden Aufgaben solltest du jeweils in zwei Schritten vorgehen:

1. Überlege dir zuerst, welcher Zusammenhang zwischen der vorliegenden Situation und dem „Binomialmodell“ besteht – was ist ein *ein* Ereignis? Wie oft wird dieses Ereignis wiederholt? Was ist ein Erfolg? Was die Wahrscheinlichkeit für Erfolg?

Notiere dir das Resultat und ordne die Zahlen den richtigen Formelzeichen zu.

2. Nun kommt der meist einfachere Teil: die konkret gesuchte Grösse ausrechnen.

Frage 7 *Die Dozentin mag dieses Jahr keine Prüfungen korrigieren und macht deshalb die Noten mit dem Würfel.*

- a) *Wie gross ist die Wahrscheinlichkeit, in 5 Prüfungen 5 Mal einen 6er zu würfeln?*
- b) *Wie gross ist die Wahrscheinlichkeit, in 5 Prüfungen 5 Mal keinen 6er zu schiessen?*

R-Frage 25 a) *Rechne nach: wie wahrscheinlich ist im Beispiel auf S. 26 („Strom ohne Atom“, 20% dafür, 80% dagegen, 100 Befragte) eine Stichprobe mit 10% ja-Stimmen?*

- b) *Versuche nun mit **R** die Zahlenwerte der Tabelle auf Seite 27 zu reproduzieren (Umfrage „Strom ohne Atom“).*
-

Das Beispiel „Strom ohne Atom“ zeigt, dass der für die schliessende Statistik sehr zentrale Prozess der Stichprobenziehung, der insbesondere häufig bei Umfragen zum Tragen kommt, exakt zum Binomialmodell passt.

R-Frage 26 *Angenommen, im Kanton X ist 40% der Bevölkerung katholisch. Wie gross ist die Wahrscheinlichkeit, dass eine Zufalls-Stichprobe von 10 Personen in diesem Kanton in Bezug auf den Katholiken-Anteil exakt repräsentativ ist?*

Frage 8 *In der Modellvorstellung eines idealen Gases besteht ein solches aus unendlich kleinen Molekülen, die sich gegenseitig nicht beeinflussen. Im Prinzip können sich daher alle Moleküle am selben Ort aufhalten. Wo sich das einzelne Molekül zu einem bestimmten Zeitpunkt genau aufhält, ist so gut wie Zufall.*

Denken wir uns ein Gefäss, das nur gerade 50 Moleküle eines idealen Gases enthält. Wir denken uns das Gefäss in zwei Hälften unterteilt (nur im Geiste, keine Trennwand).

- a) Wie gross ist die Wahrscheinlichkeit, dass sich zu einem bestimmten Zeitpunkt alle 50 Moleküle in der rechten Hälfte des Gefässes befinden?
- b) Wie gross ist die Wahrscheinlichkeit, dass sich zu einem bestimmten Zeitpunkt 25 Moleküle in der rechten Hälfte des Gefässes befinden und die anderen 25 in der linken Hälfte?
- c) Angenommen, das Gas besteht aus 25 O₂-Molekülen und 25 N₂-Molekülen – wie gross ist dann die Wahrscheinlichkeit, dass sich zu einem bestimmten Zeitpunkt alle 25 O₂-Moleküle in der rechten Hälfte des Gefässes befinden und alle N₂-Moleküle in der linken Hälfte?

Frage 9 a) Wie selten mag die in der Zeitungsmeldung unten geschilderte Situation auf der Erde auftreten?

Beten für einen Sohn

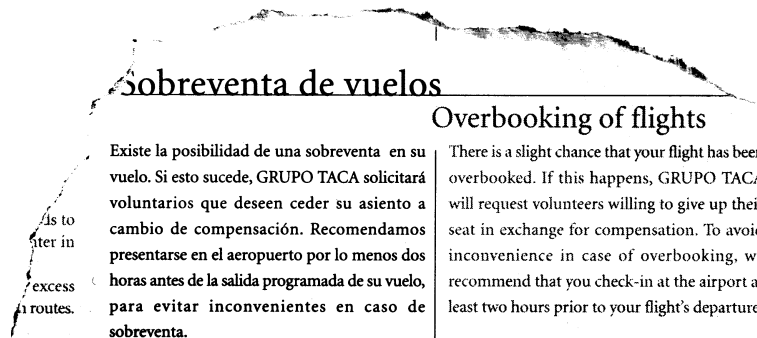
Katmandu. – Zum 20. Mal ist sie bereits schwanger, die 48-jährige Subhadra Dangi aus dem kleinen nepalesischen Dorf Hajipur. Und die Hindufräule hofft und betet, es möge doch ein Sohn werden – nach 19 Töchtern. Mehr als die Hoffnung bleibt ihr nicht, da es in dieser rückständigen Gegend noch keine Ultraschall-Möglichkeiten gibt.

Die Familie ist arm und lebt vom Einkommen aus zwei Dutzend Kühen und gleich viel Ziegen. Dennoch will Subhadra Dangi unbe-

dingt noch einen Sohn, denn dieser besitzt im Hinduismus – wie in den meisten asiatischen Religionen – eine besondere Stellung und wird den Mädchen oft vorgezogen. Er ist auch Träger des Familienerbes. Eine Braut, die nach der Heirat in die Familie des Mannes zieht, wird erst durch die Geburt eines Sohnes Vollmitglied in dessen Gemeinschaft. Die Entscheidung habe Subhadra Dangi selbstständig getroffen, «ohne jeglichen äusseren Druck», so der Ehemann. (vks.)

- b) Wie gross ist die Wahrscheinlichkeit, dass es noch einmal ein Mädchen gibt?

Frage 10 Fluggesellschaften „überbuchen“ die Flüge normalerweise, d.h. sie verkaufen (sofern sich genügend Käufer finden) mehr Sitze als das Flugzeug in Wirklichkeit hat. Die Gesellschaften rechnen nämlich fest damit, dass die eine oder andere Passagierin nicht oder nicht rechtzeitig eintrifft.



Angenommen, eine Fluggesellschaft geht von folgenden Vorgaben aus:

- das Flugzeug hat 280 Plätze;
- die langjährige Erfahrung zeigt, dass 4% der Reservationen nicht wahrgenommen werden;
- es werden maximal 285 Tickets verkauft.

Fragen:

- a) Was sind hier k , p , n , und was ist hier überhaupt binomialverteilt?
- b) Wie viele Passagiere werden mit der grössten Wahrscheinlichkeit eintreffen, wenn 285 Tickets ausgegeben sind?
- c) Wie gross ist die Wahrscheinlichkeit, dass exakt 280 Passagiere eintreffen?
- d) In wie vielen Prozent der maximal „ausgebuchten“ Flüge müssen Passagiere mit einer gültigen Reservation umgebucht/vertröstet werden?

(Wir berechnen dies noch nicht definitiv, aber wenigstens formal.)

Anders als in den vorangegangenen Aufgaben ging es in der letzten Teilaufgabe in Frage 10 nicht um einen einzigen, festgelegten k -Wert, sondern um einen k -Bereich, wobei k hier grösser als ein Grenzwert $x_k = 280$ ist: $n \geq k > x_k$. Dieser Fall ist in den Anwendungen viel häufiger. Es handelt sich eigentlich um Quantile der Binomialverteilung (analog zu den Quantilen der Normalverteilung, die dir im letzten Semester begegnet sind). Die Wahrscheinlichkeit mehr als x_k mal Erfolg zu erhasen ergibt sich dann zu

$$P(k > x_k) = \sum_{k=x_k+1}^n \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

Obige Summe (Gleichung 3.9) können wir bequem mittels des **R**-Befehls

`1-pbinom(xk,n,p)=pbinom(xk,n,p,lower.tail=FALSE)`

berechnen. Das zusätzliche Argument `lower.tail=FALSE` erlaubt uns die **Verteilunganteile rechts von** x_k zu berechnen. Lies hierzu bitte die Definition des Befehls mit `?pbinom` nach. Im Flugzeugbeispiel ist `xk=280`, `n=285` und `p=0.96`.

Standardmässig berechnet **R** die Verteilungsanteile links von x_k und inklusive x_k (`lower.tail=TRUE`, \leq). Die Wahrscheinlichkeit weniger als x_k mal oder genau x_k mal Erfolg zu haben ergibt sich zu

$$P(k \leq x_k) = \sum_{k=0}^{x_k} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k},$$

was wir in **R** einfach mittels

`pbinom(xk,n,p)`

berechnen können.

R-Frage 27 *Nochmal das Flugzeug-Business von S. 37. Wie gross ist die Wahrscheinlichkeit, dass jemand überzählig ist?*

R-Frage 28 *Bei der Produktion von Pommes-Chips-Säcken wird 3% Ausschuss produziert. Wie gross ist die Wahrscheinlichkeit, in einer Stichprobe aus 500 produzierten Säcken mehr als 20 Ausschussstücke zu finden?*

Frage 11 *Wie gross sind Mittelwert und Standardabweichung in der Flugzeug -Frage 10? Skizziere diese Wahrscheinlichkeitsverteilung.*

CHECKLISTE

Verstehst du jetzt:

- Wie der Zufall bei Stichproben hineinspielt und dass man diesen Zufall mit berechenbaren Wahrscheinlichkeiten beschreiben kann?
- Was eine Binomialverteilung ist und warum sie oft eine Rolle bei Stichproben spielt?

Kannst du in eigenen Worten erklären

- welches Modell hinter einer Binomialverteilung steckt?

Kannst du folgende Begriffe auf eine Situation mit binomialverteilten

Wahrscheinlichkeiten anwenden?

- Ereignis, Erfolg, Misserfolg, Wahrscheinlichkeit des Erfolgs im Einzelereignis

Kannst du jetzt:

- Die Wahrscheinlichkeiten einer Binomialverteilung mit und ohne **R** berechnen?
- Die Fragen ohne Hilfe lösen?

Kapitel 4

Schätzungen aus Stichproben

4.1 Konfidenz und Signifikanz

Der Zweck einer Stichprobe ist meist – um es etwas abstrakt zu formulieren – die eine oder andere Eigenschaft der Verteilung eines Merkmals in der Population (auch „Grundgesamtheit“ genannt) zu schätzen, aus der die Stichprobe gezogen ist. Beispiel: ich frage 11 Studierende nach ihrer Körpergrösse, um daraus möglichst exakt abzuleiten, wie gross die ~~zu~~ Studierenden im Mittel sind: hier wäre die Studentenschaft die Population, um die es geht, die Körpergrösse das Merkmal, und mit der untersuchten Eigenschaft der Grössenverteilung ist die Durchschnittsgrösse gemeint.

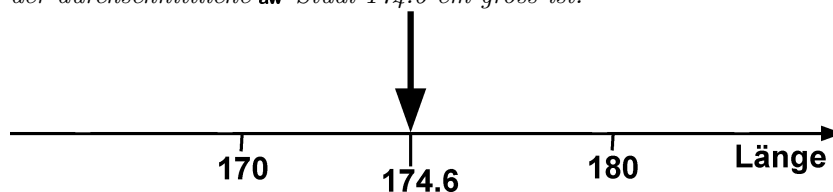
Man kann versuchen, eine Grösse aus einer Stichprobe *möglichst exakt* zu schätzen, aber auf den tatsächlichen Wert wird man nicht kommen, es sei denn mit viel Glück. Das ist der entscheidende Unterschied zwischen Stichproben und Vollerhebungen. Stichproben liefern nur *Schätzwerte*.

Wie man auf den Schätzwert kommt, ist schnell gesagt und nicht überraschend: der beste Schätzwert für die Population ist der entsprechende Wert der Stichprobe. Im Beispiel: Der Durchschnitt der Körpergrössen in der Stichprobe ist gleichzeitig der Schätzwert für die Durchschnittsgrösse aller ~~zu~~ Studierenden.

Beispiel: 11 zufällig ausgewählte Studierende haben diese Körperlängen (in cm):

160, 167, 168, 170, 171, 175, 175, 178, 180, 182, 195.

Der Durchschnitt ist 174.6 cm. Aufgrund dieser Stichprobe behaupten wir nun, dass der durchschnittliche ~~zu~~ Studi 174.6 cm gross ist.

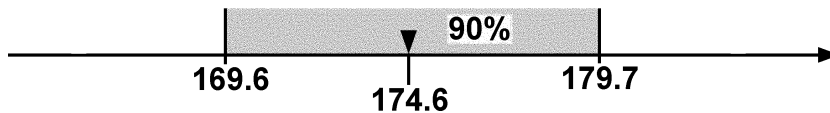


Aber damit ist die Arbeit noch keineswegs getan. Wie (un)genau ist jetzt dieser Schätzwert? Trifft er die Wahrheit bis auf ein paar Millimeter, oder müssen wir damit rechnen, dass die

Studentenschaft in Wahrheit 20 cm grösser ist? Nur wenn wir die (Un)genauigkeit quantifizieren können, ist der Schätzwert brauchbar.

Wie soll man eine Ungenauigkeit quantifizieren? Eine Schätzung der Körpergrössen könnte z.B. in folgender Art formuliert werden.

Die Studierenden sind im Durchschnitt etwa 174.6 cm gross, wobei wir mit 90%iger Sicherheit sagen können, dass der wahre Wert zwischen 169.6 cm und 179.7 cm liegt (wie sich dieses Zahlenwerte berechnen, wird in Abschnitt 4.3 erläutert.)



Um die Genauigkeit der Schätzung zu quantifizieren, kommen zur Schätzung selber (174.6 cm) also noch zwei Angaben hinzu: einerseits muss man sagen, in welchem Bereich der wahre Wert *sehr wahrscheinlich* sein wird (169.6 cm ... 179.7 cm), andererseits, wie gross diese Wahrscheinlichkeit ist, was man also mit „sehr wahrscheinlich“ genau meint (hier 90%).

Das erste (169.6 cm ... 179.7 cm) nennt man das „Konfidenz-“ oder „Vertrauensintervall“ (oder –„Bereich“) *KI*. Man „vertraut darauf“, dass die Wirklichkeit irgendwo in diesem Vertrauensbereich liegt.

Das zweite (die 90%) nennt man „Signifikanz-“ oder „Verlässlichkeitsniveau“ oder „Konfidenzniveau“.

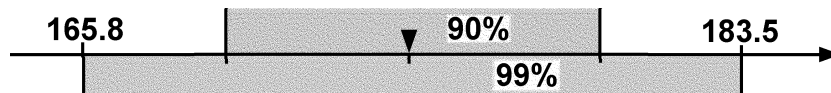
- ⊗ **Notation:** Das Formelzeichen für das Signifikanzniveau ist γ . Häufiger als γ gibt man jedoch die „Irrtumswahrscheinlichkeit“ (oder das „Irrtumsniveau“) an: $\alpha = 1 - \gamma$ ($\alpha = 0.1 = 10\%$ im Beispiel oben). (α und γ sowie die zugehörigen Begriffe, werden in der Praxis oftmals verwechselt, doch ergibt sich meist aus dem Kontext, ob das Signifikanzniveau γ oder die Irrtumswahrscheinlichkeit α gemeint ist). Wenn man „signifikant“ schreibt, ohne weitere Präzisierung, ist $\alpha = 0.05 = 5\%$ gemeint, „sehr signifikant“ heisst $\alpha = 0.01 = 1\%$.

- o! **Merke!** Das Konfidenzintervall (auch Vertrauensintervall genannt) gibt die Präzision der Lageschätzung eines Parameters (zum Beispiel eines Mittelwertes) an. Das Konfidenzintervall ist der Bereich, der bei unendlicher Wiederholung eines Zufallsexperimentes mit einer gewissen Häufigkeit (dem Signifikanzniveau γ) die wahre Lage des Parameters einschliesst.

Konfidenz(intervall) und Signifikanz(niveau) gehören immer zusammen, eine Angabe alleine sagt nicht viel aus. Sobald eine der beiden Grössen gewählt wurde, folgt daraus die andere (zu gegebenen Daten). Vielleicht ist man mit einer 90%igen Sicherheit (Signifikanz) nicht zufrieden. Dann wählt man ein grösseres Signifikanzniveau γ , z.B. $\gamma = 99\%$. Nun wird aber auch auch das resultierende Konfidenzintervall breiter sein: Was die Aussage an Sicherheit gewinnt, verliert sie an Schärfe.

- o! **Merke!** Je grösser das Signifikanzniveau γ (je kleiner die Irrtumswahrscheinlichkeit $\alpha = 1 - \gamma$), desto breiter ist das zugehörige Konfidenzintervall.

In unserem Beispiel liegt die tatsächliche Körperlänge der ~~zu~~ Studierenden mit 99%iger Sicherheit zwischen 165.8 cm und 183.5 cm:



Eigentlich ist es etwas ganz Alltägliches, dass eine Vorhersage um so sicherer ist, je weniger scharf sie gefasst ist. Bei einer Wettervorhersage kann der Meteorologe vielleicht aus folgenden Aussagen auswählen:

- „mit 100% Wahrscheinlichkeit findet das Wetter heute statt.“
- „mit 80%iger Wahrscheinlichkeit wird es heute eher kalt sein.“
- „mit 5%iger Wahrscheinlichkeit wird die Tageshöchsttemperatur heute zwischen 2°C und 3°C liegen.“

100%ige Sicherheit hat man bei statistischen Resultaten meist nur für nutzlose Erkenntnisse, z.B. dass die Grösse der Studierenden irgendwo zwischen 0 und 3 Metern liegen muss. Wir müssen in der Statistik damit leben, dass immer ein gewisses „Restrisiko“ bleibt, dass unsere Resultate falsch sind. Aber: Wir können das Restrisiko wenigstens beziffern!

Für dichotome Merkmale können wir mit unseren bisherigen Methoden berechnen, wie gross bei bekanntem Populationswert (auch Wert der Grundgesamtheit genannt) die Wahrscheinlichkeit für dieses oder jenes Resultat einer Stichprobe ist. Wir können auch für einen ganzen Bereich (d.h. für ein Konfidenzintervall) ausrechnen, mit welcher Wahrscheinlichkeit das Resultat einer Stichprobe in dieses Intervall fällt (das wäre dann das Signifikanzniveau). In der Realität weiss man aber im Voraus nicht, was der Populationswert ist, sondern versucht den Populationswert mittels Stichproben zu ermitteln.

- ! *Merke! Schätzungen aus Stichproben sind wertlos, wenn das Konfidenzintervall fehlt. In den Statistik-Prüfungen gibt es bei Schätzaufgaben keine Punkte, wenn das Konfidenzintervall fehlt.*

Doch noch all diesen Vorabüberlegungen wollen wir im folgenden Abschnitt nun zeigen, wie man das Konfidenzintervall zu gegebenem Signifikanzniveau berechnet. Das jeweilige Verfahren hängt vom Datentyp Deiner Stichprobe ab.

4.2 Anteilsschätzungen

4.2.1 Dichotome Daten

Wie genau ist nun z.B. das Resultat einer Abstimmungs-Umfrage, oder ganz allgemein die Schätzung eines dichotomen Merkmals bzw. eines Anteils?

Zuerst die Begriffe klären: wir bezeichnen bei dichotomen Merkmalen (z.B. „ja/„nein“) im folgenden jeweils einen der beiden Werte (z.B. „ja“) als „Erfolg“. Welchen der beiden Werte man dafür wählt, ist willkürlich – analog zum Binomial-Modell (Abschnitt 3.2).

Eine Stichprobe des Umfangs n hat dann einen Erfolgsanteil

$$\hat{p} = \frac{k}{n},$$

wobei k die Zahl der Erfolge in der Stichprobe ist. Bei dichotomen Merkmalen geht es immer um Anteile. Umgangssprachlich würde man sagen: „in k von n Fällen ist ...“, das ist genau die gleiche Information, und in diese Information steckt alles, was in einer Stichprobe aus dichotomen Daten enthalten ist. **Der Erfolgsanteil der Stichprobe ist gleichzeitig unsere beste Schätzung für den Erfolgs-Anteil π der ganzen Population:**

WENN IN EINER STICHPROBE **dichotomer** DATEN MIT UMFANG n k ERFOLGE GEZÄHLT WERDEN, SO IST DER ERFOLGSANTEIL DER STICHPROBE

$$\hat{p} = \frac{k}{n} \quad (4.1)$$

DIE BESTE SCHÄTZUNG FÜR DEN ERFOLGSANTEIL DER GANZEN POPULATION π . DAS KONFIDENZINTERVALL (IN **R** MIT `conf.level` BEZEICHNET) ZUM SIGNIFIKANZNIVEAU γ (IN **R** `gamma`) EINER STICHPROBE MIT n EREIGNISSEN UND k ERFOLGEN, KANN MITTELS DES **R**-BEFEHLS

`binom.test(k,n,conf.level= gamma)`

BERECHNET WERDEN.

BEI GROSSEN STICHPROBEN ($n \cdot \hat{p} \cdot (1 - \hat{p}) \geq 9$) KANN MAN DIE GRENZEN DES INTERVALLS APPROXIMATIV MIT

$$\pi_{u/o} = \frac{k}{n} \pm z_{\alpha/2} \cdot \sqrt{\frac{k}{n^2} \cdot \left(1 - \frac{k}{n}\right)} = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} . \quad (4.2)$$

BERECHNEN.

IN (4.2) GEHT DAS SIGNIFIKANZNIVEAU $\gamma = 1 - \alpha$ ÜBER DEN KOEFFIZIENTEN $z_{\alpha/2}$ EIN. FÜR DIE ÜBLICHSTEN IRRTUMSWAHRSCHEINLICHKEITEN $\alpha = 1 - \gamma$ GILT:

γ	0.90	0.95	0.99
α	0.10	0.05	0.01
$z_{\alpha/2}$	1.64	1.96	2.58

(4.3)

IN ANDEREN FÄLLEN KANN MAN $z_{\alpha/2}$ MIT DER **Q**UANTILFUNKTION DER STANDARDNORMALVERTEILUNG (DER UMKEHRFUNKTION DER WAHRSCHEINLICHKEITSDICHTEFUNKTION DER STANDARDNORMALVERTEILUNG) IN **R** MITTELS

`qnorm(1-alpha/2)`

ODER ALTERNATIV MIT

`abs(qnorm(alpha/2))`

BERECHNEN.

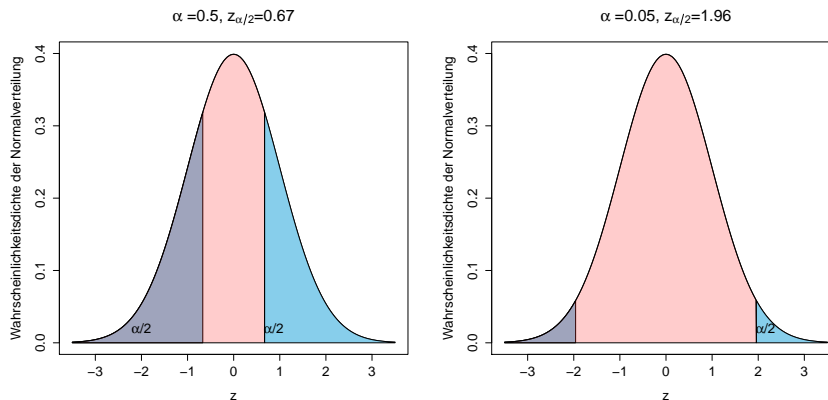


Abbildung 4.1: Wahrscheinlichkeitsdichtefunktion der Standardnormalverteilung. Die rechte blau unterlegte Fläche mit Flächeninhalt $\alpha/2$ ($\alpha = 0.5$ (links) und $\alpha = 0.05$ (rechts)) und $z_{0.25} = 0.67$ (links) und $z_{0.025} = 1.96$ (rechts) kann mit `1-pnorm($z_{\alpha/2}$)` berechnet werden. Kennst Du umgekehrt nur die Irrtumswahrscheinlichkeit α , kannst Du das zugehörige $z_{\alpha/2}$ mit `qnorm(1-alpha/2)` berechnen.

R-Frage 29 Du kontrollierst in einer Produktionslinie für Pommes-Chips-Säcke 100 Stück und findest, dass 3 davon die Spezifikationen nicht erfüllen. Was kannst du auf dem Signifikanzniveau $\gamma = 95\%$ aussagen?

Frage 12 Was ist der Unterschied der letzten Frage zur ersten Pommes-Chips-Frage in Abschnitt 3.2?

Frage 13 Abstimmungssonntag, es wird über die Initiative „Tempo 180 innerorts“ abgestimmt. Es ist 12h. Du sollst exklusiv für „~~sw~~TV“ eine Hochrechnung machen. Zur Verfügung hast du eine Stichprobe von genau 1000 bereits ausgezählten Stimmzetteln. Laut dieser Stichprobe wird die Initiative mit 67% ja-Stimmen angenommen werden.

a) Was kannst du aus deiner Stichprobe mit 99%iger Sicherheit lesen?

b) Und auf dem 90%-Niveau?

c) Auf welchem Signifikanzniveau kannst du deine Abstimmungsprognose auf $67\% \pm 1\%$ zuspitzen?

Hinweis: Berechne zuerst das zugehörige $z_{\alpha/2}$. Um auf das Signifikanzniveau zu schließen, benutze die **R**-Funktion

`pnorm($z_{\alpha/2}$)`,

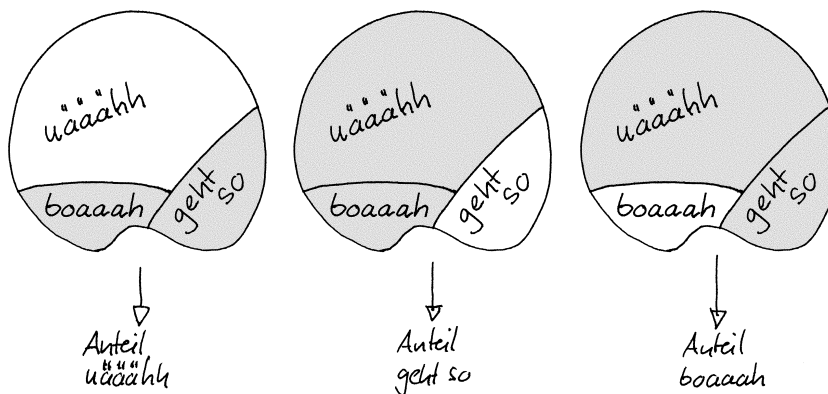
die die Fläche (das Quantil) $\Phi(z_{\alpha/2})$ unter der Dichtefunktion der Standard-Normalverteilung (Mittelwert 0, Standardabweichung 1) für alle Werte, die kleiner oder gleich als $z_{\alpha/2}$ sind, berechnet ¹.

4.2.2 Nominale und ordinale Daten

Von 67 Studierenden finden meine wunderbare neue Krawatte

59 ☒ uäääh, wie hässlich 4 ☒ geht so 4 ☒ boaaah, so schön!

Was diese 67 Leute meinen, wissen wir damit, aber sie sind zum Glück auch nur eine (unbedeutende!) Stichprobe. Ich möchte auch hier Konfidenzintervalle angeben. Nominale (rein qualitative, ohne natürliche Ordnung wie z.B. das Geschlecht oder Farbe) und ordinale (qualitative mit natürlicher Ordnung) Häufigkeiten kann man auf dichotome Häufigkeiten reduzieren. Um den „uäääh“-Anteil zu beurteilen, betrachtet man dies als Erfolg, dann fasst man alles andere (d.h. „geht so“ und „boaaah“) zu einer gemeinsamen Klasse zusammen (nicht-„uäääh“-Misserfolg). Damit ist alles auf die Situation dichotomer Daten reduziert, und man kann mit den Methoden des vorherigen Abschnitts 4.2.1 vorgehen. In dieser Weise bearbeitet man zuerst die Klasse „uäääh“, dann bearbeitet man analog die Klasse „geht so“ und dann die Klasse „boaaah“:



BEI NOMINALEN UND ORDINALEN DATEN SCHÄTZT MAN DIE ANTEILE DER EINZELNEN KLASSEN, INDEM MAN EINE KLASSE ALS ERFOLG UND ALLE ANDEREN KLASSEN ZUSAMMENGEFASST ALS MISSERFOLG ANNIMMT UND DANN MIT DEN METHODEN FÜR DICHOTOME DATEN (ABSCHNITT 4.2.1) VERFÄHRT.

R-Frage 30 Führe dies mit den Krawatten-Meinungen durch ($\alpha = 10\%$).

Frage 14 *Drei Wochen vor den Nationalratswahlen werden 3000 Personen befragt, wie sie wählen werden (unentschiedene und nicht auskunftswillige Personen werden nicht mitgezählt.) Resultat:*

<i>Kommunationalistische Aktion (KA)</i>	<i>740</i>
<i>Himmelblaues Bündnis (HB)</i>	<i>950</i>
<i>Evolutionäre Volks-Affen (EVA)</i>	<i>910</i>
<i>Trottinett-Partei (TP)</i>	<i>400</i>

Kann man zu 95% sicher sein, dass das HB die wählerstärkste Partei sein wird?

4.3 Schätzung eines Durchschnitts (metrische Daten)

Ebenso wichtig wie die bisherigen Situationen ist die Messung metrischer Größen zur Bestimmung eines Durchschnitts.

WENN DIE METRISCHEN DATEN EINER STICHPROBE VOM UMFANG n EINEN MITTELWERT \bar{x} HABEN, SO IST

$$\bar{x}$$

DIE BESTE SCHÄTZUNG FÜR DEN MITTELWERT μ DER POPULATION.

DAS ZUGEHÖRIGE KONFIDENZINTERVALL KANN IN **R** MIT

```
t.test(x, conf.level = ...)
```

BERECHNET WERDEN, WOBEI **x** EIN DATEN-VEKTOR IST, DER AUS DEN WERTEN DER STICHPROBE BESTEHT.

ALTERNATIV KANN MAN DIE GLEICHUNG

$$\mu_{u/o} = \bar{x} \pm t_{\alpha/2,n} \cdot \frac{s}{\sqrt{n}} , \quad (4.4)$$

BENUTZEN.

HIERBEI IS s IST DABEI DIE STANDARDABWEICHUNG DER STICHPROBE UND IST EINE SCHÄTZUNG FÜR DIE STANDARDABWEICHUNG σ DER GRUNDGESAMTHEIT. DAS VERHÄLTNIS

$$SE = s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

BEZEICHNET MAN ALS „STANDARDFEHLER“ (ENGLISCH: STANDARD ERROR, ABGEKÜRZT „SE“) DER SCHÄTZUNG DES MITTELWERTES. DER STANDARDFEHLER IST DIE GESCHÄTZTE STREUUNG DER STICHPROBENMITTELWERTE UM DEN WAHREN MITTELWERT DER GRUNDGESAMTHEIT. DER SE NIMMT MIT ABNEHMENDER STREUUNG DER EINZELWERTE IN DER GRUNDGESAMTHEIT UND MIT STEIGENDEM STICHPROBENUMFANG AB.

$t_{\alpha/2,n}$ IST EIN KOEFFIZIENT, DER SOWOHL VON DER IRRTUMSWAHRSCHEINLICHKEIT α ALS AUCH (FÜR $n > 30$ NUR NOCH SCHWACH) VOM STICHPROBENUMFANG n ABHÄNGT. IN **R** KÖNNEN WIR $t_{\alpha/2,n}$ MIT DER QUANTILFUNKTION DER t -VERTEILUNG, ALSO DER UMKEHRFUNKTION DER t -VERTEILUNGSFUNKTION $f(t)$ MITTELS

```
qt(1 - alpha/2, n - 1)
```

BERECHNEN. $n - 1 = df$ (IM 1 - STICHPROBENFALL) SIND DIE ANZAHL DER FREIHEITSGRADE (ENGLISCH "DEGREES OF FREEDOM", ABGEKÜRZT „df“).

t – Verteilungen und Normalverteilung

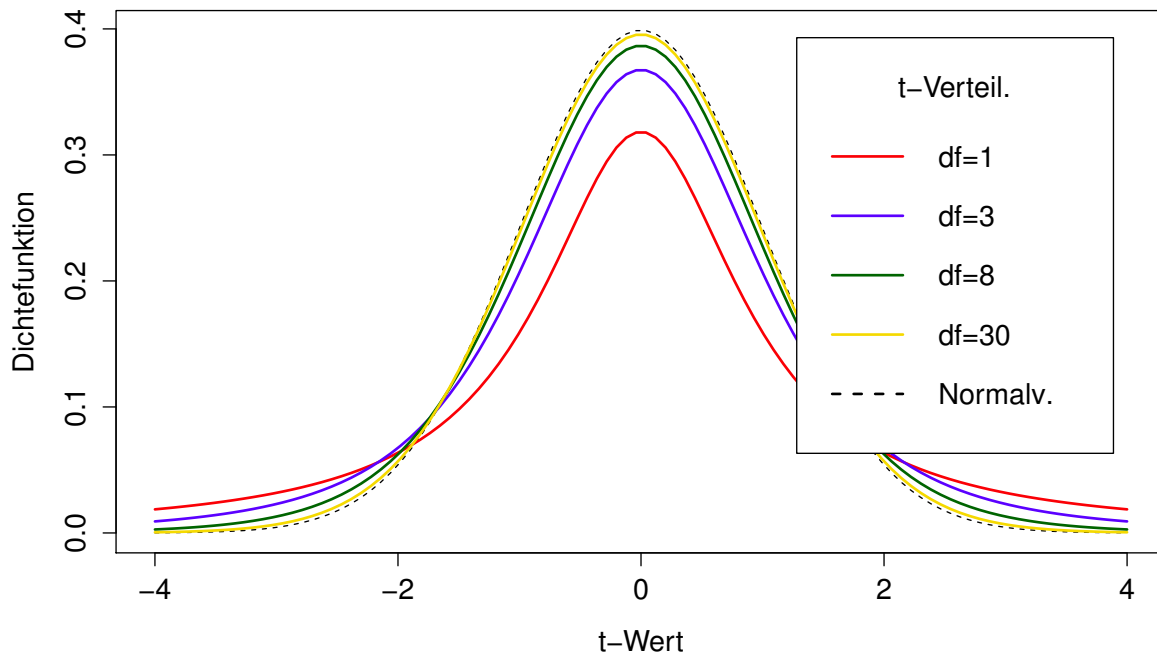


Abbildung 4.2: t-Verteilungen für unterschiedliche grosse Anzahl von Freiheitsgraden $df = n - 1$ und Normalverteilung.

Abbildung 4.2 illustriert die Abhängigkeit der Dichtefunktion der t-Verteilung von der Anzahl der Freiheitsgrade. Die t-Verteilung wird mit wachsendem $n = df + 1$ schmaler und geht für $n \rightarrow \infty$ in die Normalverteilung über. Für grosse n ($n > 30$) gilt: $t_{\alpha/2, n} \approx t_{\alpha/2} \approx z_{\alpha/2}$ ².

R-Frage 31 Seit ein paar Jahren sammle ich Angaben zu Grösse, Gewicht, Geschlecht, Augen- und Haarfarbe der Studierenden. Ich wähle daraus mit einem Zufallsmechanismus 10 Frauen aus. Deren Grösse (in cm) ist:

168 163 160 161 163 170 169 178 164 164

Ist die Grössenschätzung, die du aus dieser Stichprobe ableiten würdest, mit 95%- Sicherheit kompatibel mit dem tatsächlichen Mittelwert der Körpergrösse aller Frauen in der Daten-Sammlung, nämlich 167.6 cm?

²Hier der Vollständigkeit halber die Formel für die Wahrscheinlichkeitsdichte der t-Verteilung:

$$f_n(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}},$$

wo die Γ -Funktion definiert ist als

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

Diese recht komplexe Formel wird nicht von Euch verlangt.

4.4 Schätzung einer Anzahldichte

Bei einer Anteilsschätzung zählt man die Anzahl k der Erfolge, aber auch die Anzahl $n - k$ der Misserfolge. Es gibt jedoch auch Situationen, wo man die Misserfolge nicht erkennen und somit auch nicht zählen kann, so denn überhaupt definiert ist, was man unter Misserfolg verstehen soll.

Beispiel: du zählst die Bäume auf einem Hektar Waldboden. Die Bäume sind so etwas wie Erfolge, aber du kannst nicht gut die Misserfolge („Nicht-Bäume“) zählen. Es handelt sich gewissermassen um einen Datentyp mit nur einer einzigen Ausprägung („Baum“), während dichotome Merkmale, für die man Anteile zählen kann, zwei Werte haben („ja/„nein-...“). Oder etwas anders ausgedrückt: es ist wie bei den Anteils-Schätzungen, aber man kann nicht wissen, was n ist:

Insbesondere gibt es für k keine Einschränkung. Du kannst nicht sagen, wie viele Bäume auf einem Hektar maximal zu finden sind. (Während in einer 1000er-Stichprobe eines dichotomen Merkmals ganz klar höchstens 1000 Erfolge enthalten sein können.) Natürlich ist klar, dass du pro ha nicht z.B. 10^{50} Bäume zählen wirst; aber wo genau die Grenze des Möglichen ist, lässt sich nicht sagen.

Es geht jetzt also einfach um eine Anzahl, um etwas, das man zählt, ohne Gegengrösse. Trotzdem ist diese Anzahl in der Regel nicht ohne Bezug. Der Bezug besteht jedoch nicht in einer anderen Anzahl, sondern in so etwas wie der Waldfläche im Beispiel mit den Bäumen. Man zählt die Bäume *pro Hektar*. Es geht meist um eine Art Dichte, eine Anzahl pro Strecke, pro Fläche, pro Volumen, pro Zeiteinheit.

Die Schätzung einer Anzahl ist ein Spezialfall der Schätzung einer metrischen Grösse ansehen. Ein Spezialfall, in welchem das Konfidenzintervalle enger gewählt werden kann als mit Formel (4.4). Der Spezialfall liegt nur dann vor, wenn die gezählten Objekte natürliche Einheiten darstellen, von denen jedes entweder dabei oder nicht dabei sein kann, aber nie halb dabei. Also nicht z.B. die Anzahl cm, die ein Tisch hoch ist. Meist kommt man mit dem folgenden Rezept zu einem engeren Konfidenzintervall als mit einer metrischen Schätzung (4.4).

ZÄHLT MAN IN EINER STICHPROBE EINE ANZAHLDICHTEN x VON OBJEKTEN EINER BESTIMMTEN ART, SO IST DER SCHÄTZWERT FÜR DIE ANZAHLDICHTEN λ DER GANZEN POPULATION EBENFALLS

$$x. \quad (4.5)$$

DAS KONFIDENZINTERVALL BERECHNET MAN MIT

```
poisson.test(x, conf.level=...)
```

BEI EINER GROSSEN STICHPROBE ($x \gtrsim 100$) KANN MAN DAS KONFIDENZINTERVALL APPROXIMATIV MIT DER FORMEL

$$\lambda_{u/o} = x \pm z_{\alpha/2} \cdot \sqrt{x} . \quad (4.6)$$

BERECHNEN.

Frage 15 *2012 starben auf Schweizer Strassen 339 Menschen bei Verkehrsunfällen, 2013 waren es „nur“ noch 269. Ist dieser Rückgang real, oder ist es einfach eine Zufalls-Schwankung?*

Dann noch ein Hinweis, der eigentlich für alle Schätzmethoden gilt, aber bei den Anzahlschätzungen erfahrungsgemäss besonders dringlich ist: Das Konfidenzintervall muss man *direkt* aus den *Originaldaten* bestimmen. Falls eine Umskalierung verlangt ist (z.B. Umrechnung von „Autos pro Tag“ auf „Autos pro Jahr“), muss man dies *nach* der Berechnung des Konfidenzintervalles machen und die Intervallgrenzen mitskalieren (im Beispiel alles mit 365 multiplizieren).

R-Frage 32 *Du willst herausfinden, wie häufig die „Grüne Kupferdistel“ auf Magerwiesen vorkommt. Dazu zählst du auf einem Stück Magerwiese mit $200 \text{ m} \times 300 \text{ m}$ Fläche die darauf wachsenden Grünen Kupferdisteln. Wie gross schätzt du den Bestand pro km^2 ($\gamma = 0.95$), wenn deine Zählung 84 Exemplare ergibt?*

Falsch wäre in dieser Aufgabe folgende Herangehensweise: Zuerst die 84 auf $(200 \cdot 300) \text{m}^2$ gezählten Kupferdisteln auf $\frac{10^6}{200 \times 300} \times 84 = 16\frac{2}{3} \times 84 = 1400$ Kupferdisteln pro Quadratkilometer skalieren und dann damit das Konfidenzintervall berechnen:

$$\lambda_{u/o} = x \pm z_{\alpha/2} \cdot \sqrt{x} = 1400 \pm 1.96 \cdot \sqrt{1400} = 1400 \pm 73 .$$

D.h. das Intervall reicht nach dieser Rechnung von $\lambda_u = 1327$ bis $\lambda_o = 1473$. Dieses Vorgehen liefert ein anderes, zu kleines Konfidenzintervall, weil mit der zu grossen Input-Zahl 1400 eine zu grosse Stichprobe und damit eine zu grosse Genauigkeit impliziert wurde.

4.5 Standardfehler

Die folgende Tabelle fasst die bisher vorgestellten Formeln für die Schätzungen zusammen:

Datentyp	Formel	R-Befehl
dichotom, ordinale, nominal	$\pi_{u/o} = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$ (4.2)	<code>binom.test(k,n,conf.level=...)</code>
metrisch	$\mu_{u/o} = \bar{x} \pm t_{\alpha/2,n} \cdot \frac{s}{\sqrt{n}}$ (4.4)	<code>t.test(x,conf.level=...)</code> , wo <code>x=c(...)</code> : Datenvektor
Anzahldichte	$\lambda_{u/o} = x \pm z_{\alpha/2} \cdot \sqrt{x}$ (4.6)	<code>poisson.test(x,conf.level=...)</code> , wo <code>x</code> : Anzahldichte

Worum es mir nun geht: die drei Formeln beschreiben das Konfidenzintervall mit einer ganz ähnlichen Struktur. Nämlich:

$$\text{untere/obere Grenze} = \text{Mitte} \pm \left\{ \begin{array}{l} z_{\alpha/2} \\ t_{\alpha/2,n} \end{array} \right\} \cdot \text{noch etwas}$$

Das „noch etwas“ $\left(\sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}, \frac{s}{\sqrt{n}} \text{ oder } \sqrt{x}, \text{ je nach Datentyp} \right)$ hat einen Namen: Das ist der „Standardfehler“ (in obiger Tabelle in rot dargestellt). Nicht zu verwechseln mit der Standardabweichung. Die Standardabweichung sagt etwas über die einzelnen Werte einer Stichprobe aus, nämlich, wie stark sie streuen. Der Standardfehler sagt etwas über die Genauigkeit der Schätzung aus. Es ist gewissermaßen die Masseinheit, in der man die Breite des Konfidenzintervalles misst. Im Gegensatz zur Standardabweichung kommt es bei der Breite des Konfidenzintervalles auch noch auf den Stichproben-Umfang an.

Das übliche Formelzeichen für den Standardfehler ist SE (für Englisch „standard error“). Die Masseinheit Standardfehler wird multipliziert einem Koeffizienten ($t_{\alpha/2,n}$ im Fall metrischer Daten, sonst $z_{\alpha/2}$), der also angibt, wie viele SE das Konfidenzintervall breit ist. Diese Koeffizienten hängen von α ab, bzw. über diese Koeffizienten fließt α ins Konfidenzintervall ein. $z_{\alpha/2}$ hängt sonst von nichts ab, $t_{\alpha/2,n}$ hängt noch leicht von n ab.

CHECKLISTE

Weisst du jetzt, wie man Konfidenzintervalle schätzt:

- bei nominalen Daten?
- bei ordinalen Daten?
- bei metrischen Daten?

Kannst du jetzt die Fragen dieses Abschnittes ohne Hilfe lösen?

4.6 Konfidenzintervall der univariaten Regression

Auch die Datenpunkte, die in eine Regression einfließen, sind nur eine (zufällige) Stichprobe aus der Grundgesamtheit. Wir können ja beispielsweise nicht die Höhe und den Umfang aller Bäume auf der Welt (der Grundgesamtheit) messen, um nach einem allgemeingültigen linearen Zusammenhang zwischen Höhe und Umfang von Bäumen zu suchen. Würde eine Trendgerade nicht etwas anders herauskommen, wenn andere Elemente in die Zufalls-Stichprobe hineingerutscht wären? Wohl schon – aber wieviel anders?

R-Frage 33 *Wir nehmen wieder den `baum.csv`-Datensatz als Beispiel. Lade den Datensatz nochmal, falls du ihn nicht mehr im `Workspace` hast.*

Wiederhole dann die lineare Regression:

```
plot(baum$Dicke,baum$Hoehe)
trend=lm(baum$Hoehe~baum$Dicke)
abline(trend,col="blue")
coefficients(trend)
```

Soweit alles wie gehabt. Wir können aber noch mehr aus dem Regressionsresultat `trend` herausholen. Gib hierzu als erstes

```
confint(trend,level=0.95)
```

ein. Die Konfidenz-Intervalle für a und b wurden damit ausgegeben (der y -Achsenschnittpunkt b liegt mit 95% Sicherheit zwischen 14.8 m und 22.1 m, die Steigung a zwischen 0.018 m/cm und 0.238 m/cm).

Was bedeuten diese beiden Konfidenz-Intervalle in Kombination jetzt für die Gerade? Dies kann man graphisch darstellen:

Mit 95% Wahrscheinlichkeit liegt die wahre Gerade so, dass sie zwischen den beiden türkisfarbenen Kurven in Abbildung 4.3 hindurchpasst. Man nennt diese Kurven Trompetenkurven oder Fehlertrumpeten. Die grünen Trompeten beziehen sich auf das 99%-Niveau.

R-Frage 34 *Und so kannst du die Fehlertrumpeten selber malen:*

```
trumpet95=predict(trend,interval="confidence",level=0.95)
points(baum$Dicke,trumpet95[,2],type="l",lty=1,col="cyan")
points(baum$Dicke,trumpet95[,3],type="l",lty=1,col="cyan")
```

Die grünen Trompeten werden analog mit `level=0.99` in der ersten Zeile und `col="green"` in Zeilen 2 und 3 obiger Aufgabe erzeugt. Nun noch zum Konfidenzintervall des Korrelationskoeffizienten. In welchem Bereich liegt der Korrelationskoeffizient zu 95%? Auch dafür gibt es ein Konfidenz-Intervall:

R-Frage 35 `cor.test(baum$Dicke,baum$Hoehe,level=0.95)` gibt das Konfidenz-Intervall von r zwischen 0.11 und 0.92 an bei gegebenem Signifikanzniveau von $\gamma = 0.95$. (Ausserdem wird dir die Wahrscheinlichkeit der Nullhypothese (die behauptet, dass die Korrelation eigentlich 0 wäre) angezeigt: ($p=0.03$). Mehr zur Nullhypothese im nächsten Abschnitt.

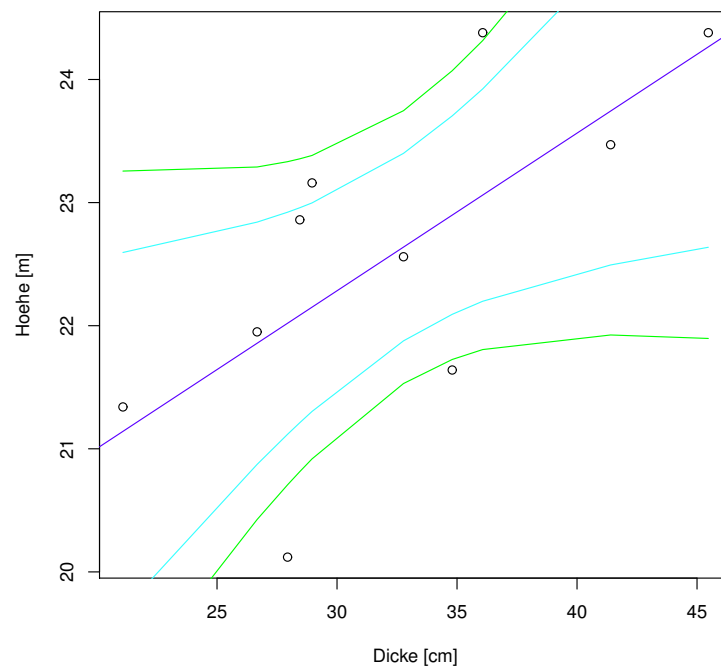


Abbildung 4.3: Trendgerade (blau) mit Trompetenkurven für $\gamma = 0.95$ in türkis und $\gamma = 0.99$ in grün

R-Frage 36 Und nun nochmals der vollständige `trees` Datensatz:

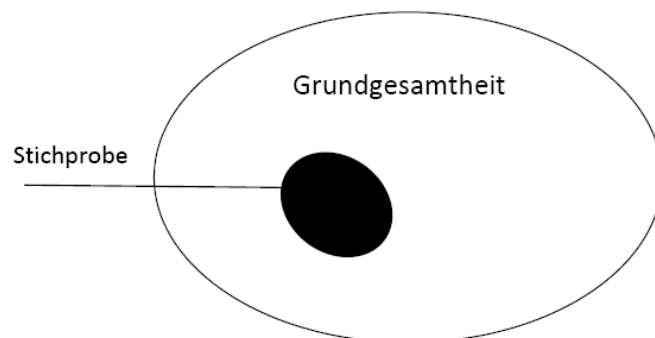
- a) Berechne das Konfidenzintervall des Korrelationskoeffizienten zwischen `Volume` und `Girth`.
 - b) Zeichne sodann die Trompetenkurven (für $\gamma = 0.99$ und $\gamma = 0.95$) der linearen Regression unter der Annahme, dass das Volumen die abhängige, `Girth` die unabhängige Variable sei.
-

Kapitel 5

Statistische Tests -Überblick

5.1 Die Grundfragen der schliessenden Statistik

Schliessende Statistik braucht man, wenn man Aussagen über Grundgesamtheiten aufgrund von Stichproben machen will. Die beschreibende Statistik dagegen kann sowohl im Falle von Stichproben, wie auch für Gesamterhebungen angewendet werden.



Von der Stichprobe ausgehend gibt es zwei Grundprobleme:

1. **Schätzen:** Man will anhand der Stichprobe auf Kennzahlen (z.B. Ja-Anteil, Mittelwert, etc.) der Grundgesamtheit schliessen (Kapitel [4](#)).
2. **Testen:** Man will beispielsweise testen, ob
 - (a) ob es zwischen zwei Merkmalen einen **Zusammenhang** gibt (Regression, Kapitel [2](#))
 - (b) zwei Stichproben von der gleichen Grundgesamtheit sind (**Unterschied** testen)

In den folgenden Kapiteln werden wir uns mit statistischen Tests, die den **Unterschied** im Fokus haben beschäftigen. Beispiel für solche Fragestellungen:

1. Es liegen zwei Stichproben vor und es soll entschieden werden, ob sie sich signifikant unterscheiden oder zur gleichen Grundgesamtheit gehören.
Beispiel: Sind die deutschen Autos anfälliger für Pannen als die japanischen? Das Problem löst man mit einem der vielen Tests für Unterschiede. Bei der Auswahl des Tests spielt vor allem die Art der Daten eine Rolle (metrisch, ordinal, nominal).
2. Es liegen mehrere Stichproben vor und es soll entschieden werden, ob sich mindestens eine von den andern signifikant unterscheidet.
Beispiel: Gibt es Unterschiede bei den Lebenserwartungen in verschiedenen Weltregionen? Hier kommt die einfaktorielle ANOVA zum Zug.

5.2 Testen von Hypothesen

Das Testen auf Unterschiede oder Zusammenhänge gehört zu den Grundfragen der Statistik (siehe Abschnitt 5.1). Bei einer statistischen Studie wird die zugrundeliegende Fragestellung in Form von Hypothesen formuliert.

Bei einer statistischen Studie werden immer gegensätzliche, einander ausschliessende Hypothesen definiert, nämlich die **Nullhypothese(n)** H_0 und die **Alternativhypothese(n)** H_1 .

1. Die **Nullhypothese** H_0 behauptet immer, dass es (eigentlich) keine Unterschiede bzw. keinen Zusammenhang gibt. Die auftretenden Unterschiede, Abweichungen oder Zusammenhänge sind also eine rein zufällige Konsequenz von statistischen Schwankungen.
2. Die **Alternativhypothese** H_1 besagt, dass der Unterschied oder der Zusammenhang nicht zufällig, sondern systematisch ist.

Hinsichtlich der Alternativhypothesen gibt es zwei verschiedene Varianten. Wir können nämlich entweder bloss interessiert sein, ob es einen Unterschied gibt (zweiseitige Alternativhypothese), oder wir können uns auf eine „Richtung“ des Unterschieds (grösser oder kleiner) festlegen (einseitige oder gerichtete Alternativhypothese)

Ob eine einseitige oder zweiseitige Alternativhypothese gewählt werden soll, entscheidet man (also die Person, welche die Analyse macht) in aller Regel selber. Standardmässig wird eine zweiseitige Hypothese aufgestellt. Eine einseitige Hypothese ist nur dann angezeigt, wenn sie sich aufgrund einer starken Vermutung und Hinweisen aus den Daten aufdrängt.

In der Praxis wird zumeist ein α -Niveau von 5% gewählt. Falls allerdings die Ablehnung der Nullhypothese mit drastischen Konsequenzen verbunden ist (z.B. das Töten aller Tiere aufgrund einer Krankheit), sollte man α besser kleiner wählen.

5.3 Parametrische und parameterfreie Tests

Es gibt eine grosse Menge statistischer Tests, um zu prüfen, ob es zwischen Merkmalen einen Zusammenhang gibt oder ob es zwischen Stichproben signifikante Unterschiede gibt.

Welchen Test man nimmt, hängt von der konkreten Fragestellung und dem Skalenniveau (metrische, ordinale oder nominell) ab. Die wichtigsten Tests werden wir besprechen und jeweils fixfertige **R**-Rezepte angeben, die man einfach ausführen muss¹. Das Problem, welchen Test man wählen soll, bleibt aber selber zu lösen. Wir werden uns zunächst auf das Testen auf Unterschiede konzentrieren.

Bemerkungen:

1. Parametrische Tests

Tests für metrische Daten, sogenannte parametrische Tests, müssen mehr Voraussetzungen erfüllen als solche für ordinale oder nominale Daten. Dafür sind metrische Tests mächtiger: **Die Macht eines Tests ist die Wahrscheinlichkeit, dass bei gegebener Stichprobe ein signifikanter Unterschied festgestellt wird, wenn H_1 gilt.** Tests mit kleinerer Macht brauchen grössere Stichproben zur Feststellung von signifikanten Unterschieden (d.h. zur Ablehnung von H_0).

2. Parameterfreie Tests

Parameterfreie Tests kommen bei ordinal- oder nominalskalierten Daten zur Anwendung. Metrische Daten kann man natürlich auch als ordinal- oder nominalskaliert betrachten und die entsprechenden Tests verwenden. Die Teststärke ist dann aber kleiner, das heisst, bei fixer Irrtumswahrscheinlichkeit α —ist es weniger wahrscheinlich, dass eine falsche Nullhypothese verworfen wird und ein signifikanter Unterschied festgestellt wird. Man läuft somit Gefahr, einen sogenannten **Fehler 2. Art** zu machen.

5.4 Fehler 1. und 2. Art

- Beim **Fehler 2. Art** wird die Nullhypothese beibehalten, obwohl sie falsch ist.
- Umgekehrt wird beim Fehler **Fehler 1. Art** die Nullhypothese abgelehnt, obwohl sie zutrifft. Die Wahrscheinlichkeit, dass man einen solchen Fehler 1. Art macht ist gleich der Irrtumswahrscheinlichkeit α . Je kleiner man α wählt, desto sicherer kann man Fehler 1. Art ausschliessen (doch die Gefahr einen Fehler 2. Art zu machen, nimmt zu.)

An den konkreten Beispielen in den nächsten Abschnitten werden wir diese (jetzt noch recht abstrakten) Konzepte des Hypothesentestens näher erläutern. Wir beginnen mit dem t —Test, dem wichtigsten Test für intervallskalierte Daten. Je nach Situation gibt es verschiedene Varianten dieses Test.

¹Für das Verständnis ist es aber sinnvoll, die Tests auch ein paar Mal von Hand gerechnet zu haben.

Kapitel 6

Parametrische Tests

6.1 Der t -Test

6.1.1 Der t -Test im 1-Stichproben-Fall

Die Grundidee des statistischer Testens wollen wir am Beispiel des „ t -Tests“ erläutern. Der t -Test ist eng verwandt mit den Konfidenzintervallen arithmetischer Mittel (Abschnitt 4.3)

¹ Es gibt mehrere Varianten des t -Tests, wir beginnen mit dem „1-Stichproben-Fall“: Diese Variante des t -Tests klärt ab, ob der Durchschnittswert aus einer Stichprobe vereinbar ist mit einem vorgegebenen Vergleichswert. Wenn Vergleichswert und Stichprobenmittelwert zufällig gerade identisch sind, ist dies offensichtlich der Fall. Aber wenn die Stichprobe in die eine oder andere Richtung abweicht, ist noch nicht automatisch klar, ob die Stichprobe dem Vergleichswert widerspricht oder nicht. Denn Stichproben-Mittelwerte fallen ja mal grösser, mal kleiner aus. Die nächste Frage soll dies illustrieren. Im Moment kannst du diese Frage wahrscheinlich noch nicht lösen. Trotzdem liest du schon einmal die Frage. Die Anleitung zur Lösung wird in Bälde folgen, und dann wirst du noch einmal Gelegenheit zur Lösung bekommen.

¹Wir werden aber später zu Tests kommen, die viele andere Möglichkeiten bieten und nicht durch Schätzungen ersetzt werden können.

R-Frage 37 Du hast einen neuen Dünger für Kartoffeln erfunden, den „Megabintje-HighNO_x“. Du probierst ihn an einer Kartoffel-Sorte aus, von der bekannt ist, dass die Kartoffeln im Durchschnitt 240 g schwer sind. Mit deinem tollen Dünger werden sie jetzt 255 g schwer, wie du nach dem Auswägen von 200 Kartoffeln ausrechnest. „Ein phänomenaler Durchbruch!“, rufst du aus und siehst vor deinem geistigen Auge eine gewaltige Dünger-Fabrik entstehen. „Unsinn“, sagt am Abend dein Gatte. Ist doch vielleicht nur ein Stichproben-Zufall. 200 andere Kartoffeln, und der Dünger erscheint vielleicht wirkungslos. Wer hat recht?

Letztlich wird sich die Meinungsverschiedenheit in der Düngerproblematik nie mit endgültiger Gewissheit klären lassen. Trotzdem kannst du etwas tun um Licht in die Geschichte zu bringen, nämlich Folgendes:

1. Du kannst ausrechnen, wie gross die Wahrscheinlichkeit ist, eine solche Stichprobe mit grösserem Gewicht als der Populations-Mittelwert rein zufällig zu ziehen. Du nimmst hier also an, dass die (im Stichprobendurchschnitt schwereren) Kartoffeln Deiner Stichprobe eigentlich aus einer Population mit Mittelwert $\mu = 240$ g stammen. In Kartoffeln ausgedrückt: wie gross ist die Wahrscheinlichkeit, **zufällig** eine 200er-Stichprobe mit $\bar{x} = 255$ g Mittelwert zu ziehen aus einer Population von Kartoffeln, die weiterhin Mittelwert $\mu = 240$ g hat (weil der Dünger nämlich gar nichts bewirkt hat).

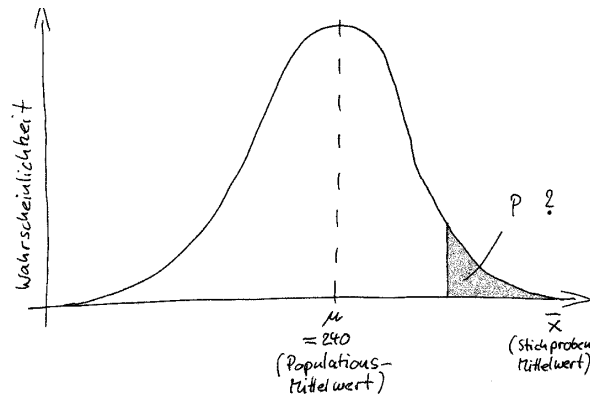


Abbildung 6.1: Beispiel für einseitiges („grösser“) Testen beim t -Test im 1-Stichprobenfall: Der graue Bereich im rechten „Schwanz“ der Verteilung ist die Wahrscheinlichkeit der Nullhypothese P . Ist P kleiner als die vorher festgelegte Irrtumswahrscheinlichkeit α wird die Nullhypothese abgelehnt: Die Kartoffeln unsere Stichprobe wären dann signifikant schwerer als der Populationsdurchschnitt.

Diese Wahrscheinlichkeit nennt sich „ P -Wert“. Es handelt sich hier um die Wahrscheinlichkeit der Nullhypothese.

2. Wenn die Wahrscheinlichkeit der Nullhypothese sehr klein ist (z.B. $p = 0.0001\%$), wirst du sagen: „Eine Stichprobe mit so schweren Kartoffeln kann kein reiner Zufall ist. Der Dünger bewirkt etwas!“ Die gedüngten Kartoffeln werden also höchstwahrscheinlich nicht aus der Population mit $\mu = 240$ g stammen, sondern die gedüngten Kartoffeln haben tatsächlich ein signifikant grösseres mittleres Gewicht. (Was aber noch nicht zwingend heisst, dass $\bar{x} = 255$ g deren wahrer (Populations-)Mittelwert ist.)
Ist die Wahrscheinlichkeit der Nullhypothese hingegen gross (z.B. $p = 30\%$), so ist die Beweiskraft minim, die Nullhypothese kann nicht abgelehnt werden, das höhere Gewicht deutet nicht darauf hin, dass der Dünger etwas bewirkt hat. x
3. Was heisst, die Wahrscheinlichkeit ist „sehr klein“? **Bei welchem P -Wert man die Grenze setzt**, ab der man die Stichprobe als Beweis gegen den Mittelwert (und unseren Referenzwert) $\mu = 240$ g akzeptiert, **muss man unbedingt vorab festlegen** – eigentlich schon bei der Planung des Projektes. Diese Grenze ist wieder die Irrtumswahrscheinlichkeit α . Wiederum bezeichnet man das Komplement zu 100% als das Signifikanzniveau $\gamma = 1 - \alpha$.
4. Die Annahme, „Die Stichprobe stammt aus der unveränderten Population mit $\mu = 240$ g“ ist hier die „Nullhypothese“ (abgekürzt H_0). Meist sagt die Nullhypothese aus, dass *kein* Unterschied (oder Zusammenhang) vorliegt. Die auftretenden Unterschiede (oder Zusammenhänge) sind also eine rein zufällige Konsequenz von statistischen Schwankungen. Es gehört zentral zum statistischen Test-Verfahren, dass man sich anfangs auf den Standpunkt dieser Nullhypothese stellt. Sie ist so etwas wie die Unschuldsumutung bei einem Gerichtsprozess. Meist formuliert H_0 gerade das Gegenteil von dem, was man (im Gerichtsprozess der Staatsanwalt) eigentlich beweisen möchte.
5. Was man eigentlich zeigen möchte, ist die Alternativhypothese H_1 . Sie besagt, dass der Unterschied (oder der Zusammenhang) nicht zufällig, sondern systematisch ist. In unserem konkreten Beispiel der Aufgabe 37 haben wir es mit einer einseitigen (oder gerichteten) Alternativhypothese zu tun: Die Alternativhypothese behauptet, dass der Mittelwert des Kartoffelgewichts unserer Stichprobe (255g) ist (auf dem gegebenen Signifikanzniveau) signifikant *grösser* als der Referenzwert 240 g.
6. Nullhypothese und Alternativhypothese ergänzen sich und schliessen sich gegenseitig logisch aus: $P(H_0) + P(H_1) = 1$

6.1.1.1 Voraussetzungen für den t -Test

Voraussetzung für die korrekte Durchführung des t -Tests ist, dass die metrischen **Stichprobenwerte normalverteilt** sind. Normalverteilung kann in einem ersten Schritt graphisch durch z. B. durch Histogramme oder Boxplots geprüft werden². Bei grossen Stichproben (Stichprobengrösse grösser 30) kann man allerdings auch bei nicht wirklich normalverteilten Daten den t -Test machen. Im folgenden wollen wir das „Rezept“ des t -Tests durchspielen.

²Ein Testverfahren hoher Güte ist der Shapiro-Wilk-Test (in R `shapiro.test(x)`), wo `x` der Datenvektor der Stichprobe ist

6.1.1.2 Das t -Test-Rezept

t -Test, 1-Stichproben-Fall

Test-Frage: Unterscheidet sich der arithmetische Mittelwert \bar{x} einer Stichprobe signifikant von einem vorgegebenen Referenzwert μ ?

Test-Rezept:

1. Formuliere die Nullhypothese H_0 und die korrespondierende Alternativhypothese H_1 .
2. Festlegung der Irrtumswahrscheinlichkeit $\alpha = 1 - \gamma$.
3. Berechnung des P-Wertes:

(a) **Berechnung des P-Wertes aus den Stichprobenwerten mit \mathbb{R} :**

Im Allgemeinen liegen die einzelnen Stichprobenwerte vor. Beispiel:

Die Stichprobe aus den drei Werten 70, 85, 31 wird in \mathbb{R} als Vektor $\mathbf{x}=\mathbf{c}(70,85,31)$ definiert und mit einer Referenzgrösse $\mu = 91$ verglichen.

i. **Zweiseitige Alternativhypothese:**

Wollen wir zeigen, dass der Stichprobenmittelwert \bar{x} sich signifikant vom Referenzwert μ unterscheidet, lautet die

Alternativhypothese H_1 : $\bar{x} \neq \mu$ und die

Nullhypothese H_0 : $\bar{x} = \mu$.

Dann lautet der t -Test in obigem Beispiel

```
t.test(x,mu=91,alternative="two.sided")
```

oder kürzer

```
t.test(x,mu=91)
```

(`alternative="two.sided"` ist die Alternativhypothese, die der `t.test` standardmässig annimmt, falls keine andere Alternativhypothese explizit spezifiziert wird.) Wir erhalten einen P-Wert von 0.21.

ii. **Einseitige Alternativhypothese:**

A. Wollen wir zeigen, dass der Stichprobenmittelwert \bar{x} signifikant **grösser** als der Referenzwert μ ist, lautet die

Alternativhypothese H_1 : $\bar{x} > \mu$ und die

Nullhypothese H_0 : $\bar{x} \leq \mu$.

In \mathbb{R} können wir den t -Test dann in unserem Beispiel folgendermassen durchführen:

```
t.test(x,mu=91,alternative="greater").
```

und erhalten einen P-Wert von 0.89. Mit dem `alternative="greater"` haben wir die Alternativhypothese H_1 spezifiziert, hier, dass die Stichprobe \mathbf{x} einen Wert nachweisen soll, der grösser (`greater`) ist als die Referenzgrösse `mu=91`.

B. Wollen wir zeigen, dass unser Stichprobenmittelwert \bar{x} signifikant **kleiner** als der Referenzwert μ ist, lautet die Alternativhypothese H_1 : $\bar{x} < \mu$

und die Nullhypothese $H_0 : \bar{x} \geq \mu$. In **R** kann der t -Test dann folgendermassen ausgeführt werden:

```
t.test(x,mu=91,alternative="less")
```

Im Output interessiert uns nur der P -Wert, die Wahrscheinlichkeit der Nullhypothese. In unserem Beispiel 0.11.

(b) **Berechnung des P -Wertes aus dem zugehörigen t -Wert:**

In Zeiten vor Statistikprogrammen wie **R** musste man sich zuerst aus Umfang n , Mittelwert \bar{x} und Standardabweichung s der Stichprobe den t -Wert berechnen:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{SE},$$

wo SE der Standardfehler ist. Beispiel: Eine Stichprobe mit Umfang $n = 3$, Mittelwert $\bar{x} = 62$, Standardabweichung $s = 27.9$, Referenzwert wieder $\mu = 91$. In **R** übertragen erhalten wir:

```
n=3; xmean=62; s=27.9; mu=91
t=(xmean-mu)*sqrt(n)/s
```

Wie im Fall (a) kommt es nun darauf an, ob wir zeigen wollen, dass der Stichprobenmittelwert unterschiedlich zum Referenzwert, grösser oder kleiner ist.

i. **Einseitiges Alternativhypothese: grösser**

Wollen wir zeigen, dass der Stichprobenmittelwert signifikant grösser als der Referenzwert ist ($H_1 : \bar{x} > \mu$, $H_0 : \bar{x} \leq \mu$), dann erhalten wir mit dem **R**-Befehl

```
pt(t,n-1,lower.tail=FALSE)
```

den zu unserem Eingangsbeispiel mit Referenzwert 91 t -Wert gehörigen P -Wert von 0.89

ii. **Einseitiges Alternativhypothese: kleiner**

Wollen wir beispielsweise zeigen, dass der Stichprobenmittelwert signifikant kleiner als der Referenzwert ist ($H_1 : \bar{x} < \mu$, $H_0 : \bar{x} \geq \mu$), dann erhalten wir mit dem **R**-Befehl

```
pt(t,n-1)
```

den zu unserem t -Wert gehörigen P -Wert von 0.11.

iii. **Zweiseitige Alternativhypothese:**

Wollen wir folgendes zeigen: $H_1 : \bar{x} \neq \mu$, $H_0 : \bar{x} = \mu$, dann müssen wir unterscheiden, ob $t < 0$ oder $t > 0$ ist.

A. In obigem Beispiel ist $t < 0$, da der Stichprobenmittelwert kleiner als der Referenzwert ist. Somit wir erhalten mit dem **R**-Befehl

```
pt(t,n-1)*2
```

den zugehörigen P -Wert von 0.21.

B. Ist hingegen $t > 0$, weil der Stichprobenmittelwert grösser als der Referenzwert ist, so erhalten wir den zugehörigen P -Wert mit

```
pt(t,n-1,lower.tail=FALSE)*2.
```

Beispiel: Referenzwert von $\mu = 50$, ansonsten wie bisher Umfang $n = 3$,

Mittelwert $\bar{x} = 62$, Standardabweichung $s = 27.9$. Dann erhalten wir den zugehörigen P -Wert zu einem neuen $t = \frac{62-50}{27.9/\sqrt{3}} = 0.75$ zu $P = 0.53$.

4. Wenn $P < \alpha$, ist ein signifikanter Unterschied zwischen arithmetischem Mittel und Referenzwert nachgewiesen und die Nullhypothese kann verworfen werden, andernfalls nicht. In unserem Beispiel: Bei einer Irrtumswahrscheinlichkeit von $\alpha = 0.05$ kann die Nullhypothese nicht verworfen werden

Jetzt kannst du die **R**-Frage 37 bearbeiten. (Für die Standardabweichung der Stichprobe nimmst du $s = 60$ g, und für die Irrtumswahrscheinlichkeit 5%.)

R-Frage 38 *In deiner Semesterarbeit untersuchst du eine neue Art Joghurtdeckel. Dabei geht es um die Kraft, die aufgewendet werden muss, um den Deckel abzuziehen. Diese soll nicht mehr als 4 N betragen. Deine Messwerte sind: 2.5 N, 3.5 N, 5.0 N, 3.8 N.*

Beweisen deine Messungen, dass die neuen Deckel der Anforderung genügen?

Achtung bei der Interpretation: Die Statistik untersucht nicht die Qualität des Experimentes. Wenn du dabei einen Fehler gemacht hast (wenn du z.B. nur die grösseren Kartoffeln für die Auswägung berücksichtigt und die kleinen zur Seite geschoben hast) bringt die Statistik kein korrektes Resultat. In unseren Methoden geht es immer rein nur um die korrekte Behandlung des Stichprobenzufalls, unter der Voraussetzung, dass der Versuch korrekt durchgeführt wurde. Ausserdem wird durch einen statistischen Test immer nur ein allfälliger Unterschied bewiesen, aber nie eine Kausalbeziehung. Vielleicht sind deine Kartoffeln ja wegen des besonders guten Bodens gross heraus gekommen und nicht wegen des Düngers. Die Statistik kann verschiedene Ursachen meist nicht auseinanderdividieren.

Beachte auch noch dies: Wenn man die Nullhypothese nicht verwerfen kann, so ist dies noch kein Beweis für H_0 . H_0 kann mit den hier dargestellten Mitteln überhaupt nie bewiesen werden. Denn vielleicht ist ja nur die Stichprobe zu klein geraten? Nur das Gegenteil kann (statistisch) bewiesen werden. Im Beispiel eben könnte man vielleicht noch 20 weitere Messungen machen, und vielleicht würde es dann reichen, um H_0 zu widerlegen.

6.1.2 Allgemeines und Varianten zur Testerei

Zwar macht jeder der Dutzenden von bekannten statistischen Tests und jede der meist mehreren Untervarianten immer wieder etwas anderes. Trotzdem klingen die Fragestellungen immer wieder ähnlich. Ich versuche in diesem Abschnitt zuerst das Gemeinsame zu formulieren, das sich durchzieht.

Das immer ähnliche:

- Es geht im Weiteren jeweils um den Vergleich von zwei oder mehr Verteilungen eines Merkmals. (Wobei eine der Verteilungen eine theoretische sein kann, eine Vorhersage, eine Vorgabe etc.)
- An diesen Verteilungen interessiert wiederum eine oder mehrere Eigenschaften, z.B. ein Lage- oder ein Streumass, das man vergleichen will.

Frage 16 Betrachten wir nochmals unser Dünger-Beispiel **R-Aufgabe 37**:

- a) Um wie viele und um welche Verteilungen geht es?
 b) Um welche Eigenschaft(en) der Verteilung(en) geht es?
-

- Als erstes schaut man, ob sich überhaupt ein nennenswerter Unterschied abzeichnet. Wenn nein, ist der Rest hinfällig.
 Dünger-Beispiel: $255 > 240 \implies$ ja, die Sache interessiert.
 - Dann formuliert man die Null-Hypothese. Allgemein: „die betrachteten Verteilungen sind zufällig ausgewählte Stichproben aus der gleichen Population.“
 H_0 beim Dünger: 255 statt 240, das ist ganz normaler Stichproben-Zufall.
 - Nun stellt man die Nullhypothese H_0 auf den Prüfstand: Wie wahrscheinlich ist es, dass die beobachtete Differenz zwischen den Stichproben rein zufällig ist? Diese Wahrscheinlichkeit, den P -Wert, vergleicht man mit dem schon im Voraus festgelegten Signifikanz-Niveau α . Der P -Wert ist somit die Wahrscheinlichkeit der Nullhypothese.
 - Die Nullhypothese H_0 wird verworfen, wenn die Wahrscheinlichkeit der Nullhypothese kleiner als die gewählte Irrtumswahrscheinlichkeit ist: $p < \alpha$.
 - Die Nullhypothese wird beibehalten, falls $p > \alpha$ ist.
- o! Merke! In neuerer Zeit wird es mehr und mehr üblich, einfach die P -Werte anzugeben. Es wird gelegentlich argumentiert, der zahlenmäßige P -Wert enthalte mehr Information als die Testentscheidung alleine. Der Leser kann dann sein Signifikanzniveau auch selber wählen. Dem ist aber entgegenzuhalten, dass diese Praxis geradezu dazu einlädt, die Fehler-Wahrscheinlichkeit einfach gleich p zu setzen, womit man dann immer zu einem signifikanten Resultat mit dem bestmögliche Signifikanzniveau kommt. Das ist aber nicht der korrekte Umgang mit Wahrscheinlichkeiten. Es ist fast schon so, als ob man beim Roulett den Einsatz hinlegt, nachdem die Kugel schon ihr Feld gefunden hat!

6.1.2.1 Freiheitsgrad

Die t-Verteilung hängt von der Anzahl der Freiheitsgrade $n - 1$ ab, doch wie erklärt sich dieser Wert?

IST AUS n BEOBACHTUNGEN DAS ARITHMETISCHE MITTEL

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{i=1}^{n-1} x_i + x_n}{n}$$

BERECHNET WORDEN, SO KÖNNEN BEI FESTEM \bar{x} UND GEGEBENEM n NUR NOCH $n - 1$ BEOBACHTUNGEN **frei** VARIIEREN, DENN DER n . - WERT KANN ZU

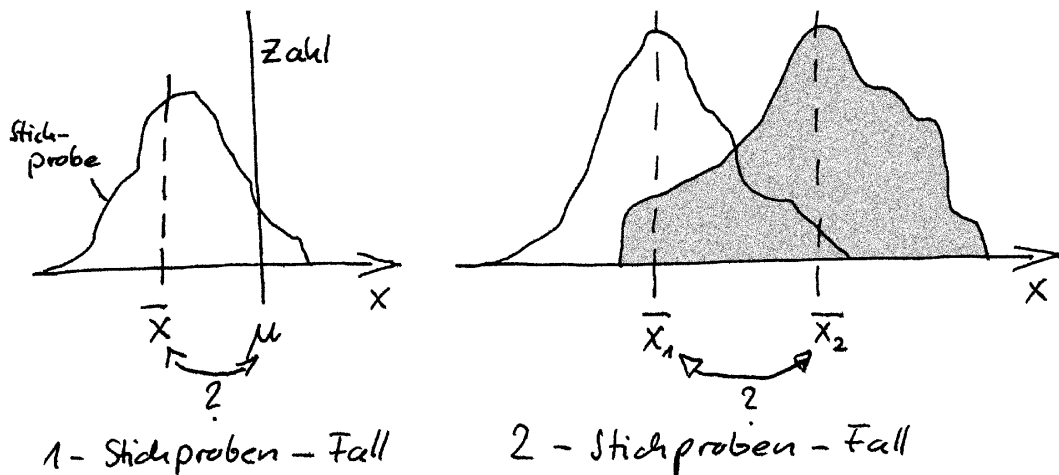
$$x_n = n \cdot \bar{x} - \sum_{i=1}^{n-1} x_i$$

BERECHNET WERDEN.

Beispiel: Drei Messungen $x_1 = 1$, $x_2 = 2$, $x_3 = 3$ mit Mittelwert $\bar{x} = 2$. Kennt man nur den Mittelwert \bar{x} und die $n - 1 = 2$ anderen Messwerte, ergibt sich der $n = 3$. Messwert zu $3 \cdot 2 - 1 - 2 = 3$.

6.1.3 Der t -Test für zwei unabhängige Stichproben

Man kann mit Hilfe des t -Tests nicht nur den Mittelwert einer Stichprobe mit einer gegebenen Zahl (Abschnitt 6.1.1) vergleichen, sondern auch zwei Stichproben, z.B. die Resultate zweier verschiedener Experimente mit Mittelwerten \bar{x}_1 und \bar{x}_2 , miteinander vergleichen.



6.1.3.1 Voraussetzungen für den t -Test im 2-Stichprobenfall

6.1.3.1.1 Normalverteilung in beiden Stichproben

Voraussetzung für die korrekte Durchführung des t -Tests im 2-Stichprobenfall ist eine **Normalverteilung der beiden Stichproben**. Bei Stichproben mit $n > 30$ können wir diese Voraussetzung stillschweigend als gegeben annehmen, da der t -Test dann robust ist gegen Abweichungen von der Normalverteilungsvoraussetzung. Ein statistischer Test (eine statistische Methode) wird als robust bezeichnet, wenn er (sie) auch dann zuverlässig bleibt, wenn die Voraussetzungen (z.B. Normalverteilung) nicht oder nicht vollständig zutreffen.

6.1.3.1.2 Varianzhomogenität: Student's t -Test und Varianzheterogenität: Welch- t -Test

Während der Vergleichswert im 1-Stichproben-Fall als exakt vorausgesetzt wird, trägt im 2-Stichproben-Fall auch die zweite Stichprobe mit ihrer Zufallsstreuung zur Unsicherheit des Resultates bei. Deshalb gehen in die Berechnung der Testgrösse t im 2-Stichproben-Fall neben den Mittelwerten und Umfängen **beider** Stichproben auch die Streuungen **beider** Stichproben ein³. Welchen Test wir anwenden dürfen, hängt nun davon ab, ob die Streuungen der beiden Verteilungen ähnlich sind (man spricht von „Varianzhomogenität“) oder ob sie sich signifikant unterscheiden („Varianzheterogenität“). Wie bei der Normalverteilung kann man das am besten einfach „mit dem Auge“ überprüfen, in dem man die Verteilungen mittels **R** plottet. In Abschnitt 6.3 werden wir den F -Test kennenlernen, mit dem man formal testen kann, ob sich die Streuungen zweier Verteilungen signifikant bei gegebenen Signifikanzniveau unterscheiden. Liegt Varianzhomogenität vor, dürfen wir den Student's t -Test anwenden, ansonsten kommt der Welch- t -Test, der keine Varianzhomogenität voraussetzt, zum Zuge. **Der grosse Vorteil des Welch- t -Tests ist somit, dass man nicht extra vorab auf Varianzhomogenität testen muss.** Da der Welch- t -Test selbst bei ähnlichen Varianzen (Varianzhomogenität) und ähnlichen oder gleichen Stichprobengrössen (also der eigentlichen Domäne des Student's t -Test) eine nur leicht schwächere Teststärke als der eigentliche Student's t -Test besitzt, gehen einige Statistiker dazu über, beim Vergleich der Mittelwerte zweier Stichproben immer den Welch- t -Test auszuführen. Dies ist der Grund dafür, dass der Welch- t -Test die Standard-Einstellung des t -Tests in **R** ist.

³**R** berechnet intern diese Testgrösse, die genaue Berechnung von t im 2-Stichprobenfall geht jedoch über den Stoff dieses Kurses hinaus.

t -Test, 2-Stichproben-Fall, unabhängige Stichproben

Test-Frage: Unterscheidet sich der arithmetische Mittelwert \bar{x}_1 einer Stichprobe 1 signifikant vom arithmetischen Mittelwert \bar{x}_2 einer anderen, unabhängigen Stichprobe 2?

Test-Rezept:

1. Formuliere H_0 und H_1 .
Die Nullhypothese H_0 lautet hier, dass der Unterschied zwischen \bar{x}_1 und \bar{x}_2 **nicht** signifikant ist, die auftretenden Unterschiede sind rein zufällige Konsequenz statistischer Schwankungen.
2. Festlegung der Irrtumswahrscheinlichkeit $\alpha = 1 - \gamma$ (normalerweise $\alpha = 0.05$.)
3. Berechnung des P -Wertes aus den Stichprobenwerten Seien die Zahlenwerte zweier Stichproben gegeben (beispielsweise 70, 85, 31 bei einer Stichprobe und 72, 61, 104 bei der anderen), dann definiert man sie in \mathbf{R} als Vektoren:

```
x1=c(70,85,31); x2=c(72,61,104)
```

- (a) Varianzheterogenität ($s_1 \neq s_2$) **Welch- t -Test** oder kurz Welch-Test

Der Welch- t -Test ist eine Modifikation des eigentlichen Student's - t -Test im 2-Stichprobenfall (siehe (b)), die keine Varianzhomogenität voraussetzt.

Wollen wir zeigen, dass die erstgenannte Stichprobe einen Stichprobenmittelwert hat, der signifikant *kleiner* ist als der Stichprobenwert der zweitgenannten ($H_1 : \bar{x}_1 < \bar{x}_2$, $H_0 : \bar{x}_1 \geq \bar{x}_2$), dann lautet der t -Test

```
t.test(x1,x2,alternative="less").
```

Im Output interessiert uns nur der P -Wert (die Wahrscheinlichkeit der Nullhypothese), in unserem Beispiel $P = 0.23$.

Mit dem `alternative="less"` haben wir \mathbf{R} die Seitigkeit unserer Alternativhypothese mitgeteilt: Die erstgenannte Stichprobe soll einen Wert nachweisen, der signifikant *kleiner* ist als die zweitgenannte.

Im umgekehrten Fall verwendet man `alternative="greater"`. Wollen wir nur einen Unterschied zwischen den beiden Stichproben nachweisen, verwenden wir `alternative="two.sided"`.

- (b) Varianzhomogenität ($s_1 \approx s_2$): **Student's t -Test**

Unterscheiden sich die Varianzen nicht signifikant voneinander, dürfen wir den Student's t -Test ausführen:

```
t.test(x1,x2,alternative="less",var.equal = TRUE).
```

Die entsprechenden Seitigkeiten `alternative="greater"` und `alternative="two.sided"` analog zum Fall des Welch-Tests.

4. Wenn $P < \alpha$, ist ein signifikanter Unterschied nachgewiesen und die Nullhypothese kann verworfen werden, andernfalls nicht.

R-Frage 39 *Du hast in deiner Semesterarbeit ja bereits eine neue Art Joghurtdeckel untersucht (R-Frage 38). Die Untersuchung verlief unbefriedigend. Jetzt probierst du es mit einer neuen Deckelart. Die neuen Werte sind 2.0 N, 2.1 N, 3.5 N, 1.8 N.*

Beweisen deine Messungen, dass die neuen Deckel besser sind als die zuerst untersuchten? (Wobei „besser“ hier heisst, dass die neuen Messwerte kleiner sind als die vorherigen. Du darfst Varianzhomogenität voraussetzen.)

R-Frage 40 *Mathematik steht im Ruf eines Faches, mit dem das männliche Geschlecht weniger Mühe bekundet als das weibliche. Ist das wirklich so? Wir werden diese Frage zwar hier nicht klären können. Aber wir können ja mal die Mathematik-Noten an der ~~an~~ untersuchen (alle Zeugnisnoten von 2 Jahrgängen UI-Studis). Du darfst Varianzhomogenität voraussetzen.*

Note	Frauen	Männer
2.0	—	—
2.3	1	1
2.5	2	2
2.8	—	1
3.0	1	5
3.3	1	2
3.5	3	5
3.8	6	1
4.0	8	13
4.3	6	8
4.5	3	16
4.8	—	9
5.0	4	6
5.3	1	—
5.5	2	4
5.8	1	1
6.0	2	4
n:	41	78

- a) Bekunden die Männer weniger Mühe als die Frauen? Was sagen denn die Daten dazu ($\alpha = 5\%$)?
- b) Ist der Durchschnitt der Damen signifikant über 4.0?
-

CHECKLISTE*Weisst du jetzt:*

- Wofür man den t -Test braucht?
- was der P -Wert besagt?
- welchen Einfluss das Signifikanz-Niveau auf das Resultat des t -Test hat?
- was der Unterschied ist zwischen 1-Stichproben-Fall und 2-Stichproben-Fall?
- welche Möglichkeiten der graphischen Darstellung der Daten es gibt?

Kannst du jetzt:

- einen t -Test durchführen?
- die bisherigen Fragen dieses Abschnittes ohne Hilfe lösen?

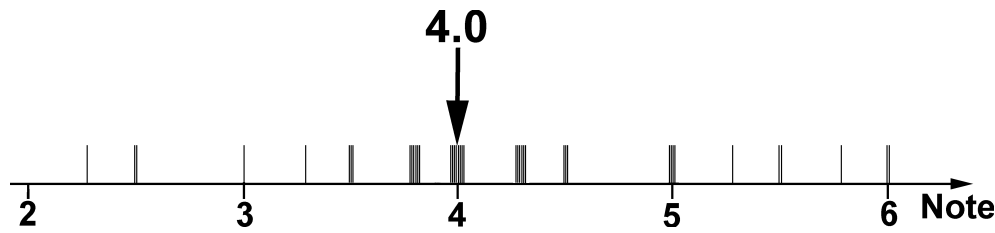
6.1.4 Visualisierung der Daten vor dem Testen

Bevor ich weitere Testvarianten aufzähle, will ich hier einen Zwischenhalt einlegen und auf einen wichtigen Aspekt hinweisen.

Ein t -Test ist schnell und problemlos durchgeführt; und er liefert ein konkretes Resultat, das sich in einem einzigen kurzen Satz formulieren lässt. Vielleicht ist es fast ein bisschen zu problemlos – jedenfalls ist es eine verbreitete Seuche, dass man Daten gar nicht mehr anschaut, sondern einfach drauflos testet. Auch in diesem Kapitel sowie in den Übungsaufgaben wird dir dieses Vorgehen zwar immer wieder vorgeführt; aber nur, weil das Testen eben den jetzt aktuellen Lernstoff darstellt. Ich möchte dir für die Praxis (z.B. in Semesterarbeiten) jedoch dringend zu einer anderen Vorgehensweise raten:

1. Zuerst schaust du die Daten an. Dafür musst du sie geeignet darstellen, am besten grafisch, so dass du sie wirklich *sehen*, dir ein *Bild* machen kannst.
2. Dann überlegst du dir aufgrund des visuellen Eindrucks, was die Daten wohl aussagen könnten. Formuliere eine oder mehrere Hypothesen!
3. Oft sprechen die Daten eine klare Sprache; in diesem Fall ist ein statistischer Test eigentlich nicht wirklich nötig (aber aus Gründen der Systematik manchmal trotzdem verlangt). Den statistischen Test braucht man eigentlich dann, wenn die Daten zwar auf einen Unterschied zwischen den verschiedenen Vergleichsseiten „hinweisen“, aber man ist nicht so recht sicher, ob es nicht auch Zufall sein könnte. Genau dies klärt ein statistischer Test dann mit mathematischen Mitteln.

Ich möchte diese Punkte noch etwas genauer erläutern. Zu 1.: *Wie* soll man die Daten darstellen? Grundsätzlich ist das beschreibende Statistik, und man kann sich in den entsprechenden Kapiteln der erstsemestrigen Mathematik inspirieren lassen. Aber nehmen wir uns doch einmal die **R**-Frage 40b) als Beispiel vor („Ist der Durchschnitt der Damen signifikant über 4.0?“). Man kann sich ein beispielsweise ein Bild machen, indem man die Werte auf einem Zahlenstrahl einträgt:



Bei grossen Datenmengen sind Säulendiagramme, Stabdiagramme oder Scatterplots geeignete Visualisierungs-Tools. Hier wird jeder einzelne Datenpunkt wiedergegeben. Aggregierte Darstellungen der Daten sind Boxplots oder Fehlerbalken wie in Abbildung 6.2: In Box-

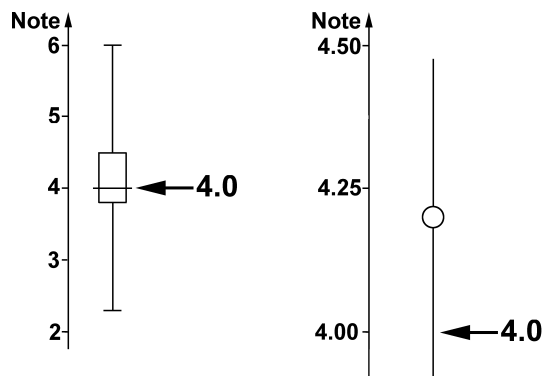


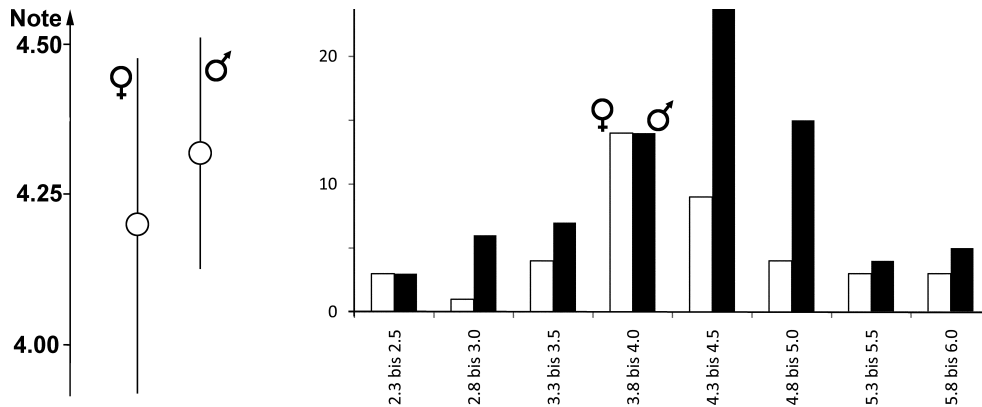
Abbildung 6.2: Links: Boxplot, Rechts: Fehlerbalken

plots und Fehlerbalken gehen nur die summarischen Streu- und Lagemasse ein. Dies ist eine Darstellungsmethode, die oft in Publikationen verwendet wird. Wie man die Länge des Fehlerbalkens definiert, ist nicht allgemeingültig festgelegt. Man kann grundsätzlich eine Standardabweichung σ (oder ein Mehrfaches davon) der Werte der Stichprobe auftragen. Dies gibt einem einen Eindruck für die Streuung der *Einzelwerte*. Im gegenwärtigen Zusammenhang interessiert aber mehr die Aussage bezüglich der Population. Für den wahren Populationswert kann man ein Konfidenzintervall berechnen (nach den Regeln von Abschnitt 4.5) und dieses als Fehlerbalken abbilden. Noch beliebter ist der Standard-Fehler oder ein Mehrfaches davon. 2 Standard-Fehler (aufgetragen jeweils nach oben und unten vom zentralen Wert) entsprechen ungefähr (!) einem 95%-Konfidenz-Intervall, 3 Standard-Fehler entsprechen ungefähr einem 99.75%-

Konfidenz-Intervall, der gesamte Fehlerbalken ist somit fdann4 bzw. 6 SE lang.

Was liest man nun von Auge aus diesen Figures heraus? Da der Referenzwert noch innerhalb des Fehlerbalkens (mit einer halben Länge von 2 SE) liegt ist die Differenz zum Stichprobenmittelwert tendenziell nicht signifikant. Auch die vorangegangenen Figures zeigen dem Auge keinen Unterschied, der nicht mehr als Zufallsschwankung zu verstehen ist.

Und in R-Frage 40a)? Der Datenvergleich kann wieder über Zahlenstrahl-, Boxplots- oder Fehlerbalkenbilder etc...gemacht werden. Z.B.:



(Links wieder mit $2 SE$). Wiederum ist ein Unterschied in der Lage der beiden Stichproben wahrzunehmen – wie wir es ja schon von den berechneten \bar{x}_i -Werten her wissen. Der Unterschied ist aber nicht gross im Vergleich zur Streuung der einzelnen Stichprobe, wesentlich kleiner als $2 SE$. Auch hier hätte man also das Resultat der Testrechnung wohl erraten können.

Frage 17 Skizziere und interpretiere von Auge die Daten der **R-Kartoffel-Aufgabe 37**.

R-Frage 41 Laden von Moodle das File `kartoffel.csv` herunter und lies es mit `kartoffel=read.csv("kartoffel.csv")` ein, hier handelt es sich um die Massen von 200 Kartoffeln mit Mittelwert 255 g und Standardabweichung 60 g, genau wie in der **R-Kartoffel-Aufgabe 37**.

a) Versuche Dir nun mit Dir bekannten **R**-Befehlen ein „Bild“ von den Daten zu machen.

b) Führe nun den einseitigen t -Test zum Referenzwert 240 g ausgehend von den einzelnen Stichprobenwerten durch. Vergleiche mit dem Resultat aus der **R-Kartoffel-Aufgabe 37**.

Die bisherigen Beispiele zum t -Test waren allesamt einseitig. Z.B.:

- Bei der Kartoffel-**R**-Frage 37 würde ein neuer Dünger, der *kleinere* Kartoffeln produziert, dazu führen, dass man den Dünger wegwirft, und nicht zu einem statistischen Test. Es interessiert ausschliesslich der Fall *grösserer* Kartoffeln.
- Bei der Mathematiknoten-**R**-Frage 40 wird die Möglichkeit, dass die *Damen besser* sind als die Herren erst gar nicht in Betracht gezogen.

R-Frage 42 Wie würde in den genannten Fällen jeweils die Fragestellung lauten, die zu einem zweiseitigen Test führt? (auch wenn die zugehörigen Alternativhypothesen dann vielleicht nicht besonders sinnvoll sind). Berechne die zugehörigen P -Werte. Was fällt Dir auf?

6.1.5 t -Test für zwei abhängige Stichproben

Wenn du in Erfahrung bringen willst, ob der Fluglärm in der Empfindung der Menschen am Zunehmen ist, wirst du dies vielleicht mit repräsentativen Umfragen im Abstand von einigen Jahren machen. Nun hast du in diesem Fall zwei Möglichkeiten:

①

Du wählst bei der ersten Umfrage eine Anzahl Personen aus. Beim zweiten Mal wählst du noch einmal neu eine Anzahl Personen aus. Die befragten Personen sind beim zweiten Mal nicht die gleichen wie beim ersten Mal.

②

Du wählst bei der ersten Umfrage eine Anzahl Personen aus. Ausser ihrer Meinung nimmst du auch ihre Adresse auf. Beim zweiten Mal befragst du noch einmal die gleichen Personen.

Frage 18 *Was glaubst du, welche Methode die exakteren Resultate liefert?*

Die Variante ② trifft man seltener an. Sie ist in vielen Fällen auch gar nicht realisierbar (z.B. kannst du eine Kartoffel ja nicht zuerst mit Dünger A wachsen lassen und anschliessend mit Dünger B). Wo sich Variante ② realisieren lässt, ist sie von der Aussagekraft her unbedingt vorzuziehen. Man spricht von „gekoppelten“, „gepaarten“, „verbundenen“ oder „abhängigen“ Stichproben.

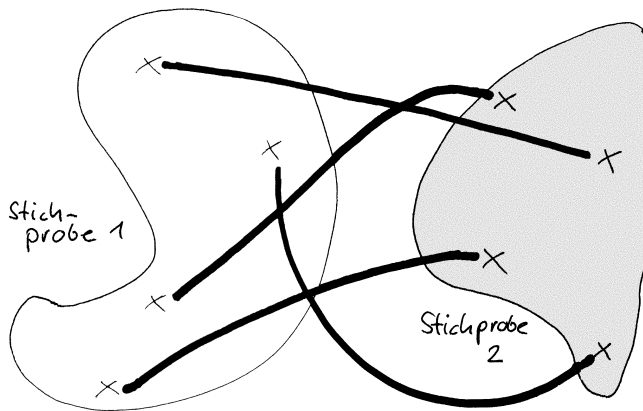
Die Variante ① ist der Normalfall. Man spricht von „unabhängigen“, „nicht gekoppelten“, „ungepaarten“ Stichproben.

Frage 19 *Wie könnte man das Beispiel mit den Mathematiknoten ([R-Frage 40](#)) so abändern, dass es zu einem Problem mit gepaarten Stichproben wird? (Die Frage darf dabei durchaus einen anderen Sinn erhalten.)*

Frage 20 Beurteile bei folgenden Situationen, welcher Fall vorliegt:

- 1-Stichproben-Fall
 - 2 unabhängige Stichproben
 - 2 abhängige Stichproben
- a) Ein Autohersteller sammelt Verbrauchsdaten einer normalen Benzinsorte mit und ohne einen speziellen Zusatz. Er wählt 10 Autos aus verschiedenen Klassen aus und dazu je einen Fahrer. Jedes Auto fährt je einmal mit und einmal ohne Zusatz eine gewisse Strecke ab.
 - b) Du planst den Bau eines Einkaufszentrums. Wichtig für die Wahl des Standorts ist u.a. das Haushalts-Einkommen der Bevölkerung in der Umgebung. Die Stadtverwaltung behauptet, dass der Durchschnitt bei 115 kFR/Jahr liegt. Du machst eine Umfrage und kommst auf nur 99 kFR/Jahr.
 - c) Du willst zwei Glühbirnen-Fabrikate vergleichen und u.a. herausfinden, welches Fabrikat die grössere Lebensdauer hat. Du misst bei je 10 Glühbirnen, wie lange sie brennen.
 - d) Ein Hersteller von Leuchtstoffröhren behauptet, seine Röhren hätten im Schnitt eine Lebensdauer von mindestens 9000 h. Du willst das nachprüfen, indem du die aufgezeichneten Lebensdauern von 15 Röhren analysierst.

Bevor wir das Testrezept angehen, noch ein nützlicher Hinweis zu gepaarten und ungepaarten Stichproben: Paarung setzt voraus, dass es zwischen den Elementen der beiden Stichproben eine objektive Zuordnung gibt:



Man muss diese Zuordnung in Worten ausdrücken können (z.B. „simultane Werte gehören zusammen“).

Dies setzt wiederum voraus, dass die beiden Stichproben gleichen Umfang haben: $n_1 = n_2$. Gleicher Umfang bedeutet noch nicht zwingend Paarung, aber bei ungleichen Umfängen ist Paarung ausgeschlossen.

Durchführung des Tests

 t -Test, 2-Stichproben-Fall, abhängige Stichproben

Test-Frage: Unterscheidet sich der arithmetische Mittelwert einer Stichprobe signifikant vom arithmetischen Mittelwert einer anderen Stichprobe, die mit der ersten gekoppelt ist?

Test-Rezept:

1. Formuliere H_0 und H_1 .
2. α wählen.
3. Zuerst definiert man die Zahlenwerte jeder einzelnen Stichprobe als zwei separate Vektoren: (z.B. 70, 85, 31 und 72, 61, 104):

```
x1=c(70,85,31)
```

```
x2=c(72,61,104)
```

Dabei muss darauf geachtet werden, dass die Reihenfolgen der Kopplung entsprechen (dass z.B. das Element 85 in **x1** das Pendant zu 61 in **x2** ist. Wollen wir beispielsweise zeigen, dass der arithmetische Mittelwert der erstgenannten Stichprobe kleiner als der arithmetische Mittelwert der zweitgenannten ist ($H_1 : \bar{x}_1 < \bar{x}_2$, $H_0 : \bar{x}_1 \geq \bar{x}_2$), braucht der t -Test nun die Option **paired=TRUE**:

```
t.test(x1,x2,alternative="less",paired=TRUE)
```

Im Output interessiert uns einfach der p -Wert, in unserem Beispiel $P = 0.31$.

Falls man erwarten würde, dass **x1** grössere Werte enthält als **x2**, würde man **alternative="greater"** wählen.

Im 2-seitigen Fall lautet die Option **alternative="two.sided"**.

4. Dann wieder α und P vergleichen.

R-Frage 43 a) In einer Anlage mit Kirschbäumen wird der Witterungseinfluss auf den Ertrag untersucht ($\alpha = 5\%$), indem man in zwei vom Wetter her verschiedenen Jahren die Erträge der Bäume vergleicht:

Baum	1	2	3	4	5	6	7	8
Ertrag 2013 [kg]	36.0	31.5	34.0	32.5	35.0	31.5	31.0	35.5
Ertrag 2014 [kg]	34.0	35.5	33.5	36.0	39.0	35.0	33.0	39.5

b) Wie sähe das Resultat aus, wenn man die Kopplung ignorierte?

6.1.5.1 Gekoppelte Stichproben als 1–Stichproben–Fall

Bei vielen Tests gibt es von der Ausgangslage her die 3 Varianten

- (A) 1–Stichproben–Fall,
- (B) 2 unabhängige Stichproben,
- (C) 2 abhängige Stichproben.

Entsprechend habe ich versucht, dir die Unterschiede zwischen diesen 3 Situation klar zu machen. Und nun muss ich dir aber sagen, dass 2 der 3 Situationen eigentlich äquivalent sind. Nicht etwa die Fälle (B) und (C), die oberflächlich betrachtet am nächsten beieinander zu liegen scheinen, sondern (A) und (C) sind äquivalent.

Im Fall (C) bildet **R** zuerst die Differenzen zwischen den Wertepaaren ($x_1 - x_2$). Dieser Differenzvektor stellt insgesamt *eine* Stichprobe dar. Fragt sich nur noch, was dann hier der Vergleichswert μ ist, mit dem man diese einzelne Stichprobe vergleicht. Antwort: $\mu = 0$, denn die Nullhypothese sagt ja, dass zwischen den Stichproben keine Differenz besteht.

R–Frage 44 *Um dich selber davon zu überzeugen, nimmst du nochmal das Kirschbaum–Beispiel R–Frage 43. Berechne den P–Wert, indem du den Differenzvektor mit 0 vergleichst.*

In vielen Statistik–Büchern und auch in den noch folgenden Testrezepten in diesem Skript werden die Fälle (A) und (C) nicht mehr beide aufgeführt, sondern nur noch einer davon.

Die Äquivalenz von 2 abhängigen Stichproben mit dem 1–Stichproben–Fall kann man auch für die Visualisierung ausnützen, indem man die Paardifferenzen d_i darstellt und mit $\mu = 0$ vergleicht.

Frage 21 a) Visualisiere die Situation der „Kirschbaum“–R–Frage 43 als 1–Stichproben–Fall.
 b) Wie sähe deine Visualisierung aus, wenn es dabei um unabhängige Stichproben ginge?

Frage 22 *Im letzten Abschnitt ging es um Schätzungen, jetzt geht es um Tests. Was ist eigentlich der Unterschied oder der Zusammenhang zwischen diesen beiden Gruppen von Methoden? Wann braucht man was?*

CHECKLISTE

Weisst du jetzt:

- was der Unterschied ist zwischen einseitigem Fall und zweiseitigem Fall?
- was der Unterschied ist zwischen gepaarten Stichproben und ungepaarten Stichproben?

Kannst du jetzt:

- einen t–Test für gepaarte Stichproben durchführen?
- die bisherigen Fragen dieses Abschnittes ohne Hilfe lösen?

6.2 Überprüfen der Normalverteilungsannahme

Der t -Test ist, insbesondere bei grossen Stichproben, robust gegen Abweichung von der Normalverteilungsannahme. Doch insbesondere bei kleinen Stichproben müssen wir manchmal die Normalverteilungsannahme überprüfen - entweder graphisch mit einem Normal-Quantil-Diagramm oder mittels eines statistischen Tests. Statistische Tests haben aber den Nachteil, dass ihr Resultat stark von der Grösse der Stichprobe abhängt.

6.2.1 Normal-Quantil-Diagramm oder QQ-Plot

Normal-Quantil-Diagramme (auch Quantil-Quantil-Plots oder kuz QQ-Plots genannt) sind eine relativ einfache Methode, um graphisch festzustellen, ob die Werte mehr oder weniger normalverteilt sind: Wir stellen die Methode anhand eines Beispiels vor:

Gegeben sei eine Stichprobe mit den Gewichten von $n = 8$ Personen (in kg).
`gewicht=c(65,62,81,88,73,75,71,79)`

1. Man bringt den Datenvektor der Stichprobe mit dem `sort`-Befehl in eine aufsteigende Reihenfolge:

```
gewichtsorted=sort(gewicht)
```

d_i ist die i -te Zahl des Datenvektors `gewichtsorted`.

2. Man teilt nun die (Standard-)Normalverteilung in $n + 1$ gleiche Flächenanteile und bestimmt die zu diesen Flächenanteilen gehörigen z -Werte. Diesen Flächenanteile ordnet man die n Quantilen $q_i = \frac{i}{n+1}$, $i = 1, \dots, n$ zu. In unserem Beispiel: $q_1 = 1/9, q_2 = 2/9, \dots, q_8 = 8/9$. Anschliessend bestimmt man mit `qnorm` diejenigen z -Werte, die den Flächenanteilen (=Quantilen) der Standardnormalverteilung entsprechen.

```
quantile=seq(1,n)/(n+1)
```

```
z=qnorm(quantile)
```

z_i ist die i -te Zahl des Datenvektors `z`.

3. Falls die Werte der Stichprobe tatsächlich von einer Normalverteilung stammen, entsprechen sie etwa den gleichen Quantilen der Standardnormalverteilung. Das bedeutet, dass **die die Datenpaare (z_i, d_i) in einem Streudiagramm auf einer Geraden liegen**. Wir können uns durch `plot(z,gewichtsorted)` davon überzeugen.
4. Lässt sich durch alle diese Zahlenpaare nur schlecht eine Gerade legen, so sind die Daten nicht normalverteilt.

6.2.1.1 QQ-Plot in R

Der QQ-Plot ist eine sehr weitverbreitete Methode zur visuellen Inspektion der Normalverteilung. Obiges Rezept ist deshalb schon implementiert und der QQ-Plot kann in **R** einfach direkt ausgehend vom Datenvektor mittels

`qqnorm(gewicht)` erzeugt werden.

`qqline(gewicht)`

legt eine Gerade durch die zentralen 25-75% der Daten (die sogenannten Q1-Q3 Quartile, die Du schon beim Boxplot kennengelernt hast.).

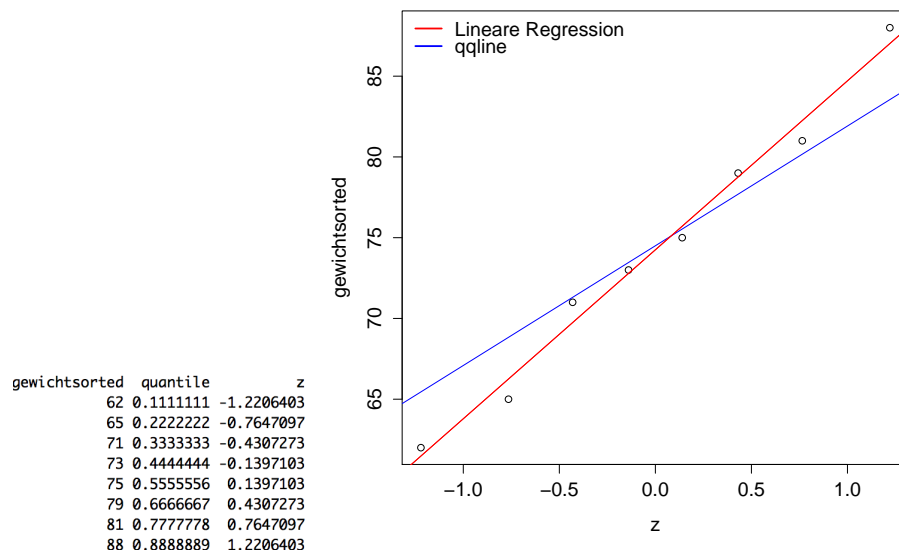


Abbildung 6.3: QQ-Plot der Gewichtsdaten. Rot: Trendgerade der linearen Regression zwischen Gewicht und den z-Werten der Quantile der Normalverteilung. Blau: qqline der Q1-Q3 Quartile.

6.2.2 Statistische Tests zur Überprüfung der Normalverteilungsannahme

Der Shapiro-Wilk-Test ist ein Test grosser Güte, um die Normalverteilung der Stichproben-daten `x` zu überprüfen. In **R**:

`shapiro.test(x)`, wo `x` der Datenvektor der Stichprobe ist.

Der Shapiro-Wilk-Test funktioniert nicht gut bei Stichproben mit vielen identischen Werten. Dann können wir alternativ den parameterfreien, weniger mächtigen Kolmogorov-Smirnov-Test anwenden:

`ks.test(x, pnorm, mean=mean(x), sd=sd(x))`.

R-Frage 45 *Erstelle den QQ-Plot der obigen Gewichtsdaten, einmal dem obigen Beispiel folgend, einmal mit `qqnorm`. Berechne sodann für die Gewichtsdaten der obigen 8 Personen (abhängige Variable) und die z -Werte der Quantile (unabhängige Variable) die Trendgerade der linearen Regression. Zeichne diese mit `abline` in Deinen QQ-Plot ein. Vergleiche mit `qqline`. Führe sodann den Shapiro-Wilk-Test aus. Was ist hier die Nullhypothese? Und was sagt Dir nun der P -Wert?*

6.3 Überprüfen der Varianzhomogenität: Der F -Test

Beim t -Test für unabhängige Stichproben haben wir unterschieden zwischen homogenen und heterogenen Varianzen. Falls nicht offensichtlich ist, welcher Fall vorliegt, so kann man es mit dem F -Test (benannt nach R.A. Fisher, dem vielleicht leuchtendsten Namen in der Statistik-Geschichte) herausfinden. Der F -Test überprüft, ob zwei Stichproben sich hinsichtlich ihrer Varianz (oder Standardabweichung) signifikant unterscheiden. Die zugehörige F -Verteilung hängt sowohl von den Varianzen als auch den Freiheitsgraden der beiden Stichproben ab. Im Folgenden werden wir die statistischen Tests meist am Beispiel zweiseiti-

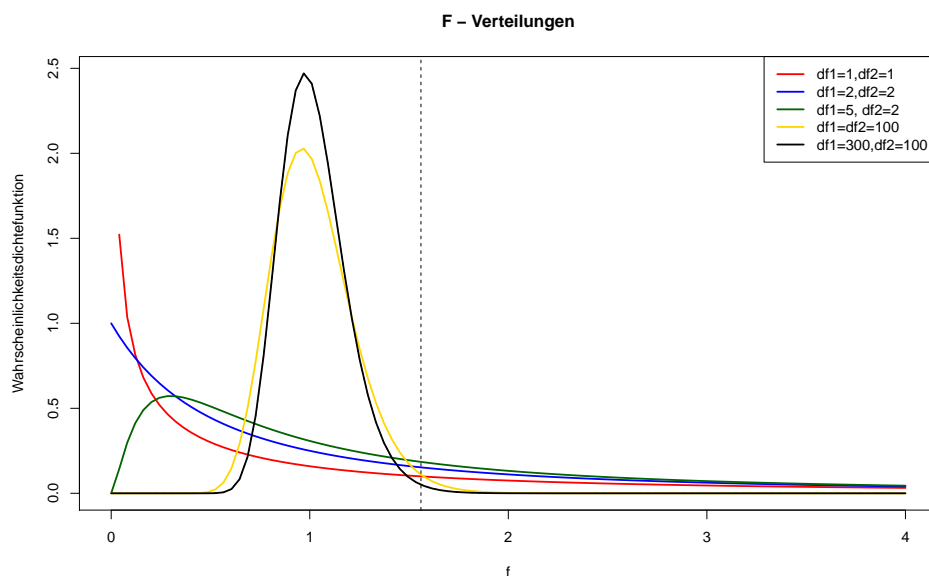


Abbildung 6.4: F -Verteilung für verschiedene Freiheitsgrade der Stichproben. Die Prüfgrösse $f = \frac{s_1^2}{s_2^2}$ (aufgetragen auf der x -Achse) ist immer positiv. Gestrichelte Linie: $f=1.56$

ger Fragestellungen (in **R** `alternative="two.sided"`) durchspielen. Die entsprechenden einseitigen Fragestellungen lassen sich in **R** einfach durch die entsprechenden Optionen

`alternative="greater"` oder `alternative="less"` berechnen.

F-Test, 2-Stichproben-Fall, unabhängige Stichproben

Test-Frage:

Unterscheidet sich

die Standard-Abweichung s_1 einer Häufigkeitsverteilung 1

signifikant von

der Standard-Abweichung s_2 einer Häufigkeitsverteilung 2?

Test-Rezept:

1. Formuliere H_0 und H_1 . 2. Lege α fest.

3. Berechnung des P-Wertes:

(a) Die einzelnen Stichprobenwerte liegen vor

Beispiel: 70, 85, 31 bei einer Stichprobe und 72, 61, 104 bei der anderen.

Dann definiert man sie in **R** als Vektor:

```
x1=c(70,85,31); x2=c(72,61,104)
```

Dann geht der F-Test so:

```
var.test(x1,x2,alternative="two.sided")
```

Im Output interessiert wieder einfach der P -Wert, in unserem Beispiel $P = 0.78$.

(b) Berechnung des P -Wertes aus Stichprobenumfängen und deren Standardabweichungen

R berechnet im Hintergrund den P -Wert folgendermassen (Zahlenwerte aus Beispiel oben):

Erste Stichprobe mit Umfang $n_1 = 3$ und Standardabweichung $s_1 = 27.88$;

zweite Stichprobe mit Umfang $n_2 = 3$ und Standardabweichung $s_2 = 22.34$;

Diejenige Stichprobe, deren Standardabweichung grösser ist, erhält die Nummer 1, die andere die Nummer 2. Zuerst berechnen wir das Verhältnis der Varianzen, die

Prüfgrösse $f = \frac{s_1^2}{s_2^2}$. In **R** übertragen:

```
n1=3; s1=27.88
```

```
n2=3; s2=22.34
```

```
f=(s1/s2)^2
```

In unserem Beispiel entspricht der P -Wert von 0.78 der (doppelten) Fläche des rechten „Schwanzes“ der F-Verteilung mit Freiheitsgraden $df_1=3-1=2$ und $df_2=3-1=2$ rechts von der Prüfgrösse $f = \frac{s_1^2}{s_2^2} = 1.56$ (blaue Linie Abbildung 6.4):

```
2*pf(f,n1-1,n2-1,lower.tail=FALSE)
```

Der Faktor 2 rührt vom zweiseitigen Testen. Im einseitigen Fall ($H_1 : s_1 > s_2$) ergibt sich der P -Wert direkt zu `pf(f,n1-1,n2-1,lower.tail=FALSE)`

4. Ist $P < \alpha$, so ist s_1 signifikant von s_2 verschieden (zweiseitiger Test).

R-Frage 46 *Eine Stichprobe mit Umfang 35 und Standard-Abweichung 3.7 wird mit einer anderen Stichprobe (Umfang 28, Standard-Abweichung 6.4) verglichen. Unterscheiden sich die Streuungen der beiden Stichproben signifikant ($\gamma = 95\%$)?*

R-Frage 47 *Zurück zu den Joghurtdeckeln (R-Aufgaben 38 und 39). Unterscheiden sich die Streuungen der beiden Messungen signifikant ($\gamma = 95\%$) ?*

R-Frage 48 *Und auch zu den den Mathematiknoten von Männern und Frauen (R-Aufgabe 40) kehren wir zurück. Unterscheiden sich die Streuungen der beiden Messungen signifikant ($\gamma = 95\%$)?*

Frage 23 *Du kennst jetzt den t -Test und den F -Test. Aber was ist eigentlich der Unterschied zwischen diesen 2 Tests?*

Kapitel 7

Parameterfreie Tests

Parametrische Verfahren setzen voraus, dass die bearbeiteten Daten erstens metrisch und zweitens normal- oder in einer anderen vorgeschriebenen Weise verteilt sind.

Daraus ergeben sich Einschränkungen:

- In vielen Anwendungsbereichen der Statistik sind die Daten eher ordinal. Häufig werden in Umfragen die Meinungen der Befragten in ordinalen Kategorien abgefragt. Beispiele:
 - Du sammelst Daten zum Geschmack einer neuen, von Dir gezüchteten Apfelsorte. Die Bewertung des subjektiven Süsseempfindens erfolgt auf einer Skala von 1=„mässig süss“, 2= „süß“ bis 3 = „sehr süß“.
 - Du willst wissen, ob das neue Leuchtkonzept eine signifikante Aufhellung der Zufriedenheit bei den Angestellten eines Büros gebracht hat. Ihr macht dafür eine Umfrage, wobei jeder Mitarbeitende 3 Monate vor und 3 Monate nach der Installation der neuen Beleuchtung gefragt wird, wie gross die Zufriedenheit mit den Lichteinstellungen ist. Die Bewertung erfolgt auf einer Skala von 1=„übehaupt nicht“ bis 10= „völlig zufrieden“.

Auch bei ordinalen Daten stellt sich die Frage, ob das Zentrum einer Stichprobe (am besten mit dem Median anzugeben) besser oder schlechter als dies oder jenes ist. Hier braucht man zwingend eine nicht-parametrische Alternative.

Übrigens ist für nicht-metrische Daten so etwas wie die Normalverteilung gar nicht definiert.

- Stell dir vor, du hast einen Stapel metrischer Zahlen vor dir, die du interpretieren sollst. Um sie parametrisch testen zu können, müsstest du je nach Verfahren (z. B. beim t -Test) korrekterweise erst nachweisen, dass sie tatsächlich normalverteilt sind. Normalverteilung ist ja keineswegs selbstverständlich; man könnte sich doch z.B. auch eine schiefe Verteilung denken; für gewisse Fragestellungen ist eine schiefe Verteilung sogar der Normalfall.

Doch all diese (Vorab)-Testerei wird somit recht mühsam. Und speziell bei kleinen Da-

tensätzen ($n < 100$) wird sich die Frage ohnehin nie in vertrauensserweckender Weise klären lassen. Vielleicht doch lieber sofort einen nicht-parametrischen Test machen?

Vor allem bei kleinen Stichproben-Umfängen sind parameterfreie Tests weniger problematisch. Bei grossen Umfängen (sagen wir dreistelligen) ist die Normalverteilungs-Voraussetzung hingegen meist bedeutungslos; bei grossen Stichproben sind die meisten Tests robust gegen die Verletzung der Normalverteilungs-Voraussetzung.

Parametrische Tests sind viel bekannter als die nicht-parametrischen Alternativen. Die meisten „Nebenbei-Statistiker“ – und solche machen wahrscheinlich den weitaus grössten Teil der statistischen Tests auf diesem Planeten — verwenden parametrische Tests ohne viel über die Normalverteilungs-Annahme nachzudenken. Den Profi-Statistikern dagegen stehen dabei die Haare zu Berge, und manche Profis sagen sogar, dass sie zumindest bei kleinen Datensätzen *nie* parametrische Tests einsetzen.

Parameterfreie Tests (auch verteilungsfreie oder nichtparametrische Tests genannt) sind nicht komplizierter als parametrische. Was ist dann ihr Nachteil? Erstens sind sie weniger bekannt, vielleicht weil sie historisch erst entwickelt wurden, als die wichtigsten parametrischen Tests schon allgemein im Gebrauch waren. Objektiv wiegt der zweite Nachteil schwerer: wie bereits in Abschnitt 5.3 erwähnt, haben parameterfreie Tests eine geringere Teststärke, oft auch in deutschsprachiger Statistik-Literatur mit dem englischen Wort „Power“ bezeichnet: *Bei gleichen Daten liefern parameterfreie Tests einen grösseren P-Wert, man kann die Aussage deshalb nur mit einer höheren Irrtumswahrscheinlichkeit machen.* Oder andersherum gedacht: man muss bei parameterfreien Tests grössere Stichproben haben, um beweisen zu können, was man gerne beweisen möchte. Aber häufig hat man im Messalltag nicht so viele Daten, wie man gerne hätte.

7.1 Anteilstest

Der Anteilstest eignet sich für dichotome Daten (beliebige Daten lassen sich auf dichotome zurückführen, indem man solange Klassen zusammenfasst, bis nur noch zwei Klassen übrig bleiben.) Wie schon bei den Parameter-Schätzungen geht es auch hier wieder um den Anteil $\hat{p} = k/n$ der Häufigkeit eines der beiden Variablen-Werten (den wir als „Erfolg“ titulieren, vgl. Abschnitt 4.2.1).

Anteilstest, 1-Stichproben-Fall

Test-Frage:

Allgemein: Unterscheidet sich ein in einer Stichprobe vom Umfang n beobachteter Erfolgsanteil \hat{p} signifikant von einem vorgegebenen Wert π_0 ?

Im Beispiel unten: Ist \hat{p} signifikant kleiner als π_0 ?

Test-Rezept:

1. Formuliere H_0 und H_1 . 2. Lege α fest.

3. Berechnung des P-Wertes

Beispiel: Wenn in einer Stichprobe 40 Erfolge und 70 Misserfolge gezählt werden, und man damit auf dem 95% Niveau nachweisen will, dass der Erfolgsanteil $\hat{p} = 40/110 = 0.36$ in der ganzen Population signifikant kleiner als 45% ist, so liefert

```
n=40+70
k=40
pi0=0.45
binom.test(k,n,pi0,alternative="less")
```

den entsprechenden P -Wert, nämlich $P = 0.04$.

Die Alternativen "**greater**" und "**two.sided**" stehen wieder zur Auswahl, analog zu dem, was du beim t -Test gesehen hast.

4. Ist $P < \alpha$ kann die Nullhypothese verworfen werden, sonst nicht.

In unserem Beispiel ist $0.04 < \alpha$, die Nullhypothese kann somit verworfen werden, die Stichprobe ist damit beweiskräftig.

-
- R-Frage 49** a) Du kontrollierst in einer Produktionslinie für Pommes-Chips-Säcke 100 Stück und findest, dass 3 davon die Spezifikationen nicht erfüllen. Der Vertrag mit dem Lieferanten besagt, dass maximal 5% „schlechte“ Säcke geliefert werden dürfen. Beweist deine Stichprobe, dass der Vertrag mit 99% iger Sicherheit eingehalten wird?
- b) Am nächsten Tag kontrollierst Du 1000 Pommes-Chips-Säcke und findest, dass 30 davon die Spezifikationen nicht erfüllen. Kannst Du an diesem Tag den Vertrag mit dem Lieferanten mit 99% iger Sicherheit einhalten?
-

Anteilstest, 2-Stichproben-Fall, 2 unabhängige Stichproben

Test-Frage: Unterscheiden sich die Anteile \hat{p}_1 und \hat{p}_2 zweier unabhängiger Stichproben vom Umfang n_1 bzw. n_2 signifikant von einander?

Test-Rezept:

1. Formuliere H_0 und H_1 . 2. Lege α fest.

3. Berechnung des P-Wertes

Beispiel: Angenommen eine erste Stichprobe hat $n_1 = 110$ Umfang und enthält $k_1 = 40$ Erfolge ($\hat{p}_1 = k_1/n_1 = 40/110 = 0.36$); bei einer zweiten Stichprobe sind es $n_2 = 120$ Umfang und $k_2 = 54$ Erfolge ($\hat{p}_2 = k_2/n_2 = 54/120 = 0.45$). Unterscheiden sich die beiden Stichproben im Erfolgs-Anteil signifikant ($\alpha = 5\%$) voneinander?

Das kann man mit **R** so herausfinden:

```
n1=110; k1=40
```

```
n2=120; k2=54
```

Fasse die Erfolge **k1** und **k2** der beiden Stichproben im Datenvektor **k** zusammen:

```
k=c(k1,k2)
```

Fasse die Stichprobenumfänge **n1** und **n2** im Datenvektor **n** zusammen:

```
n=c(n1,n2)
```

Dann liefert

```
prop.test(k,n,alternative="two.sided")
```

den zugehörigen P-Wert: $P = 0.23$.

Die Alternativen **"greater"** und **"two.sided"** stehen wieder zur Auswahl, analog zu dem, was du beim t -Test gesehen hast.

4. Ist $P < \alpha$ kann die Nullhypothese verworfen werden, sonst nicht.

In unserem Beispiel ist $P = 0.23 > \alpha$ und die beiden Stichproben unterscheiden sich im Erfolgsanteil nicht signifikant voneinander.

R-Frage 50 Von 670 Beizen im Kanton Weingau waren im Jahre 2005 deren 27 auf asiatischen Food spezialisiert. Ende 2014 waren von noch 644 Beizen deren 54 auf asiatisch spezialisiert. Kann man davon ausgehen, dass die Zunahme der asiatischen Beizen ein tatsächlicher Trend ist ($\alpha = 2.5\%$)?

Frage 24 Du hast ein alkoholfreies Ultra-Bio-Light-Bier erfunden. Du bist überzeugt, dass man es geschmacklich nicht von Normalbier unterscheiden kann. Aus deinem Bekanntenkreis wählst du 20 Leute aus, die behaupten, sie könnten jedes alkoholfreie Bier von normalem unterscheiden, und du lädst sie zur Degustation. Alle geben einen überzeugten Tipp ab, welches der beiden vorgesetzten Gläser alkoholfreies Bier enthält (nachdem sie den Mund so voll genommen haben, können sie nicht gut anders). 11 tippen richtig und 9 falsch. Interpretation?

a) Überlege dir zuerst, ob es sich um einen 1-Stichproben-Fall oder um einen 2-Stichproben-Fall handelt. Welches ist/sind die Stichprobe(n)?

b) Überlege dir, welches Resultat bei der Degustation zu erwarten ist, wenn dein Bier tatsächlich ununterscheidbar ist von Normalbier.

R-Frage 51 *Und jetzt, was sagt R dazu?*

Frage 25 *Was war in der Bier-Frage überhaupt die Nullhypothese? Und was haben die Daten bewiesen?*

Frage 26 *Worin unterscheidet sich der Anteilstest hauptsächlich von den vorher besprochenen t - und F -Tests?*

CHECKLISTE

Weisst du jetzt:

- *Wie man graphisch und mit statischen Tests die Normalverteilungsvoraussetzung überprüfen kann*
- *wann es einen F -Test braucht?*
- *wann es einen Anteils-Test braucht?*

7.2 Wilcoxon-Test

Es gibt zahlreiche parameterfreie Alternativen zum t -Test. Die bekannteste Alternative ist der Wilcoxon-Test, der (anstatt Mittelwerte wie beim t -Test) die Mediane der Stichproben vergleicht. Analog zum t -Test müssen wir wieder die Fälle unabhängiger Stichproben und abhängiger Stichproben sowie den Einstichprobenfall gegen Referenzwert unterscheiden.

- **Der Wilcoxon-Test für unabhängige Stichproben: U -Test**

Der Wilcoxon-Test für unabhängige Stichproben wird oftmals als U -Test (Eselsbrücke: „U“ für unabhängig) bezeichnet. Weitere gebräuchliche Bezeichnungen des U -Tests sind:

Wilcoxon-Rangsummen-Test, Wilcoxon-Mann-Whitney-Test, Mann-Whitney-U-Test oder Mann-Whitney-Test. All diese Namen werden synonym in der Literatur verwandt - die vielen Namen für den gleichen Test deuten schon an, wie wichtig er ist. Im Gegensatz zum t -Test gehen beim U -Test nicht die Stichprobenwerte selbst, sondern ihre **korrigierten Ränge** in die Prüfgrösse U ein (daher auch der Name Wilcoxon-Rangsummen-Test).

- **Wilcoxon-Test im Einstichprobenfall gegen Referenzwert oder Wilcoxon-Test abhängiger Stichproben:**

Wilcoxon-Vorzeichen-Rang-Test

Handelt es sich um zwei abhängigen Stichproben wird, wie beim t -Test für abhängige Stichproben, zuerst der Differenzenvektor $\mathbf{d}=\mathbf{x1}-\mathbf{x2}$ der beiden Stichproben $\mathbf{x1}$ und $\mathbf{x2}$ gebildet. Beim Berechnen der Prüfgrösse werden sodann die korrigierten Ränge der Stichprobenpaare d_i mit positiven Differenzen mit den korrigierten Rängen der Stichprobenpaare mit negativen Differenzen verglichen (die genauen Details der Berechnung ersparen wir uns hier).

Wilcoxon-Test, 2-Stichprobenfall für zahlenmässige Daten

Test-Frage: *Unterscheidet sich der Median \tilde{x}_1 einer Stichprobe signifikant vom Median \tilde{x}_2 einer anderen Stichprobe?*

Test-Rezept:

1. Formuliere H_0 und H_1 . 2. Lege α fest.
3. Berechnung des P-Wertes:
Falls die Daten zahlenmässig vorliegen (auch wenn vielleicht gar nicht echt metrisch), definierst du die Stichproben wie als Datenvektoren. Beispiel:

```
x1=c(70,85,31,36)
x2=c(71,86,35,33)
```

 Bei ungepaarten Stichproben: U-Test

```
wilcox.test(x1,x2,alternative="two.sided")
```

 den P -Wert, im vorliegenden Beispiel $P = 0.8857$.
 Wie beim t -Test gibt es wieder die Varianten

```
alternative="greater"
```

 and

```
alternative="less"
```


 Bei gepaarten Stichproben: Wilcoxon-Vorzeichen-Rang-Test

```
wilcox.test(x1,x2,alternative="two.sided",paired=TRUE)
```

 (P -Wert = 0.5807).
4. Ist $P < \alpha$ kann die Nullhypothese verworfen werden, sonst nicht.
In unserem Beispiel kann somit sowohl bei den ungepaarten als auch bei den gepaarten Stichproben die Nullhypothese nicht verworfen werden, die Mediane der Stichproben unterscheiden sich nicht signifikant.

Wilcoxon-Test, 1-Stichproben-Fall für zahlenmässige Daten:Wilcoxon-Vorzeichen-Rang-Test

Test-Frage: *Unterscheidet sich der Median \tilde{x}_1 einer Stichprobe signifikant von einem Referenzwert μ ?*

Test-Rezept:

1. Formuliere H_0 und H_1 . 2. Lege α fest.
3. Berechnung des P-Wertes:
Beispiel: Definiere wie oben die Stichprobe als Vektor

```
x1=c(70,85,31,36)
```

 Vergleiche nun $x1$ mit einem Referenzwert (beispielsweise $\mu = 50$):

```
wilcox.test(x1,mu=50,alternative="two.sided")
```

 Gibt $P = 0.625$.
 Gepaarte Stichproben als 1-Stichproben-Fall mit Referenzwert $\mu = 0$
 Mit $x1$ und $x2$ erhalten wir

```
wilcox.test(x1-x2,mu=0,alternative="two.sided"),
```

 was zum selben P -Wert ($P = 0.58$) führt wie beim gepaarten 2-Stichprobenfall.
4. Ist $P < \alpha$ kann die Nullhypothese verworfen werden, sonst nicht.

Wilcoxon-Test für ordinale Daten, die nicht in Zahlenform vorliegen

Test-Frage: Unterscheidet sich der Median \tilde{x}_1 einer Stichprobe signifikant vom Median \tilde{x}_2 einer anderen Stichprobe?

Test-Rezept:

1. Formuliere H_0 und H_1 . 2. Lege α fest.

3. Berechnung des P-Wertes:

Beispiel:

```
x1=c("nass","nass","trocken","nass","feucht")
x2=c("trocken","feucht","trocken","nass","trocken")
```

`wilcox.test` kann nicht direkt mit diesen Angaben arbeiten. Stattdessen muss man die **Worte (genauer die Zeichenketten) gemäss einer von uns definierten Reihenfolge in Zahlen konvertieren**: Nehmen wir nun an, dass die Reihenfolge "trocken" < "feucht" < "nass" gelten solle. Nun ersetzen wir "trocken" durch 1, "feucht" durch 2 und "nass" durch 3 (z. B. mit der `find` und `replace`-Funktion in Rstudio) und erhalten

```
r1=c(3,3,1,3,2)
r2=c(1,2,1,3,1)
```

Weiter wie mit metrischen Daten:

U-Test:

```
wilcox.test(r1,r2,alternative="two.sided")
```

Mit dem Resultat: $P = 0.2188$.

Analog:

Wilcoxon-Test, gepaart: Wilcoxon-Vorzeichen-Rang-Test

```
wilcox.test(r1,r2,alternative="two.sided",paired=TRUE)
```

($P = 0.1736$).

Wilcoxon Test, 1 Stichprobenfall: Wilcoxon-Vorzeichen-Rang-Test

```
x1=c("nass","nass","trocken","nass","feucht")
wilcox.test(r1,mu=3,alternative="greater")
```

($P = 0.04449$).

4. Ist $P < \alpha$ kann die Nullhypothese verworfen werden, sonst nicht.

Der `wilcox.test` liefert oft eine Warnmeldung, die besagt, dass der P -Wert nicht besonders exakt ist, meist weil die Stichproben zu klein sind. Wir nehmen es zur Kenntnis und leben bei unseren einfachen Beispielen damit.

Die P -Werte fallen mit `wilcox.test` in der Regel etwas höher aus als mit `t.test`. So äussert sich die etwas geringere Teststärke (Power) des parameterfreien Verfahrens.

R-Frage 52 Die folgende Tabelle enthält verschiedene Messungen des Proteingehaltes zweier Bohnensorten A und B [alle Angaben in g Proteine pro kg Bohnen]:

Sorte A	Sorte B
22.6	17.0
16.0	19.2
21.2	19.1
21.4	19.6
19.4	16.8
18.7	18.4
23.1	
14.3	

- a) Visualisiere die Daten geeignet.
- b) Berechne den P -Wert mit dem Wilcoxon-U-Test und dem Welch-t-Test. Wie erklärst Du Dir den Ergebnis-Unterschied? Welche Sorte hat den höheren Proteingehalt, A oder B?
- c) Waren die Voraussetzungen für den t-Test überhaupt gegeben? Prüfe auf Normalverteilung in beiden Stichproben und prüfe im Anschluss auf Varianzhomogenität. Dürftest Du den Student's t-Test für zwei unabhängige Stichproben hier anwenden?

R-Frage 53 Ein Hotel gibt den Gästen einen Fragebogen zur Zufriedenheit zum Ausfüllen. Die Tabelle gibt die Antworten wieder, die im Mai letzten Jahres bezüglich der Qualität der Küche abgegeben wurden. Um die Qualität zu verbessern, wurde unterdessen ein neuer Küchenchef eingestellt. Die zweite Zeile gibt die Resultate vom Mai dieses Jahres wieder:

	sehr zufrieden	zufrieden	unzufrieden
letztes Jahr:	65	85	4
dieses Jahr:	85	65	4

- a) Siehst du in der Tabelle eine Verbesserung der Zufriedenheit?
- b) Ist die Verbesserung signifikant ($\alpha = 0.05$)?

R-Frage 54 Unten angegeben sind die Absatzzahlen eines neuen Gerätes in 12 Filialen einer Geschäftskette. Kann man auf dem 5%-Niveau sagen, dass der Median der verkauften Geräte grösser ist als 10 Geräte pro Filiale? Die Zahlen:

8 18 9 12 10 14 16 7 14 11 10 20

R-Frage 55 Eine Gruppe von Konsumenten vergleicht zwei neue Arten Cola. Die Konsumenten werden gebeten, den Geschmack auf einer Skala von 1 bis 30 anzugeben:

Tester	A	B	C	D	E	F	G	H	J	K	L	M	N	O
MiCola	20	24	28	24	20	29	19	27	20	30	18	28	26	24
Coopola	16	26	18	17	20	21	23	22	23	20	18	21	17	26

Gibt es einen Unterschied ($\gamma = 90\%$)?

CHECKLISTE

Weisst du jetzt:

- was der Sinn von parameterfreien Verfahren ist?

Kannst du jetzt:

- alle bisher behandelten Tests durchführen?
- die bisherigen Fragen dieses Abschnittes ohne Hilfe lösen?

7.3 Der χ^2 -Test

Ein recht vielfältiges Werkzeug ist der χ^2 -Test (χ ist der griechische Buchstaben Chi). Er vergleicht zwei oder mehr Häufigkeits-Verteilungen ganz allgemein in ihrer Form (sehen sich die Verteilungen ähnlich, oder haben sie signifikant verschiedene Form?), ohne sich auf einen einzelnen Parameter einzuschränken.

Voraussetzung des χ^2 -Tests ist, dass die Daten nominal- oder ordinalskaliert sind. Der χ^2 -Test involviert nicht einzelne Variablen-Werte, sondern geht von Häufigkeiten aus. Häufigkeiten für eine (möglichst nicht allzu grosse) Anzahl Klassen, wie sie bei nominalen oder ordinalen Daten gewöhnlich von selbst entstehen, bei metrischen durch das Zusammenfassen von Daten in Klassen erst künstlich gebildet werden müssen. Und noch etwas Vorteilhaftes: der χ^2 -Test kann auch mehr als zwei Verteilungen gleichzeitig vergleichen. Diese fast schon universellen Möglichkeiten eröffnen eine breite Palette an Anwendungsrichtungen. Der χ^2 -Test ist der Allrounder unter den statistischen Tests. Das ist aber nicht nur ein Vorteil. Spezialisierteren Tests ist der χ^2 -Test auf deren Spezialgebieten unterlegen, indem er weniger Power hat; das heisst, für ein beweiskräftiges Ergebnis sind beim χ^2 -Test Stichproben mit grösserem Umfang erforderlich als bei entsprechenden Spezial-Tests.

7.3.1 Vergleich zweier oder mehrerer Verteilungen in ihrer Form: χ^2 -Test, 2-oder mehr Stichproben-Fall

Wir wollen das Prinzip des χ^2 -Test an einem Beispiel erläutern: Tragen Frauen genauso oft einen Schirm bei sich wie Männer? Um diese Frage zu beantworten stellst Du Dich in die Bahnhofsstrasse und zählst wieviele Männer und Frauen einen Schirm dabei haben. Dein Ergebnis fasst Du in folgender Tabelle, auch Kontingenztafel genannt, zusammen.

	Frauen	Männer	Zeilensumme
mit Schirm	30	15	45
ohne Schirm	20	55	75
Spaltensumme	50	70	120

Die **Kontingenztafel** hat folgende Struktur.

	y_1	y_2	
x_1	B_{11}	B_{12}	$B_1 = \sum_j B_{1j}$
x_2	B_{21}	B_{22}	$B_2 = \sum_j B_{2j}$
	$B^1 = \sum_i B_{i1}$	$B^2 = \sum_i B_{i2}$	$n = \sum_{i,j} B_{ji}$

In der rechten Spalte stehen jeweils die Zeilensummen B_i , in der letzten Zeile stehen die Spaltensumme B^j . Das Feld unten rechts enthält die totale Anzahl Objekte. In unserem Beispiel hatten wir z. B. total $B_1 = 45$ Menschen mit Schirm (Zeilensumme Zeile 1) und total $B^2 = 70$ Männer (Spaltensumme Spalte 2) und total $n = 120$ Menschen beobachtet.

Die Kontingenztafel als solches ist nur eine Darstellungsform der Daten. Sie dient als Ausgangslage, um zum Beispiel die Frage zu beantworten, ob es einen signifikanten Unterschied gibt zwischen Frauen und Männern hinsichtlich ihrer Schirmgewohnheiten. Bei solchen Fragen stellt man jeweils die **Nullhypothese auf, dass die auftretenden Häufigkeiten gleichverteilt sind**. Was heisst das? Es gibt 45 von 120 Leuten mit Schirm und 50 von 120 Leuten waren Frauen. Rein theoretisch müssten also bei Gleichverteilung der Schirmgewohnheiten von Männern und Frauen $50/120 \cdot 45 = 18.75$ Frauen einen Schirm dabei haben. Wenn der tatsächlich beobachtete Wert stark von diesem Wert abweicht, dann muss man davon ausgehen, dass Frauen und Männer nicht gleichhäufig einen Schirm dabei haben. Diese Idee wird systematisch mit dem χ^2 -Test erfasst.

Rezept für χ^2 -Test**Frage:** Unterscheiden sich die Verteilungen zweier Stichproben signifikant?

1. Formuliere H_0 und H_1 . Die Nullhypothese lautet hier, dass sich die Verteilungen der beiden Stichproben nicht signifikant voneinander unterscheiden.
2. Lege α fest.
3. Jetzt berechnest du die Prüfgrösse

$$\chi_b^2 = \sum_{i,j} \frac{(B_{ij} - E_{ij})^2}{E_{ij}} \quad (7.1)$$

n ist der Umfang der Stichproben (nicht die Anzahl der Klassen).
 B_{ij} ist die beobachtete (absolute) Häufigkeit in Zeile i / Spalte j .
 B_i ist die Zeilenhäufigkeit und B^j die Spaltenhäufigkeit.

$$E_{ij} = \frac{B_i B^j}{n} \quad (7.2)$$

ist der erwartete Absolutwert.

Der erwartete Absolutwert ist symmetrisch im Zeilenindex i und Spaltenindex j , d.h. es ist egal welche Deiner Beobachtungen Du in den Zeilen bzw. Spalten anordnest.

Die Anzahl Freiheitsgrade ist

$$df = (c - 1) \cdot (r - 1) \quad (7.3)$$

c ist die Anzahl Spalten (columns) und r ist die Anzahl Zeilen (rows).

Die Prüfgrösse χ_b^2 (Gleichung 7.1) summiert also die auf den Erwartungswert normierten Abweichungsquadrate der beobachteten Häufigkeiten und der erwarteten Häufigkeiten. Je mehr die beobachteten Werte B_{ij} von den erwarteten Werten abweichen, um so grösser ist die Prüfgrösse χ_b^2 .

Der P-Wert ist die Fläche der χ^2 - Wahrscheinlichkeitsdichtfunktion (Abbildung 7.1) rechts vom berechneten χ_b^2 -Wert. Je grösser die Prüfgrösse χ_b^2 , umso kleiner ist der zugehörige P-Wert. In \mathbb{R} können wir P mittels

`pchisq(χ_b^2 ,df,lower.tail=FALSE)`

berechnen.

4. Falls $P < \alpha$, kann H_0 verworfen werden: der Unterschied zwischen den Verteilungen ist dann signifikant.

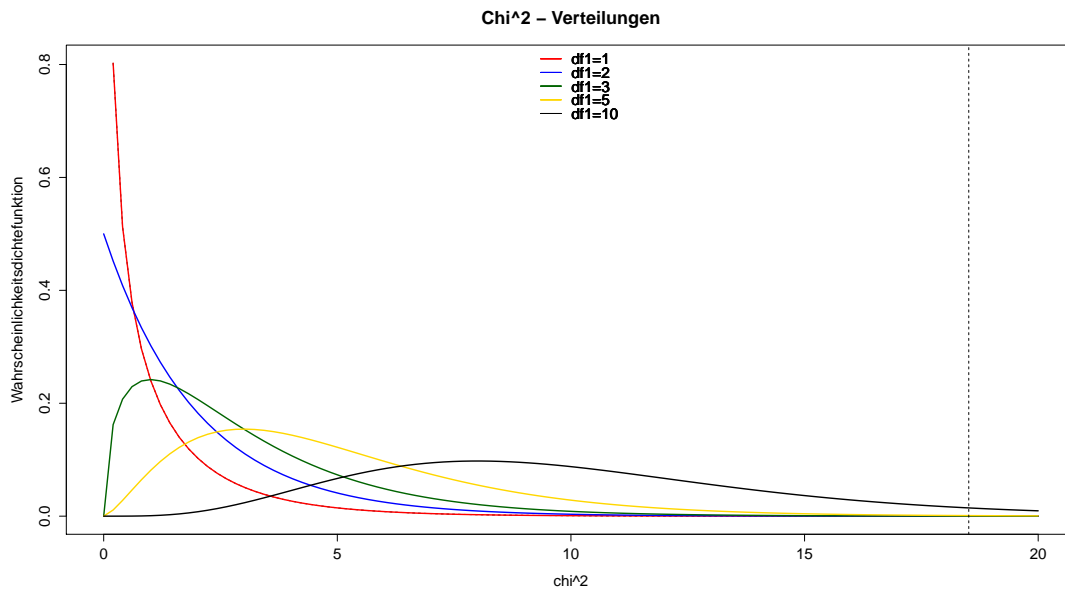


Abbildung 7.1: χ^2 -Verteilung für verschiedene Freiheitsgrade der Stichproben. Die Prüfgrösse χ^2 (aufgetragen auf der x-Achse) ist immer positiv. Gestrichelte Linie: $\chi_B^2 = 18.51$ des Schirmbeispiels. Rote Linie: $df = 1$ wie im Schirm-Beispiel.

Spielen wir nun den Test anhand des Schirmbeispiels durch:

1. H_0 : Frauen und Männer haben gleichhäufig den Schirm dabei.
 H_1 : Es gibt einen signifikanten Unterschied bei den Schirmgewohnheiten von Frauen und Männern.
2. $\alpha = 0.05$
3. Bevor wir die Prüfgrösse χ_B^2 berechnen können, erstellen wir die Kontingenztafeln für die beobachteten Häufigkeiten und für die erwarteten Daten.

Beobachtete Häufigkeiten:

	Frauen	Männer	Zeilensumme
mit Schirm	30	15	45
ohne Schirm	20	55	75
Spaltensumme	50	70	120

Erwartete Häufigkeiten:

	Frauen	Männer	Zeilensumme
mit Schirm	$\frac{45 \cdot 50}{120} = 18.75$	$\frac{45 \cdot 70}{120} = 26.25$	45
ohne Schirm	$\frac{75 \cdot 50}{120} = 31.25$	$\frac{75 \cdot 70}{120} = 43.75$	75
Spaltensumme	50	70	120

Damit haben wir:

$$\chi_b^2 = \frac{(30 - 18.75)^2}{18.75} + \frac{(15 - 26.25)^2}{26.25} + \frac{(20 - 31.25)^2}{31.25} + \frac{(55 - 43.75)^2}{43.75} = 18.51 \quad (7.4)$$

$$df = (2 - 1) \cdot (2 - 1) = 1.$$

`pchisq(18.51, 1, lower.tail=FALSE)`

ergibt einen P-Wert von $1.69 \cdot 10^{-5}$.

4. $P < \alpha$, damit wird H_0 verworfen. Es gibt also zwischen Frauen und Männer einen Unterschied bei der Häufigkeit der Schirmbenutzung.

7.3.1.1 χ^2 -Test in R

Wie so oft ist in **R** bereits ein fix-und fertiges Rezept zur Berechnung des P-Wertes beim Vergleich zweier oder mehrerer Verteilungen implementiert, das uns die mühsame Berechnung der Prüfgrösse χ_b^2 erspart. In **R** können wir somit den Schritt 3 des obigen Rezepts viel einfacher ausführen. Wieder das Schirmbeispiel:

```
3. mitschirm=c(30,15)
   ohneschirm=c(20,55)
   kontingenztabel=rbind(mitschirm,ohneschirm)
```

`rbind` ordnet die Vektoren `mitschirm` und `ohneschirm` zeilenweise (row) untereinander an ^a.

Der eigentliche Test lautet dann:

```
chisq.test(kontingenztabel,correct=FALSE)
```

Der **R**-Output lautet `X-squared = 18.5143, df = 1, p-value = 1.686e-05`. Das gleiche Resultat wie bei unsere Berechnung oben.

Ohne den Zusatz `correct=FALSE` macht **R** bei nur einem Freiheitsgrad (wie in unserem Beispiel) den χ^2 -Test mit der sogenannten Yates-Korrektur (siehe nächster Abschnitt 7.3.1.1.1) und liefert ein leicht anderes Resultat als die Handrechnung. Ist $df > 1$ macht es keinen Unterschied, ob `correct=FALSE` oder nicht gesetzt wurde, da die Yates-Korrektur nur bei $df = 1$ zum Tragen kommt. Verwende zukünftig bei allen Freiheitsgraden immer (den genaueren) Befehl:

```
chisq.test(kontingenztabel).
```

^aWir erhalten dasselbe Resultat, wenn wir die Daten spaltenweise (column) mit `cbind` verbinden: `kontingenztabel=cbind(mitschirm,ohneschirm)`, da die Berechnung der Prüfgrösse χ_b^2 symmetrisch in Spalten und Zeilen ist.

7.3.1.1.1 Voraussetzungen des χ^2 -Tests

- Die Daten sind nominal- oder ordinalskaliert.
- Der χ^2 -Test setzt eine gewisse Stichproben-Grösse voraus (deren Kriterien allerdings von Buch zu Buch variieren). Nehmen wir hier an, dass die Stichprobengrösse mindestens 50 sein soll.

- Mindestens 80% der erwarteten Zellhäufigkeiten der Kontingenztafel sind > 5 . Falls die Häufigkeiten zu niedrig sind, gibt `chisq.test` eine Warnmeldung aus, berechnet aber trotzdem einen P -Wert, der in diesem Fall mit Vorsicht zu geniessen ist. Sind die Häufigkeiten zu klein, sollte man die Werte zu breiter gefassten Klassen zusammenlegen oder einen Test wählen, der auch für kleine Stichproben geeignet ist wie den „exakten Test von Fisher“ (`fisher.test`) (wird hier nicht besprochen).
- Die Freiheitsgrade des Tests df sollten grösser als 1 sein. Ist dies nicht der Fall, führt `chisq.test` bei $df = 1$ automatisch die sogenannte Yates-Korrektur aus. In obigem Schirm-Beispiel haben wir mit dem Zusatz `correct=FALSE` diese Yates-Korrektur aus pädagogischen Gründen unterdrückt, um das Resultat der Handberechnung des „klassischen“ χ^2 -Test zu reproduzieren. Bei der Handrechnung wäre es einfach zu mühsam gewesen, die Erwartungswerte einer Kontingenztafel mit mehr als zwei Spalten zu berechnen. **Du solltest** jedoch im Allgemeinen **zur Berechnung des P -Wertes des χ^2 -Test** (Schritt 3) immer

3. `chisq.test(kontingenztafel)`

verwenden.

Ergänzungen:

- Der zwei- oder mehr-Stichproben χ^2 -Test kann auch dazu benutzt werden, die Abhängigkeit zweier Merkmale voneinander zu untersuchen. Wenn z.B. der Ausgangspunkt folgende Tabelle ist, in der die Spalten für verschiedene Augenfarben stehen und die Zeilen für verschiedene Haarfarben:

Haarfarbe / Augenfarbe	blau	braun	grün
<i>blond</i>	190	50	40
<i>braun</i>	40	220	120
<i>schwarz</i>	10	150	50,

dann bedeuten ungefähr gleiche Verteilungsformen in den Spalten, dass die Augenfarbe für alle Haarfarben gleich verteilt sind, d.h. Augen- und Haarfarbe sind voneinander unabhängige Variablen. Die Nullhypothese kann daher auch von der Form sein: Variable x ist unabhängig von Variable y .

- Wichtig ist, dass man *absolute* Häufigkeiten (absolute Anzahl Personen, Objekte ...) verwendet, keine Prozentangaben oder andere Verhältnisgrössen.
- So bequem und nützlich es ist, mehrere Verteilungen auf einen Schlag vergleichen zu können, so hat das auch eine Kehrseite, nämlich, dass man jetzt trotzdem nicht wirklich weiss, wo genau der Unterschied liegt (ist jede von jeder verschieden, oder fällt einfach eine aus dem Rahmen?). Man kann den Unterschied genauer lokalisieren, indem man doch wieder nur einen Teil der Daten testet, z.B. die erste gegen die zweite Zeile.

R-Frage 56 Zurück zu obiger Tabelle mit den Verteilungen der Augen- und Haarfarben. Kannst Du aufgrund dieser Erhebung behaupten, dass Augen- und Haarfarbe voneinander unabhängig sind?

R-Frage 57 In Wädwyla kandidieren 3 Parteien für den Gemeinderat: die „Muttertagspartei“ (MP), die „Grauen Elefanten“ (GE) und die „Fernseh-Fussball-Partei“ (FFP). Nach der Wahl macht das Institut „Demo-Vox“ eine Analyse, welche Bevölkerungsgruppen wie gewählt haben. Eine Stichprobe ergibt:

Bevölkerungs-Gruppe	MP	GE	FFP
Frauen unter 65	50	40	30
Männer unter 65	40	30	50
Menschen über 65	10	90	20

Haben die verschiedenen Bevölkerungsgruppen signifikant ($\gamma = 99\%$) verschieden gewählt?

Falls Du einen signifikanten Unterschied festgestellt hast, versuche ihn zu lokalisieren.

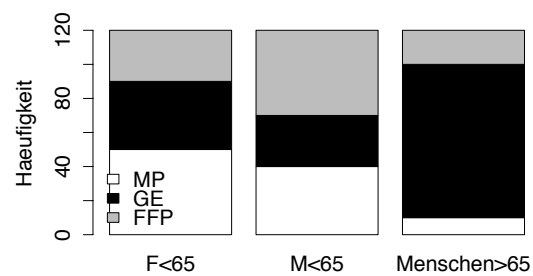


Abbildung 7.2: Wahlverhalten in Wädwyla

R-Frage 58 *Die Wirkung einer neuen Bewässerungs-Methode auf Auberginen wird untersucht. Von den 80 mit der neuen Methode bewässerten Pflanzen tragen 54 überdurchschnittlich viel. Kontrollgruppe mit normaler Bewässerung: hier tragen 34 von 60 Pflanzen überdurchschnittlich viel. Ist die Verbesserung durch die neue Methode signifikant ($\alpha = 5\%$)?*

a) *Beantworte die Frage zuerst mit dem χ^2 -Test.*

b) *Doch welcher Test erscheint Dir viel geeigneter? Warum? Führe den Test durch.*

Frage 27 *Zusatzfragen zur Auberginen-Aufgabe oben: Welchen Verbesserungsvorschlag machst du dem Auberginen-Forscher für sein nächstes Experiment?*

7.3.2 Vergleich mit einer vorgegebenen Verteilung: χ^2 -Test, 1-Stichproben-Fall

Auch wenn es nicht um einen einzelnen Parameter geht, sondern um eine ganze Verteilung, gibt es den Fall, dass eine der beiden Vergleichsseiten vorgegeben ist, durch Zielsetzungen, Theorien, bekanntes Wissen etc.

χ^2 -Test, 1-Stichproben-Fall

Test-Frage: *Entspricht die Verteilung einer Stichprobe einer vorgegebenen (erwarteten) Verteilung oder gibt es signifikante Unterschiede?*

Test-Rezept:

1. Formuliere H_0 und H_1 . 2. Lege α fest.
3. Berechnung des P-Wertes:
Zuerst die beobachtete Verteilung eingeben

`m=c(422,280,575)`

(Augenfarben blau/braun/grün von Studierenden).

Angenommen, wir wollen überprüfen, ob die Daten verträglich sind mit der Vorstellung dass die Augenfarben gleichverteilt sind. Dann formulieren wir eine Gleichverteilung mit ebenfalls 3 Elementen:

`erwarteteVerteilung=c(1,1,1)`

Und nun der Test:

`chisq.test(m,p=erwarteteVerteilung,rescale.p=TRUE)`

Alternative Formulierung

`chisq.test(m,p=c(1/3,1/3,1/3))`

4. Signifikanter Unterschied, wenn $P < \alpha$.

-
- Frage 28** *Ein Kühlschrankhersteller bietet drei Produktlinien an, die man als niedrig-, mittel- und hochpreisig bezeichnen kann. Die prozentualen Anteile an den Stückzahlen in den drei Produktkategorien waren jeweils 45%, 30% und 25%, bevor eine Werbekampagne die Vorteile der hochpreisigen Kühlschränke hervorhob. In einer Zufallsstichprobe von Kühlschränken, die nach der Kampagne verkauft wurden, waren die Verkaufszahlen in der niedrig-, mittel- und hochpreisigen Kategorie jeweils 15, 15 und 20 Stück. Teste die Nullhypothese auf dem 5%-Niveau, dass sich das jetzige Verkaufsmuster nicht von dem früheren unterscheidet.*
- a) Was genau ist hier die beobachtete Stichprobe?
 - b) Wo steht etwas über eine erwartete Verteilung?
-

- R-Frage 59** *Jetzt kannst du den χ^2 -Test durchführen.*
-

R-Frage 60 Laut Volkszählung setzt sich die Bevölkerung Kaliforniens aus folgenden ethnischen Gruppen zusammen: 50.7% Weisse, 6.6% Schwarze, 30.6% lateinamerikanischer Herkunft, 10.8% Asiaten, 1.3% andere. Die Zeitung „Investor’s Business Daily“ zählte bei 1000 College-Abgängern:

Weisse:	679
Schwarze:	51
Latinos:	77
Asiaten:	190
andere:	3

Was sind die Erwartungswerte, die man diesen Zahlen in einem χ^2 -Test gegenüberstellen würde? Führe sodann den χ^2 -Test durch.

R-Frage 61 Bei einem Würfelspiel hat jede Spielerin einen eigenen Würfel. Du hast dir deine 120 Würfe notiert:

1er	2er	3er	4er	5er	6er
23	23	21	26	21	6

„Ihr habt mir einen gezinkten Würfel gegeben, ich kriege praktisch nie einen 6er!“, folgerst du entnervt- Zurecht?

R-Frage 62 Das 2. Mendel’sche Gesetz sagt für die Kreuzung zweier Pflanzen mit rosa Blüten voraus, dass rote, weisse und rosa Nachkommen im Verhältnis 1:1:2 entstehen. Passen die folgenden 600 Nachkommen rosaroter Eltern zum 2. Mendel’sche Gesetz ($\gamma = 99\%$)?

rot	weiss	rosa
161	145	294

CHECKLISTE

Weisst du jetzt:

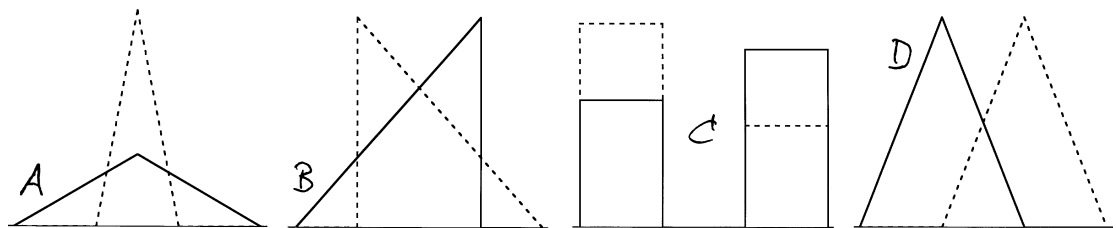
- wann es einen χ^2 -Test braucht?
- wie man beim χ^2 -Test auf die Erwartungswerte kommt

Kapitel 8

Rückblick auf die verschiedenen statistischen Tests

Für dich liegt das Problem jetzt vielleicht darin, die verschiedenen Test-Verfahren zu überblicken und zu sehen, welcher Test wann zum Einsatz kommt. Die folgende Frage könnte helfen, Dir einen Überblick zu verschaffen:

Frage 29 *In den vier Diagrammen sind jeweils die Häufigkeits-Verteilungen zweier Stichproben skizziert:*



- In welcher charakteristischen Eigenschaft unterscheiden sich die beiden Verteilungen jeweils?
 - Mit welchem/welchen Test(s) könnte man kontrollieren, ob der Unterschied zwischen den beiden Verteilungen signifikant ist?
 - Welche(r) Datentyp(en) wird/werden dabei vorausgesetzt?
 - Angenommen, es werden jeweils die Klimata zweier Länder verglichen, genauer gesagt die Häufigkeitsverteilungen der Tagestemperaturen. Mache dazu je ein Beispiel: Was könnte das untersuchte Merkmal sein, wie unterscheiden sich die beiden Länder konkret?
-

Frage 30 *Welche Verfahren hast du in diesem Kapitel bisher kennen gelernt, die*

- a) auf metrische Daten anwendbar sind?*
 - b) auf ordinale Daten anwendbar sind?*
 - c) auf nominale Daten anwendbar sind?*
 - d) auf dichotome Daten anwendbar sind?*
 - e) auf Anzahldichten anwendbar sind?*
 - f) gepaarte Stichproben anwendbar sind?*
 - g) auf mehr als 2 Stichproben anwendbar sind?*
 - h) Unterschiede in der Lage nachweisen?*
 - i) Unterschiede in der Streuung nachweisen?*
 - j) unterschiedliche Prozentsätze nachweisen?*
 - k) unterschiedliche Verteilungen nachweisen?*
-

Kapitel 9

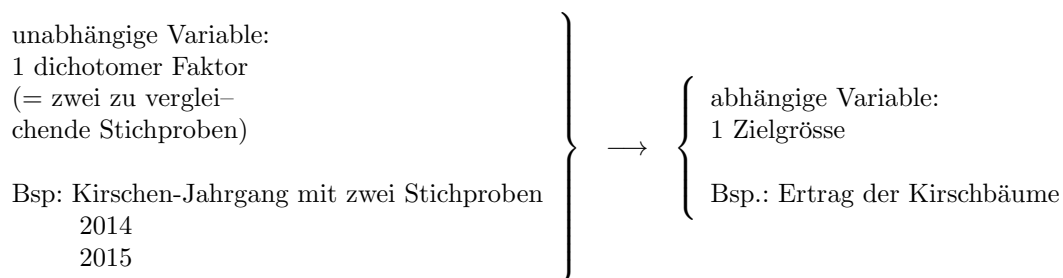
Überblick multivariate Methoden

Die bisher besprochenen statistischen Tests können (mit Ausnahme des χ^2 -Test) nur zwei Stichproben miteinander vergleichen, multivariaten Analysemethoden erlauben uns hingegen mehrere Stichproben miteinander zu vergleichen. Das vielleicht häufigst angewandte Verfahren dieser Art ist die „Varianz-Analyse“, die wir hier in ihrer einfachsten Form besprechen werden. Weitere Verfahren werden im Anschluss nur skizziert.

Bei den statistischen Tests ging es meist um *eine* einzige abhängige Variable (z.B. um den Ertrag von Kirschbäume in **R**-Aufgabe 43) in Abhängigkeit von einerer (dichotomen) unabhängigen Variablen (z. B. zwei verschiedenen Erntejahren), wir haben also den Ertrag der Kirschbäume *als Funktion* des Erntejahres untersucht.

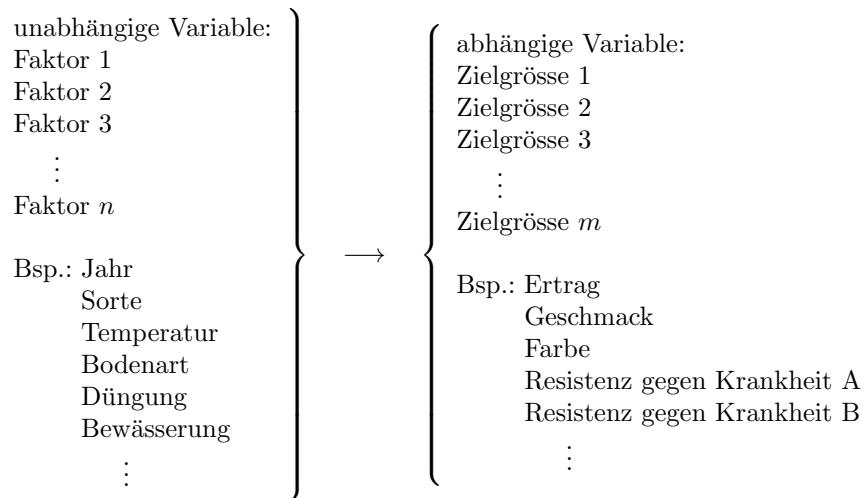
Vergleicht man mehr als zwei Stichproben, wird die unabhängige Variable auch als Faktor, die einzelnen Ausprägungen eines Faktors als (Faktor-)Stufen bezeichnet. Vergleicht man beispielsweise den Ertrag dreier Kirschjahrgänge, so wäre das Erntejahr der Faktor, die Erntejahre (z.B. 2014, 2015 und 2016) die entsprechenden Faktorstufen. Die abhängige Variable wird im Rahmen der multivariaten Statistik häufig auch als „erklärte“ Variable oder „Zielvariable/„Zielgrösse“ bezeichnet

In einfachen statistischen Tests geht es um eine einzige abhängige Variable, und eine einzige, nur dichotome unabhängige Variable (entspricht zwei Stichproben)



Multivariate Methoden dagegen untersuchen Situationen, bei denen mehrere unabhängige

und abhängige Variablen involviert sein können. Etwa bei einer Untersuchung von Gemüsesorten:



Je nach dem wie viele Faktoren und wie viele Zielgrößen welchen Datentyps involviert sind, kommen unterschiedliche multivariate Verfahren zur Anwendung. In dieser Vorlesung fokussieren wir uns auf Verfahren, die eine Zielgröße als Funktion eines oder zweier Faktoren untersuchen. Wie auch bei den „einfachen“ statistischen Tests unterscheiden wir wieder grundsätzlich zwischen parametrischen Tests, bei denen die Zielgröße metrisch sein muss und parameterfreien Tests, die diese Bedingung nicht stellen.

Abbildung 9.1 gibt einen Überblick über die Entscheidungskriterien beim Durchführen der in dieser Veranstaltung besprochenen Verfahren der Varianzanalysen. In den folgenden Abschnitten werden wir alle in diesem Flussdiagramm auftauchenden Tests erläutern.

9.1 Einfaktorielle Verfahren zum Vergleich von Mittelwerten und Medianen

Betrachtet man nur **eine Zielgröße als Funktion eines Faktors mit mehreren Faktorstufen**, vereinfacht sich obiges Schema zu



Tabelle 9.1 gibt einen Überblick aller in dieser Veranstaltung besprochenen statistischen-Verfahren, die es erlauben Mittelwerte bzw. Mediane einer **Zielgröße als Funktion eines Faktor mit zwei oder mehr Faktorstufen**) zu vergleichen.

Tabelle 9.1: Vergleich von Mittelwerten und Medianen einer Zielgrösse bei zwei oder mehr Stichproben

Datenskala abhängige Variable	Normal- verteilung?	Varianz- homogenität?	2 Stichproben	>2 Stichproben
mindestens metrisch	Ja	Ja	Student's t-test	einfaktorielle ANOVA (aov)
mindestens metrisch	Ja	-	Welch-t-Test	oneway.test
mindestens ordinal	-	-	unabhängig: Wilcoxon- Rangsummen-Test (U- Test), abhängig: Wilcoxon- Vorzeichen-Rang-Test	Kruskal-Wallis-Test

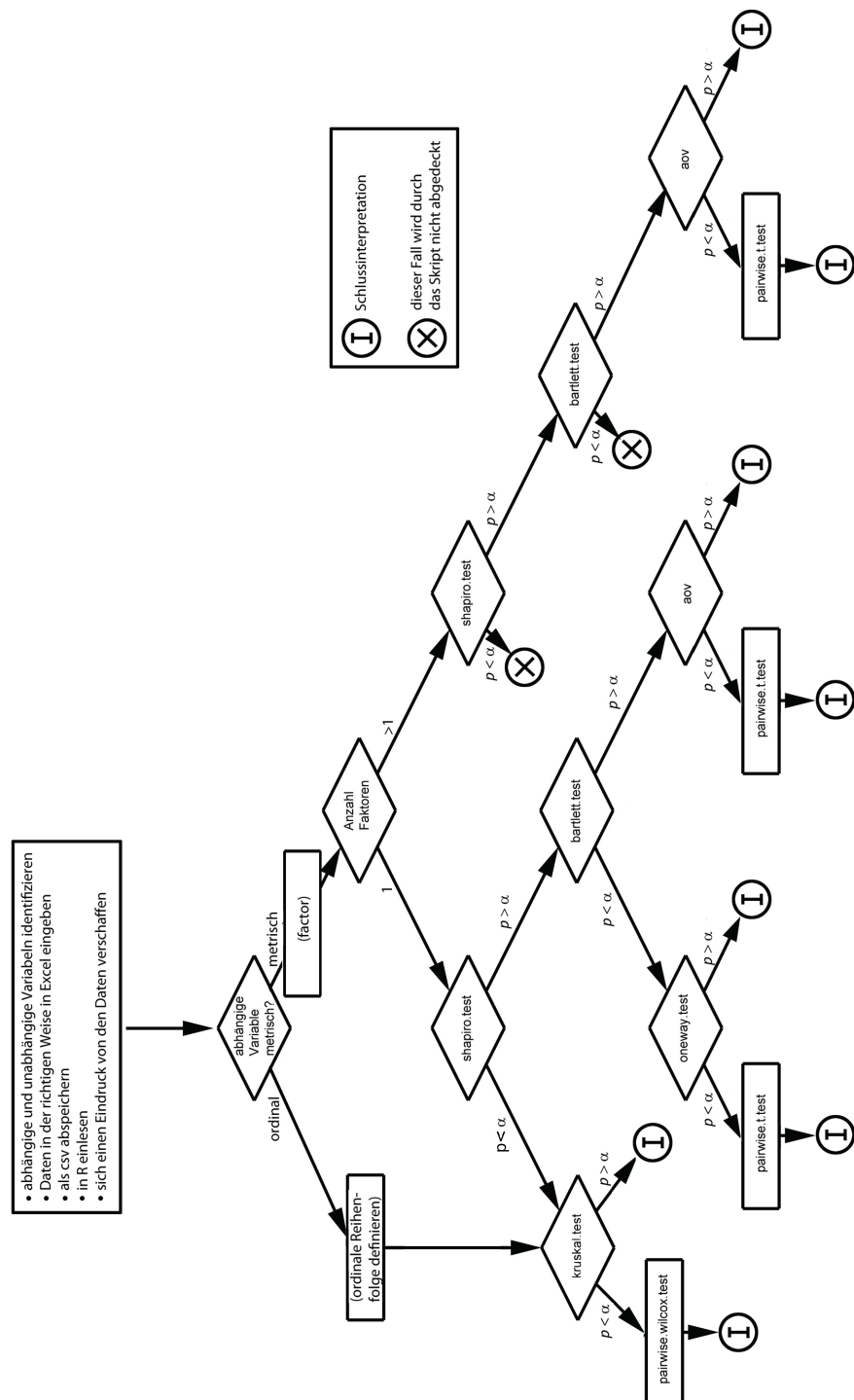


Abbildung 9.1: ANOVA-Flussdiagramm

Kapitel 10

Parametrische einfaktorielle Varianz–Analyse

10.1 Einfaktorielle ANOVA

Die einfaktorielle Varianzanalyse oder einfaktorielle ANOVA („analysis of variance“) ist die Verallgemeinerung des Student’s- t –Tests für mehr als zwei Stichproben. Die Bezeichnung „Varianz–Analyse“ rührt daher, dass die Varianzen der einzelnen Stichproben in die Berechnung der Prüfgrösse der ANOVA eingeht.

10.1.1 Voraussetzungen

Wie der Student’s- t –Test setzt die ANOVA voraus, dass

- die abhängige Variable metrisch ist, während an die unabhängige Variable(n) im Prinzip keine Anforderungen gestellt werden (am häufigsten ist/sind sie in der Praxis nominal),
- die abhängige Variable innerhalb jeder Stichprobe normalverteilt ist (bei mehr als 25 Messwiederholungen pro Stichprobe sind Verletzungen der Normalverteilungsvoraussetzung in der Regel unproblematisch, man sagt, die ANOVA sei robust gegen Abweichungen von der Normalverteilungsvoraussetzung),
- Varianzhomogenität in allen Stichproben (auch Homoskedazität genannt) gegeben ist.

10.1.2 Vorüberlegungen

Die Grundidee der ANOVA wollen wir am Beispiel der durchschnittliche Lebenserwartung von Männern in verschiedenen Ländern der Erde erläutern (Tabelle 10.1): Für jede der fünf betrachteten Weltregionen (Faktorstufen) ist die durchschnittliche Lebenserwartung

von Männern in 5-7 Länder (Anzahl der Wiederholungen pro Faktorstufe) angegeben. Diese ausgewählten Länder repräsentieren somit eine Region der Erde.

Europa	Pazifik/Asien	Afrika	Mittlerer Osten	Latein-Amerika
73	66	55	68	68
74	75	51	65	59
75	76	55	67	71
72	73	54	62	69
69	63	62	65	74
65		41	69	67
			65	57
$\bar{x}_1 = 71,3$	$\bar{x}_2 = 70,6$	$\bar{x}_3 = 53,0$	$\bar{x}_4 = 65,9$	$\bar{x}_5 = 66,4$
$s_1 = 3,72$	$s_2 = 5,77$	$s_3 = 6,90$	$s_4 = 2,34$	$s_5 = 6,21$
$n_1 = 6$	$n_2 = 5$	$n_3 = 6$	$n_4 = 7$	$n_5 = 7$

Tabelle 10.1: Lebenserwartung in den verschiedenen Weltregionen

Nun fragen wir uns, ob es zwischen den fünf verschiedenen Regionen Unterschiede in der Lebenserwartung gibt. Hier handelt es sich um den Vergleich mehrerer Stichproben, somit um das Einsatzgebiet der ANOVA. Ein Boxplot verschafft uns einen ersten Überblick über die Daten. In Abbildung 10.1 haben wir dem Boxplot zusätzlich einen sogenannten Jitter-Plot überlagert, der jeden einzelnen Messwert widerspiegelt. Natürlich kann und soll man

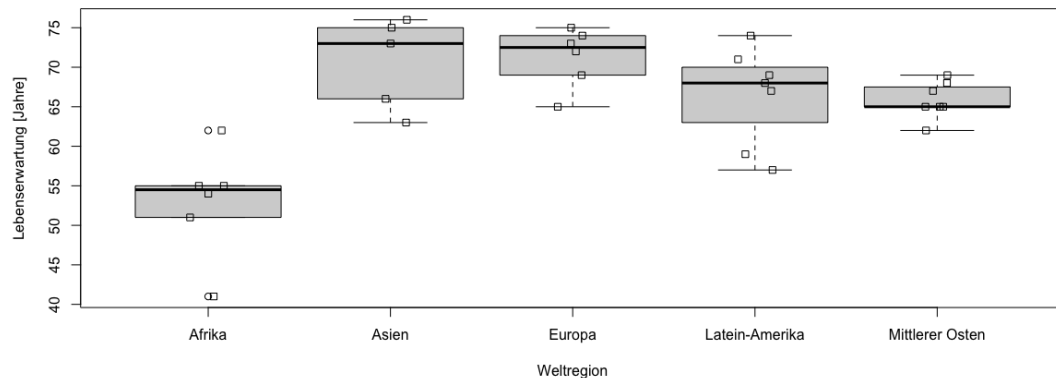


Abbildung 10.1: Boxplot: Lebenserwartung in verschiedenen Weltregionen

nun auf dieser Darstellung basierend einen Antwort-Versuch wagen – die Lebenserwartung in der ersten Stufe (Afrika) scheint signifikant tiefer zu sein als in den anderen Stufen. Die ANOVA erlaubt uns, diese Vermutung quantitativ anzugehen.

10.1.2.1 Das Problem multipler Paarvergleiche

Doch warum braucht es überhaupt eine neue Methode, um eine quantitative Antwort zu finden? Prinzipiell könnte man immer je zwei Stichproben herausnehmen (laut Kombinatorik gibt es bei 5 Regionen $\binom{5}{2} = 5 \cdot 4 / 2 = 10$ Möglichkeiten) und sie mit einem t -Test vergleichen. Das ergibt sich aber folgendes Problem: Bei einer Vielzahl von Stichproben kann es schon rein zufällig Extremwerte und somit auch signifikante Unterschiede zwischen zwei Stichproben geben. Bei jedem einzelnen t -Test zwischen zwei Stichproben ist die Wahrscheinlichkeit für einen Fehler 1. Art (wir lehnen die Nullhypothese ab, obwohl sie eigentlich gegolten hätte) maximal gleich der Irrtumswahrscheinlichkeit α . Die Wahrscheinlichkeit des Nicht-Eintretens eines Fehlers 1. Art ist dann mindestens $\gamma = 1 - \alpha$. Führt man nun N -Tests beträgt die Wahrscheinlichkeit keinen Fehler 1. Art zu machen mindestens γ^N . Dann ergibt sich die Wahrscheinlichkeit einen Fehler 1. Art zu machen beim N maligen Testen (auch "Familywise Error Rate") genannt zu

$$1 - (1 - \alpha)^N = 1 - \gamma^N$$

Führt man nun beispielsweise 10 t -Tests durch ergibt sich bei einer Irrtumswahrscheinlichkeit von $\alpha = 0.05$ eine "Familywise Error Rate" von $1 - 0.95^{10} = 40.2\%$! Extremwerte oder Ausreisser, wenn man sie sucht, denn es gibt sie von Natur aus, aus Gründen der Wahrscheinlichkeit. Aber wenn es sie aus Gründen der Wahrscheinlichkeit sowieso gibt, welche Aussagekraft haben sie dann noch? Es ist also nicht so klar, wie man es interpretieren soll, wenn wir jetzt z.B. eine Region mit kleinerer Lebenserwartung identifiziert haben. Die ANOVA wird uns helfen Signifikantes von Zufälligem zu unterscheiden.

10.1.3 Die Grundidee der ANOVA

Die Nullhypothese der ANOVA lautet, dass sich die Mittelwerte der einzelnen Stichproben nicht signifikant voneinander unterscheiden, die Alternativhypothese, dass mindestens ein Paar von Mittelwerten sich unterscheidet. Vergleiche die folgenden beiden Situationen, in der jeweils drei verschiedene Stichproben skizziert sind:



Abbildung 10.2: Verteilung von Stichproben. Links: Die Werte innerhalb der einzelnen Stichproben streuen stärker als zwischen den Stichproben. Kein signifikanter Unterschied zwischen den Stichproben ist zu erwarten. Rechts: Die Stichprobenmittelwerte streuen stärker als die Werte innerhalb der einzelnen Stichproben.

Es ist offensichtlich, dass sich die Stichproben in der rechten Situation signifikant voneinander unterscheiden. In der linken dagegen streuen die Werte innerhalb der einzelnen Faktorstufen

stärker als von Faktorstufe zu Faktorstufe – der Unterschied von Faktorstufe zu Faktorstufe könnte daher gut ein Zufallsprodukt sein, die Faktorstufen unterscheiden sich vielleicht nicht signifikant.

Die ANOVA quantifiziert diesen qualitativen Eindruck, indem das Verhältnis der Varianzen zwischen den Faktorstufen und den Varianzen innerhalb der Faktorstufen berechnet wird:

$$F = \frac{\text{Varianz zwischen den Faktorstufen}}{\text{Varianz innerhalb der Faktorstufen}}$$

Mit dieser Prüfgrösse F wird im Anschluss ein F -Test durchgeführt. In die Berechnung des finalen P -Wertes des F -Tests gehen neben der Prüfgrösse F auch die Freiheitsgrade der einfaktoriellen Varianzanalyse ein. Sie hängen sowohl von der Gesamtzahl aller Messungen (Wiederholungen) N als auch von der Anzahl der Faktorstufen k ab:

$$df_1 = k - 1, df_2 = N - k$$

df_1 sind die Freiheitsgrade der Varianz zwischen den Faktorstufen, df_2 die Freiheitsgrade für die Varianz innerhalb der Faktorstufen. Beim F -Test (siehe Abschnitt 6.3) gilt: Je grösser der F -Wert, um so kleiner ist die Wahrscheinlichkeit der Nullhypothese P . Ist die Varianz zwischen den Faktorstufen somit viel grösser als die Varianz innerhalb der Faktorstufen, so ist ein statistisch signifikanter Unterschied wahrscheinlich.

Die Berechnung der Prüfgrösse F ist komplizierter als alle vorangehenden Rezepte. Wegen des grossen Rechenaufwands nimmt man deshalb in der Praxis (und unserer Vorlesung) für eine ANOVA immer den Computer zu Hilfe nehmen. Neugierige können in Abschnitt B (freiwillig) nachlesen, welche Schritte zur Durchführung der ANOVA notwendig sind und von R im Hintergrund ausgeführt werden.

10.1.4 Einfaktorielle Varianzanalyse mit R

10.1.4.1 Datenvorbereitung

10.1.4.1.1 In Excel

In EXCEL hast du eine gewisse Freiheit, wie du die Daten genau anordnest. Statistiker verwenden dagegen eine standardisierte Anordnung, die ich dir auch empfehle. Der Import und die Weiterbearbeitung der Daten in einem echten Statistik-Programm gelingt damit viel problemloser.

Der Standard sieht so aus:

- Jede abhängige und unabhängige Variable belegen je eine Spalte. Die Reihenfolge der Spaltenanordnung ist nicht von Bedeutung. Es empfiehlt sich, in der ersten Zeile den Namen der Variablen einzugeben.
- Jede Zeile stellt einen „Fall“ dar, d.h. ein Set von zusammengehörigen Variablen-Werten. Ein Fall entspricht oft einer Person oder einem Zeitpunkt oder einer Wiederholung eines Experimentes.

Im Beispiel unserer Lebenserwartungen sieht eine geeignete Anordnung der Daten wie in Abbildung 10.3 aus.

Region	Lebenserwartung
Afrika	41
Afrika	51
Afrika	54
Afrika	55
Afrika	55
Afrika	62
Asien	63
Asien	66
Asien	73
Asien	75
Asien	76
Europa	65

Abbildung 10.3: Ausschnitt des Excel-Sheets leberw.xlsx. Unabhängige und abhängige Variable sind je in einer Spalte angeordnet.

Im EXCEL-File leberw.xlsx sind die Daten schon entsprechend angeordnet.

Diese Anordnung der Daten ist nicht nur bei der ANOVA nützlich, sondern auch für Tests wie den χ^2 -Test:

Frage 31 Welche der folgenden drei Darstellungen ist/sind somit empfohlen?

a)

	A	B	C	D	E	F
1	Person	1	2	3	4	5
2	Augenfarbe	blau	braun	braun	blau	braun
3	Haarfarbe	blond	braun	braun	braun	braun
4						

b)

	A	B	C	D
1			Augenfarbe	
2			blau	braun
3	Haarfarbe	blond	1	0
4		braun	1	3

c)

	A	B	C	D
	Person	Augen-	Haar-	
		farbe	farbe	
1				
2	1	blau	blond	
3	2	braun	braun	
4	3	braun	braun	
5	4	blau	braun	
6	5	braun	braun	
7				

10.1.4.1.2 In R

Sind nur wenige Daten vorhanden, kannst Du auch direkt in **R** einen `data.frame` erstellen. Im Beispiel der Lebenserwartungen erstellst Du zuerst einen Datenvektor für die Weltregionen:

```
eur=rep('Europa',6); paz=rep('Pazifik/Asien',5); afr=rep('Afrika',6)
mo=rep('Mittlerer Osten',7); la=rep('Latein-Amerika',7)
Region=c(eur,paz,afr,mo,la)
```

und einen für die zugehörigen Lebenserwartungen:

```
leur=c(73,74,75,72,69,65); lpaz=c(66,75,76,73,63); lafr=c(55,51,55,54,62,41);
lmo=c(68,65,67,62,65,69,65) lla=c(68,59,71,69,74,67,57)
Lebenserwartung=c(leur,lpaz,lafr,lmo,lla)
```

und fasst die beiden Vektoren sodann in einen `data.frame` zusammen:

```
leberw=data.frame(Region,Lebenserwartung)
```

Wieder entspricht jede Zeile des `data.frame` einer Wiederholung oder Messung.

10.1.4.2 Testrezept der einfaktoriellen ANOVA

Einfaktorielle Varianz-Analyse

Test-Frage: *Unterscheiden sich die Mittelwerte der Faktorstufen signifikant voneinander?*

Test-Rezept:

1. Formuliere H_0 : Die Mittelwerte der einzelnen Faktorstufen unterscheiden sich nicht signifikant voneinander unterscheiden.
 H_1 : Mindestens ein Paar von Mittelwerten der verschiedenen Faktorstufen unterscheidet sich signifikant.
2. Lege α fest.
 Lies die Daten in ein `data.frame` mit abhängiger und unabhängiger Variable je als Spalte. Im Beispiel nennen wir das `data.frame` `leberw`.
 Überprüfe die Voraussetzung der ANOVA:
 - (a) Normalverteilung in den einzelnen Stichproben (Test in der Regel nur bei Stichproben mit weniger als 25 Wiederholungen pro Faktorstufe notwendig).
 Nullhypothese H_{0N} : Die Daten jeder einzelnen Stichprobe sind normalverteilt. In `R` lässt sich der Shapiro-Wilk-Test auf Normalverteilung in den einzelnen Stichproben folgendermassen ausführen:

```
shapiro.test(rstandard(aov(leberw$Lebenserwartung~leberw$Region)))
```

 Ist $P > \alpha$, verwerfen wir die Nullhypothese nicht.
 - (b) Varianzhomogenität über alle Stichproben („Homoskedastizität“).
 Nullhypothese H_{0V} : Homoskedastizität gegeben.

```
bartlett.test(leberw$Lebenserwartung~leberw$Region)
```

 Ist $P > \alpha$, verwerfen wir die Nullhypothese nicht.
3. Nun kommt die eigentliche Varianzanalyse-Befehl `aov` zum Zug. Hier :

```
summary(aov(leberw$Lebenserwartung~leberw$Region))
```

 Wie bei den statistischen Tests ist am Output für uns einfach der P -Wert ausschlaggebend. Es ist die Zahl, die unter dem Spaltentitel `Pr(>F)` steht.
4. Ist $P < \alpha$, verwerfe wir die Nullhypothese.
 Im vorliegenden Fall ist $P = 1.36e-05$. Diese Zahl ist kleiner als jedes vernünftige α , es sind damit signifikante Unterschiede nachgewiesen.

Wir wollen nun die ANOVA an weiteren Beispiel erläutern:

Frage 32 Die folgende Tabelle enthält Messungen des Nährwertes [in kJ/100 g] von Käse aus verschiedenen Käseereien (jede Zahl entspricht einem Produkt):

Gruyère	Tilsiter	Edamer	Appenzeller
1780	1120	1080	1630
1690	1180	990	1720
1670	1110	1050	1610
1710	1070	1100	
	1020	1110	

Was ist jetzt was im Beispiel der Tabelle oben?

- Was ist hier die abhängige Variable (=Zielgrösse), was ist hier die unabhängige Variable (=Faktor)?
- Was ist hier z.B. eine Stichprobe?
- Warum ist dies eine Stichprobe und nicht eine Vollerhebung?
- Was ist ein Element einer Stichprobe (=1 Wiederholung)?

Vorab ist nicht klar, wie viele und welche der Stichproben etwas Interessantes an den Tag bringen könnten. Gibt es z.B. einfach eine der vier Sorten, die einen signifikant höheren Nährwert hat als alle anderen, während diese anderen ungefähr übereinstimmen, oder unterscheidet sich jede Stufe von jeder?

Frage 33 Lässt sich diese Frage klären, indem wir den χ^2 -Test anwenden?

R-Frage 63 Führe die ANOVA für die Käsesorten in **R** durch.

- Ordne die Daten in einem `data.frame` so an, das Du anschliessend eine ANOVA ausführen zu kannst. Vergleiche die Struktur Deines `data.frame` mit `Kaese.xlsx` auf Moodle.
 - Analysiere in einem ersten Schritt die Daten graphisch mit einem Box-Plot.
 - Führe die ANOVA durch. Prüfe vorab, ob die Voraussetzungen zur Durchführung gegeben sind.
-
-

R-Frage 64 Die Tabelle unten gibt die Erträge dreier Sorten einer Nutzpflanze wieder, die in jeweils drei, bzw. vier Versuchen ermittelt wurden:

Sorte A	Sorte B	Sorte C
2.4	1.5	1.5
2.8	1.9	2.2
2.3	1.7	1.8
	1.7	

Was sagt die ANOVA dazu? ($\alpha = 5\%$.)

10.1.5 Post-hoc-Test

Falls die ANOVA einen signifikanten Unterschied bei den Mittelwerten ergibt, stellt sich natürlich die Frage, welche Faktorstufen sich signifikant unterscheiden. Normalerweise hat man dafür eine Vermutung (etwa Afrika und Europa im Lebenserwartungsbeispiel). Wenn es nicht augenfällig ist, findet man es heraus, indem man die Faktor-Stufen nach einem speziellen Plan paarweise vergleicht. Man spricht von „Post-hoc-Tests“.

Wichtig: Ob es *überhaupt* Unterschiede gibt, sagt dir die Varianz-Analyse. Erst wenn Unterschiede bewiesen sind, darfst du ein post-hoc-Verfahren anwenden. Es kann nämlich sein, dass ein post-hoc-Verfahren auch Unterschiede anzeigt, wenn das Gesamtbild der ANOVA keine solchen bestätigt.

Es gibt zahlreiche post-hoc-Verfahren; alle berechnen alle kombinatorisch möglichen paarweisen Vergleiche zwischen den Faktorstufen (englisch: group levels). Um dem Problem der erhöhten Wahrscheinlichkeit eines Fehlers 1. Art bei multiplen Tests entgegenzutreten (siehe Abschnitt 10.1.2), werden unterschiedliche Korrekturverfahren verwandt. Hier möchte ich nur die relativ häufig benutzte „Bonferroni-Holm-Korrektur“ vorstellen:

Post-hoc-Test zur einfaktoriellen Varianz-Analyse

Test-Frage: *Falls und nur falls(!) eine Varianz-Analyse signifikante Unterschiede nachgewiesen hat, zwischen welchen Faktorstufen liegen diese signifikanten Unterschiede dann vor?*

Test-Rezept: α von der vorhergehenden Varianz-Analyse übernehmen. Angenommen,
`pairwise.t.test(leberw$Lebenserwartung, leberw$Region, p.adjust.method="holm")`

R gibt nun eine Tabelle aus, in der die 5 Weltregionen einander gegenüber gestellt werden. Der Vergleich Latein-Amerika versus Pazifik/Asien ergibt beispielsweise einen P -Wert von 0.5491. Dieser Wert liegt deutlich über α , die Lebenserwartungen in Latein-Amerika und der Region Pazifik/Asien unterscheiden sich somit nicht signifikant. Signifikante Unterschiede liegen nur zwischen Afrika und allen anderen Weltregionen vor. Wie vermutet.

R-Frage 65 *Führe die post-hoc-Analyse im Käsebeispiel aus.*

R-Frage 66 *Zwischen welchen Sorten der oben analysierten Nutzpflanze liegen signifikante Unterschiede im Ertrag vor?*

10.1.6 Varianzanalyse in R: Die unabhängigen Variablen müssen vom Datentyp `factor` sein

Wie weit voneinander entfernt sollte man Kohlköpfe pflanzen, damit die Köpfe möglichst gross werden? Du probierst es einmal mit verschiedenen Abständen aus und misst folgende Kohlkopfgewichte:

		Abstand			
		54 cm	46 cm	38 cm	30 cm
Feld:	A	1106 g	992 g	879 g	879 g
	B	1021 g	951 g	794 g	652 g
	C	936 g	1011 g	964 g	851 g

(Die Felder sind hier kein eigentlicher Faktor, sondern zeigen nur die unabhängigen Wiederholungen innerhalb jeder Faktorstufe des Abstandes an.) Nun möchtest Du mit einer einfaktoriellen Varianzanalyse herausfinden, ob das mittlere Gewicht der Kohlköpfe abhängig vom Pflanzabstand ist. Die Daten findest Du auf Moodle unter [kohleinfaktoriell.xlsx](#) schon so angeordnet, dass Du direkt eine Varianzanalyse in **R** ausführen könntest. Nachdem Du die Daten in ein csv-File verwandelt hast, lies die Daten in ein `data.frame` namens `Kohl` in **R** ein. **R** entscheidet sich nun beim Einlesen einer Datei automatisch für einen passenden Datentyp. Eine Variable deren Ausprägungen durchgehend Zahlen sind (wie hier die Variable „Abstand“) wird als `integer` oder `numeric` klassiert. Dieser Abstand ist jedoch die unabhängige Variable (der Faktor) der angepeilten ANOVA. Obwohl die Faktoren einer ANOVA typischerweise nominal sind, ist ein anderer Datentyp (wie der – metrische – Abstand der Kohlköpfe) prinzipiell nicht ausgeschlossen. Sind jedoch abhängige **und** unabhängige Variable metrisch (wie Gewicht und Abstand im Kohlbeispiel), so könnten wir prinzipiell auch eine Regressionsanalyse machen (siehe Abbildung 11.1). Und jetzt kommt die Tücke: Die Varianz-Analyse-Funktion `avov` entscheidet sich – mit viel Eigeninitiative und ohne dir eine Silbe von ihrem Entschluss mitzuteilen – dafür, mit metrischen Fakto-

ren eine Regressions- und nicht eine Varianz-Analyse zu machen. Auch wenn diese beiden Verfahren grundlegend verwandt sind, sind die P -Werte nicht identisch. Willst Du **R** also dazu zwingen, eine Varianz-Analyse zu machen, so musst Du die unabhängige Variablen vom Datentyp **integer** oder **numeric** mit dem Befehl **as.factor** in den Datentyp factor verwandeln. Im Kohl-Beispiel:

```
Kohl$Abstand=as.factor(Kohl$Abstand)
```

Diese Umwandlung des Datentyps musst Du bei allen vorgestellten Varianten der Varianz-Analyse machen, so einer der unabhängigen Faktoren vom Datentyp **integer** oder **numeric** sein sollte.

-
- R-Frage 67** a) *Führe eine Varianzanalyse mit den Kohldaten in **kohleinfaktoriell.xlsx** aus. Verschaffe Dir zuerst mit einer Graphik einen Überblick und prüfe die Voraussetzungen. Falls die Varianz-Analyse einen signifikanten Unterschied zwischen mindestens einer der Faktorstufenkombinationen findet, führe einen posthoc-Test durch.*
- b) *Was passiert, wenn Du die unabhängige Variable nicht in einen **factor** verwandest? Probier es aus und vergleiche Dein Resultat mit dem Resultat einer linearen Regression.*
-

10.2 oneway.test

Der **oneway.test** ist die Verallgemeinerung des Welch- t -Tests (Abschnitt 6.1.3.1.2) auf mehr als zwei Stichproben; er setzt somit eine metrische abhängige Variable und Normalverteilung in allen Faktorstufen voraus, verzichtet aber auf die Bedingung der Varianzhomogenität in allen Stichproben. Das Test-Prozedere läuft ansonsten analog zur ANOVA ab. Insbesondere lautet die Nullhypothese wieder, dass sich die Mittelwerte der Faktorstufen nicht signifikant unterscheiden. Der Welch-**oneway.test** im Lebenserwartungsbeispiel lautet dann

```
oneway.test(leberw$Lebenserwartung~leberw$Region)
```

Wieder interessiert uns im Output nur der P -Wert **p-value = 0.002837**.

-
- Frage 34** *Was bedeutet der p -Wert 0.002837?*
-

CHECKLISTE

Kannst du jetzt:

- *die Daten Deiner Stichproben so anordnen, dass Du eine ANOVA mit R einfach durchführen kannst?*
- *die Voraussetzungen einer einfaktoriellen ANOVA mit R überprüfen?*
- *den P-Wert im R-Output der ANOVA finden und interpretieren?*
- *eine post-hoc-Analyse durchführen und interpretieren?*

Kapitel 11

Parameterfreie einfaktorielle Varianz–Analyse

11.1 Kruskal–Wallis–Test (H–Test) bei metrischen Daten

Der Kruskal–Wallis–Test (auch H–Test genannt) ist die Verallgemeinerung des Wilcoxon–Tests für mehr als zwei Stichproben. Die abhängige Variable muss somit mindestens ordinal sein, Normalverteilung oder Varianzhomogenität in allen Faktorstufen muss jedoch nicht gegeben sein. Wie der Wilcoxon–Test reduziert der H –Test die Daten auf korrigierte Ränge (daher auch der Name „Rang–Varianz–Analyse“) und tested auf den Unterschied in der Lage des Medians von drei oder mehr ungepaarten Stichproben. Auch dies kann man mühelos von **R** erledigen lassen. Im Lebenserwartungsbeispiel:

```
kruskal.test(leberw$Lebenserwartung~leberw$Region)
```

R–Frage 68 *Vergleiche die P –Werte von*
`aov(leberw$Lebenserwartung~leberw$Region),`
`oneway.test(leberw$Lebenserwartung~leberw$Region)` *und*
`kruskal.test(leberw$Lebenserwartung~leberw$Region)`
miteinander. Was fällt Dir auf?

Das post–hoc–Verfahren sollte konsequenterweise auch parameterfrei sein. Damit geht es:

```
pairwise.wilcox.test(leberw$Lebenserwartung,leberw$Region,p.adjust.method="holm")
```

R–Frage 69 *Vergleiche das Resultat mit den parametrischen post–hoc–Werten.*

-
- R-Frage 70** *Du hast 6 verschiedene Insektizide auf Versuchsfeldern versprüht und im Anschluss die Anzahl der Insekten pro Feld gezählt. Den zugehörigen Datensatz kannst Du einfach mit `data(InsectSprays)` einlesen.*
- a) Inspiziere die Daten mit den Befehlen `str(InsectSprays)` und `View(InsectSprays)`.*
 - b) Plote die Daten mit `boxplot`. Was fällt Dir auf?*
 - c) Teste nun die Voraussetzungen der ANOVA.*
 - d) Mit welchem Test kannst Du nun signifikante Unterschiede in der Anzahl der gezählten Insekten nachweisen?*
 - e) Führe einen geeigneten post-hoc-Test durch.*
-

11.2 Kruskal–Wallis–Test (H–Test) bei ordinalen Daten

Im Lebenserwartungsbeispiel haben wir einen metrischen Datensatz parameterfrei analysiert, obwohl die Voraussetzungen für einen parametrischen Test gegeben waren. Das parameterfreie Verfahren war somit nicht zwingend (und auch nicht das geeigneteste Verfahren aufgrund der geringeren Teststärke). Im Falle nicht normalverteilter Daten (wie im Insektizidbeispiel [R-Frage 70](#)) oder bei ordinalen Messwerten kommt man um ein parameterfreies Verfahren wie den Kruskal–Wallis-Test nicht herum:

R-Frage 71 3 Lebensmittelgeschäfte werden verglichen in Bezug auf die Frische der angebotenen Gemüse und Früchte. Eine Inspektorin besucht jedes der Geschäfte und bewertet jeweils 9 zufällig ausgewählte Produkte auf ihre Frische hin. Die Bewertungen:

Geschäft A	Geschäft B	Geschäft C
sehr frisch	nicht frisch	nicht frisch
frisch	frisch	frisch
sehr frisch	frisch	frisch
sehr frisch	frisch	nicht frisch
frisch	sehr frisch	nicht frisch
frisch	sehr frisch	verdorben
nicht frisch	frisch	frisch
frisch	sehr frisch	nicht frisch
sehr frisch	frisch	nicht frisch

Im Excel-File [lebensmittelzahlen.xlsx](#) sind die Bewertungen der Inspektorin folgendermassen in Zahlen übersetzt:
 „sehr frisch“ – > 4, „frisch“ – > 3, „nicht frisch“ – > 2, „verdorben“ – > 1.
 Führe nun den Kruskal–Wallis-Test und eventuell ein geeignetes post-hoc-Verfahren aus.

In **R-Frage 71** haben wir die Bewertungen gemäss einer von uns definierten Reihenfolge in Zahlen übersetzt. Dieser Schritt ist immer notwendig, so die ordinalen Daten der abhängigen Variablen (die Messwerte) in „Wortform“ vorliegen. **Alle Varianten der Varianzanalyse in R setzen voraus, dass die abhängigen Variable vom Typ `int` oder `num` ist.**

Abbildung 11.1 gibt einen Überblick über alle bisher vorgestellten Testverfahren mindestens dreier unabhängiger Stichproben.

CHECKLISTE

Kannst du jetzt die Voraussetzungen für

- eine ANOVA,
- den `oneway.test`,
- den Kruskal–Wallis-Test benennen?

Kannst du jetzt die zugehörigen Alternativ- und Nullhypothesen formulieren?

Kannst du jetzt

- diese Tests in **R** ausführen?

Kannst du den zugehörigen **R** -Output interpretieren?

Kannst du das für den jeweiligen Test

- geeignete post-hoc-Verfahren auswählen,
- ausführen und
- interpretieren?

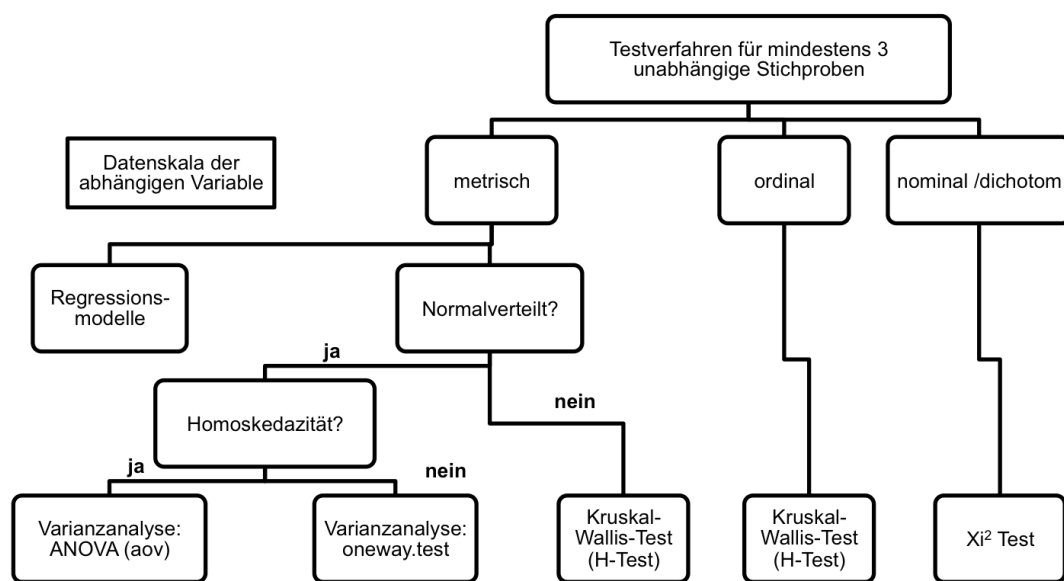


Abbildung 11.1: Entscheidungsdiagramm zu Mehrstichprobenverfahren (unabhängige Stichproben)

Kapitel 12

Zweifaktorielle Varianzanalyse

Meist interessiert nicht nur ein einziger Einflussfaktor. Die Grösse der Kohlköpfe wird auch noch von der Bewässerung, der Düngung, der Temperatur, der Kohlsorte usw. abhängen.

Die Stärke der multivariablen Methoden

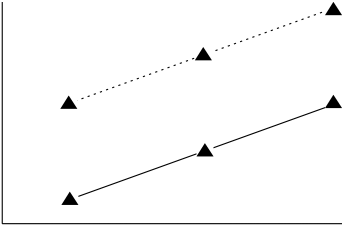
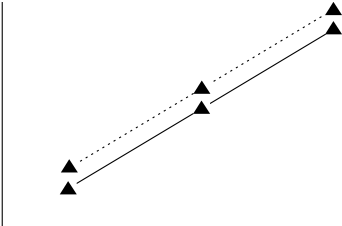
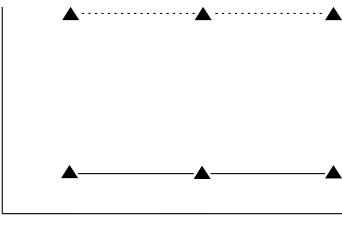
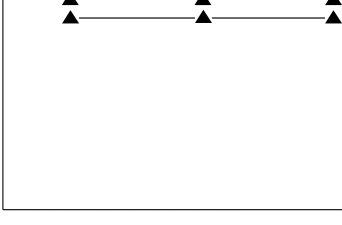
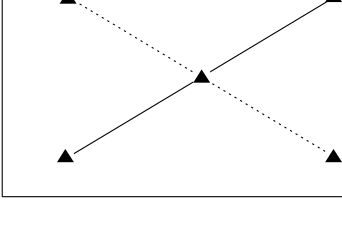
Grundsätzlich könnte man jeden dieser Faktoren in einem eigenen Versuch optimieren. Die Varianz-Analyse bietet jedoch auch die Möglichkeit, zwei oder mehr Faktoren gleichzeitig zu erforschen, wobei man in der Praxis ungern über zwei Faktoren hinausgeht, weil das Verfahren ab dem dritten Faktor an Übersichtlichkeit verliert. Je ein eigener Versuch für jeden der Faktoren oder zwei Faktoren gleichzeitig – was ist vorzuziehen? Die zweifaktorielle ANOVA bietet gegenüber zwei Einzelbehandlungen zwei entscheidende Vorteile:

1. Effizienz: Jeder Kohlkopf wird sozusagen doppelt genutzt, für die Beurteilung zweier Faktoren gleichzeitig. Bei gleicher Aussagenschärfe/-sicherheit ist der Gesamtumfang (und damit der Versuchsaufwand) bei der zweifaktoriellen ANOVA deutlich kleiner als bei zwei einfaktoriellen ANOVA.
2. Wechselwirkung: Einfaktoriell kann man die Möglichkeit einer Wechselwirkung der unabhängigen Faktoren nicht untersuchen und macht damit potentiell einen Fehler.

12.1 Wechselwirkung der Faktoren

Kann man zwei Faktoren (sagen wir: Sorte und Düngung) wirklich einzeln, voneinander isoliert betrachten, sind die beiden Faktoren voneinander unabhängig? Das ist keineswegs selbstverständlich. Es kann ja sein, dass ein Dünger bei Sorte A ganz anders wirkt, als bei Sorte B; Sorte A gedeiht vielleicht am besten mit wenig Dünger, Sorte B mit viel. Die Sorte beeinflusst dann gewissermassen die Reaktion auf die Düngungsstufen. Man spricht hier von einer „Wechselwirkung“ zwischen den beiden Faktoren.

Frage 35 *In den nachfolgenden Diagrammen sind jeweils nach oben die Resultate eines zweifaktoriellen Experimentes aufgetragen (Ertrag einer Nutzpflanze). Die Dreiecke stellen die Stufenmittelwerte (von Faktorstufen-Kombinationen) dar, einer der beiden Faktoren (Düngungsmenge) ist nach rechts aufgetragen, die Stufen des anderen Faktors (Sorte) bestehen in den beiden Linien (Sorte A ausgezogene Linie, Sorte B gestrichelt). Was ist vermutlich signifikant? Und wo besteht eine Wechselwirkung? Generell sind in einem solchen Diagramm die Linien immer dann mehr oder weniger parallel, wenn keine Wechselwirkung besteht, und umgekehrt.*

	Abhängigkeit von der Düngung	Abhängigkeit von der Sorte	Wechselwirkung Sorte \longleftrightarrow Düngung
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Das Beispiel mit den Kohlköpfen wird nun um einen zweiten Faktor ergänzt:

		Abstand [cm]			
		54	46	38	30
Bewässerung häufig	Feld A	1106	992	879	879
	Feld B	1021	951	794	652
	Feld C	936	1011	964	851
Bewässerung selten	Feld D	964	794	567	595
	Feld E	851	907	709	680
	Feld F	907	680	765	510

R–Frage 72 Die Kohl-Daten sind bereits digitalisiert in [kohlzweifaktoriell.xlsx](#). Lies sie in **R** ein und gib dem eingelesenen Datensatz den Namen `df`. Dann `attach(df)`. Inspiziere die Daten, indem du dir die Werte am Bildschirm mit den Befehlen (`str`, `table`) sowie mit `head` ansiehst. Anschliessend probierst du diese Grafikbefehle aus:

```
par(mfrow=c(1,2))
plot(Kopfgewicht~Abstand)
plot(Kopfgewicht~Bewaesserung)
```

Und dann noch diese:

```
par(mfrow=c(1,2))
interaction.plot(Bewaesserung,Abstand,Kopfgewicht)
interaction.plot(Abstand,Bewaesserung,Kopfgewicht)
```

Und nun die Analyse. Der zentrale Befehl lautet

```
summary(aov(Kopfgewicht~Abstand*Bewaesserung))
```

R–Frage 73 Führe die Varianz-Analyse in **R** aus. Kannst du den Output deuten?

Nun wenden wir uns wieder der Formelsprache zu, mit der die Abhängigkeiten eingegeben werden. Der Stern (das Multiplikationszeichen) in `Abstand*Bewaesserung` bedeutet:

sowohl **Abstand** als auch **Bewaesserung** sind Faktoren, und auch eine potentielle Wechselwirkung soll in Betracht gezogen und analysiert werden. Entsprechend enthält der Output drei *P*-Werte.

Statt `Kopfgewicht~Abstand*Bewaesserung` kann man ausführlicher auch eingeben:

`Kopfgewicht~Abstand+Bewaesserung+Abstand:Bewaesserung`

Das Resultat ist das gleiche. `Abstand:Bewaesserung` bezeichnet die Wechselwirkung.

In unserem Kohl-Beispiel hat sich die Wechselwirkung als unbedeutend erwiesen. Das heisst, dass wir sie aus dem Modell eliminieren können und auch sollen.

R-Frage 74 *Mache die Varianz-Analyse noch einmal ohne Wechselwirkung. Wie und warum haben sich die verbleibenden *p*-Werte verändert?*

Und nun der post-hoc-Test, wo sitzen denn die Unterschiede?

R-Frage 75 *Was sagt dir*

`pairwise.t.test(Kopfgewicht,Abstand,p.adjust.method="holm")`

Wie steht es analog mit der Bewässerung?

12.2 Voraussetzungen für die zweifaktorielle ANOVA

Wie schon bei der einfaktoriellen ANOVA müssen die Daten homoskedastisch, normalverteilt, echt metrisch (und echt unabhängig) sein.

Die zweifaktorielle Varianz-Analyse stellt allerdings noch eine weitere Anforderung an die Daten: Sie müssen ausgewogen (balanced) sein, d.h. **jede Faktorstufe muss aus gleich vielen Wiederholungen bestehen**. Das muss natürlich schon in die Versuchsplanung einfließen.

R-Frage 76 *Überprüfe, ob die Kohlköpfe der geforderten Normalverteilung entsprechen:*

`shapiro.test(rstandard(aov(Kopfgewicht~Abstand*Bewaesserung)))`

R-Frage 77 *Überprüfe, ob die Kohlköpfe der geforderten Varianz-Homogenität entsprechen.*

`bartlett.test(Kopfgewicht~interaction(Abstand,Bewaesserung))`

Was macht man nun, wenn eine der Voraussetzungen verletzt ist? Dann muss man mit ei-

ner Rang-Varianz-Analyse vorlieb nehme, der Kruskal-Wallis-Test (H-Test) ist auch hier einsetzbar. Allerdings mit einem entscheidenden Nachteil: Wechselwirkung kann nicht mehr nachgewiesen und auch nicht berücksichtigt werden. Meines Wissens gibt es kein parameter-freies Verfahren, das auf Wechselwirkungen eingehen kann. Der Kruskal-Wallis-Test kann auch nur je einen Faktor gesondert testen. Man muss also zuerst den einen Faktor testen und dann den anderen.

12.3 Zweifaktorielle ANOVA ohne Wiederholung

Kann man eine ANOVA auch machen, wenn pro Stufe/Stufenkombination nur ein einziger Messwert vorliegt? Mit einem gewissen Gefühl für Statistik wirst du jetzt vielleicht sagen: nein, denn dann fehlt die essentielle Information, wie stark die Werte streuen. Im einfaktoriellen Fall hast du damit auch ganz recht. Aber im zweifaktoriellen Fall hat man eben doch zu jeder Stufe mehrere Werte, nämlich entsprechend den Stufen des jeweils anderen Faktors. Tatsächlich ist eine zweifaktorielle Varianzanalyse auch möglich, wenn man nur einen einzigen Wert pro Faktorstufenkombination hat ¹.

Ganz ohne Abstriche verzichtet man aber nicht auf die Wiederholungen. Bei einer „zweifaktoriellen Varianz-Analyse ohne Wiederholungen“ (d.h. mit nur einem Messwert pro Faktorstufenkombination) fehlt die Information für die Wechselwirkung. Man muss sich also darauf verlassen können, dass keine Wechselwirkung vorliegt (z.B. aufgrund inhaltlicher Überlegungen), ansonsten riskiert man einen Fehler.

R-Frage 78 Auf 5 verschiedenen Böden wird das Auftreten von Nematodenzysten untersucht. Dabei werden 5 verschiedene Nachweismethoden angewandt, deren Wahl die gewonnenen Resultate möglicherweise auch stark beeinflusst. Man findet:

		Boden				
		1	2	3	4	5
Methode	1	127	162	155	124	169
	2	166	156	140	95	147
	3	136	123	125	88	166
	4	182	136	115	97	157
	5	133	127	117	98	169

(Du findest die Daten in [nematoden.xlsx](#).) Was ergibt die ANOVA? Wie steht es in diesem Fall mit der Wechselwirkung?

¹Bei mehr als zwei Faktoren geht es sogar mit gewissen Löchern in der Tabelle („incomplete design“), was hier aber nicht thematisiert werden soll.

R-Frage 79 *Untersuche das Nematoden-Beispiel mit dem Kruskal-Wallis-Test.*

Anhang A

Lösungen zu den Fragen im Text

A.1 Lösungen zu den Fragen in Lektion 2

Lösung zu Frage 1

Polynom 2. Grades: Parabel

Lösung zu Frage 2

Gerade parallel zur x-Achse durch die mittlere Reihe könnte ein Versuch sein. Doch ist das wirklich sinnvoll? Nein, denn die Daten legen keinerlei Korrelation nahe (siehe Diskussion im Skript im Anschluss an Frage).

Lösung zu Frage 3

Jeweils von links nach rechts: Obere Reihe: 0.96, 0, -1. Untere Reihe: -0.13, 1, -0.79

A.2 Lösungen zu den Fragen in Lektion 3

Lösung zu Frage 4

Bei $k = 15$ ist die Kurve auf knapp 5% Höhe. Die Wahrscheinlichkeit dafür, dass eine Umfrage unter den geschilderten Umständen $k = 15$ Befürworter ergibt, ist knapp 5%. Wenn man 1000 solcher Umfragen durchführt, werden ungefähr knapp 50 davon $k = 15$ als Resultat haben.

Lösung zu Frage 5

- n ist gleich 6, denn du kannst in der Zeichnung abzählen, dass die einzelne Kugel 6mal auf ein Hindernis trifft, bis sie durch ist.

- Wie gross k ist, lässt sich so nicht sagen, das entscheidet ja eben der Zufall.
- Die Kugel wird mit 50%-iger Wahrscheinlichkeit nach rechts abgelenkt: $p = 0.5 = 50\%$.

Lösung zu Frage 6

Beim Kugelbrett ist $n = 6$ und $p = 0.5$. Nun kommt noch hinzu, dass der Fall „Kugel in der Mitte“ interessiert, d.h. $k = 3$. Dazu kann man nun P berechnen:

$$P(k = 3) = \frac{n!}{k! \cdot (n - k)!} \cdot p^k \cdot (1 - p)^{n-k} = \frac{6!}{3! \cdot (3)!} \cdot 0.5^3 \cdot 0.5^3 = 0.3125 = 31.25\%.$$

A.3 Lösungen zu den Fragen in Lektion 4

Lösung zu Frage 7

a)

Ereignis:	eine Note gewürfelt	Anzahl der Ereign. in einer Kette:	$n = 5$
Erfolg:	6 gewürfelt	Anzahl der Erfolge:	$k = 5$
Erfolgswahrscheinlichkeit			
im Einzelereignis:	6er		$p = \frac{1}{6}$

Hier handelt es sich um den Spezialfall, dass die Anzahl der Ereignisse n in der Kette gleich der Anzahl der Erfolge k ist. Somit ergibt sich mit Gleichung (3.1)

$$\begin{aligned} P(k = n) &= \frac{n!}{k! \cdot (n - k)!} \cdot p^k \cdot (1 - p)^{n-k} = \frac{n!}{n! \cdot 0!} \cdot p^k \cdot (1 - p)^0 \\ &= p^k. \end{aligned} \quad (\text{L.2})$$

Mit $k = 5$ und $p = \frac{1}{6}$ folgt somit

$$P(k = n) = \left(\frac{1}{6}\right)^5 = 0.00013 = 0.013\%.$$

b)

Ereignis:	eine Note gewürfelt	Anzahl der Ereign. in einer Kette:	$n = 5$
Erfolg:	6 gewürfelt	Anzahl der Erfolge:	$k = 0$
Erfolgswahrscheinlichkeit			
im Einzelereignis:	6er		$p = \frac{1}{6}$

Hier handelt es sich um den Spezialfall, dass die Anzahl der Erfolge 0 ist. Somit ergibt sich mit Gleichung (3.1)

$$\begin{aligned}P(k=0) &= \frac{n!}{0! \cdot n!} \cdot p^0 \cdot (1-p)^{n-0} \\&= (1-p)^n.\end{aligned}$$

Mit $n=5$ und $p=\frac{1}{6}$ folgt somit

$$P(k=0) = \left(\frac{5}{6}\right)^5 = 0.40 = 40\%.$$

Lösung zu Frage 8

- a) (1) Ereignis: Molekül in Gefäss $n = 50$
- Erfolg: Molekül in rechter Hälfte $k = 50$
- Erfolgswahrscheinlichkeit: Molekül in rechter Hälfte $p = 0.5$

(2) Es gilt wieder $k = n$. Mit Gleichung (L.2) folgt somit

$$P(k = n) = p^k = 0.5^{50} = 8.9 \cdot 10^{-16}$$

- b) (1) Ereignis: Molekül in Gefäss $n = 50$
- Erfolg: Molekül in rechter Hälfte $k = 25$
- Erfolgswahrscheinlichkeit: Molekül in rechter Hälfte $p = 0.5$

$$(2) P(k) = \frac{n!}{k! \cdot (n - k)!} \cdot p^k \cdot (1 - p)^{n - k} = \frac{50!}{25! \cdot 25!} \cdot 0.5^{25} \cdot 0.5^{25} = 0.11$$

Im Gegensatz zum Resultat unter a) ist die ausgeglichene Verteilung also nicht unwahrscheinlich. Bei grösseren Molekülzahlen ist der Unterschied viel deutlicher (probiere das aus, in dem Du n erhöhst). Man kann also die gleichmässige Verteilung eines Gases in einem Gefäss – die für uns gewissermassen selbstverständlich ist – als rein statistischen Effekt verstehen. Alle Moleküle auf einer Seite des Gefässes ist ungeheuer viel unwahrscheinlicher als ein ungefähr ausgeglichener Zustand. Es braucht dafür überhaupt keine Kräfte, auch keinen Druckausgleich oder ähnliches.

- c) Hier ist wieder die Anzahl der Ereignisse gleich der Anzahl der Erfolge. Die Wahrscheinlichkeit dafür, dass sich alle 25 O₂-Moleküle in der rechten Hälfte des Gefässes befinden, ist somit mit $n = k = 25$ und $p = \frac{1}{2}$ und Gleichung (L.2)

$$(2) P(k = n) = p^k = 0.5^{25} = 2.98 \cdot 10^{-8}.$$

Die Wahrscheinlichkeit dafür, dass sich alle 25 N₂-Moleküle in der linken Hälfte des Gefässes befinden, ist ebenfalls $2.98 \cdot 10^{-8}$. Die Wahrscheinlichkeit dafür, dass beide Fakten gleichzeitig zutreffen, ist das Produkt der beiden Wahrscheinlichkeiten, also

$$(2.98 \cdot 10^{-8})^2 = 8.9 \cdot 10^{-16}.$$

Wiederum kommt man auch hier auf eine exorbitant viel grössere Wahrscheinlichkeit für eine ungefähr gleichmässige Verteilung. Dieselbe Art statistischer Überlegungen wie oben erklärt also auch, warum sich ein Gas nicht von selbst entmischt!

Lösung zu Frage 9

- a) Die Wahrscheinlichkeit dafür, dass von 19 Kindern alle 19 weiblich sind, lässt sich mit unserem Modell berechnen:

① Ereignis:	Kind	$n = 19$
Erfolg:	Tochter	$k = 19$
Erfolgswahrscheinlichkeit:	Tochter	$p = 0.5$

Wieder gilt $k = n$. Somit folgt mit Gleichung (L.2)

$$\textcircled{2} \quad P(k = n) = p^k = 0.5^{19} = 1.9 \cdot 10^{-6}$$

Das ist zwar eine kleine Wahrscheinlichkeit, aber sooo klein, dass deswegen in den Schweizer Zeitungen berichtet werden müsste, wenn der Fall in Nepal einmal eintrifft, ist diese Wahrscheinlichkeit dann doch auch nicht.

Wir sind aber auch noch gar nicht fertig mit der Rechnung. Denn man muss diese Zahl noch multiplizieren mit der Wahrscheinlichkeit, dass eine Familie 19 Kinder hat. Das dürfte noch einmal eine sehr kleine Zahl sein, und insgesamt handelt es sich in der Tat um ein bemerkenswertes Ereignis.

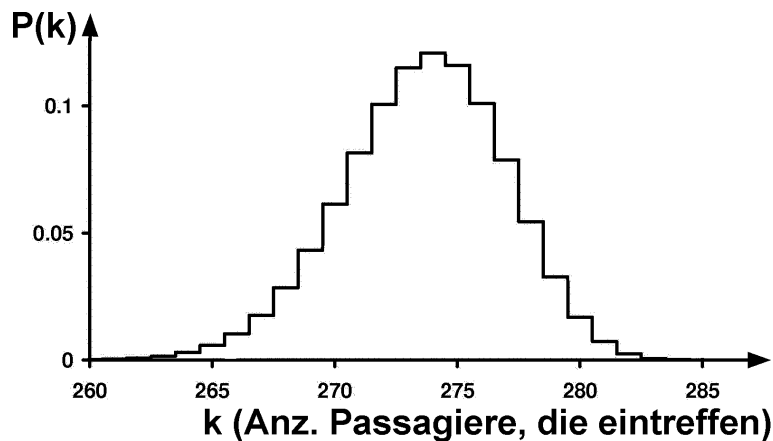
- b) Die Wahrscheinlichkeit für das 20. Mädchen ist 50%. Die 19 vorangegangenen Geburten ändern daran nichts. Der Embryo weiss nicht, wer vor ihm geboren wurde.

Lösung zu Frage 10

- a) Man kann wieder das Binomial-Modell heranziehen:

Ereignis:	verkaufter Sitzplatz	$n = 285$
Erfolg:	besetzter Sitzplatz	k : nicht festgelegt
Erfolgswahrscheinlichkeit:	besetzter Sitzplatz	$p = 0.96$

Im Gegensatz zu vorangehenden Fragestellungen ist hier aber nicht mehr der Fall eines konkreten k -Wertes angesprochen, sondern es geht um den allgemeinen Zusammenhang. k ist jetzt die unabhängige Variable – so wie in der Grundlagen-Mathematik das x in $f(x)$ –, als Funktion dessen die Wahrscheinlichkeit betrachtet wird. Oder als Diagramm:



- b) Der wahrscheinlichste Wert ist (im Rahmen dessen, was wir hier behandeln) immer das, was man auch mit gesundem Menschenverstand findet, in diesem Fall:

$$96\% \text{ von } 285 \quad \text{oder allgemein} \quad p \cdot n = 0.96 \cdot 285 = 273.6 \approx 274.$$

So sieht man es auch im Plot bei a), der mit Gleichung (3.1) berechnet wurde. (Runden: k ist eine ganzzahlige Grösse, Passagiere gibt es schliesslich nur ganze, aber $n \cdot p$ kann auch eine gebrochene Zahl ergeben, also muss man gegebenenfalls runden.)

- c) Das ist nun wieder die bereits durchexerzierte Art Fragestellung, nämlich mit festgelegtem k :

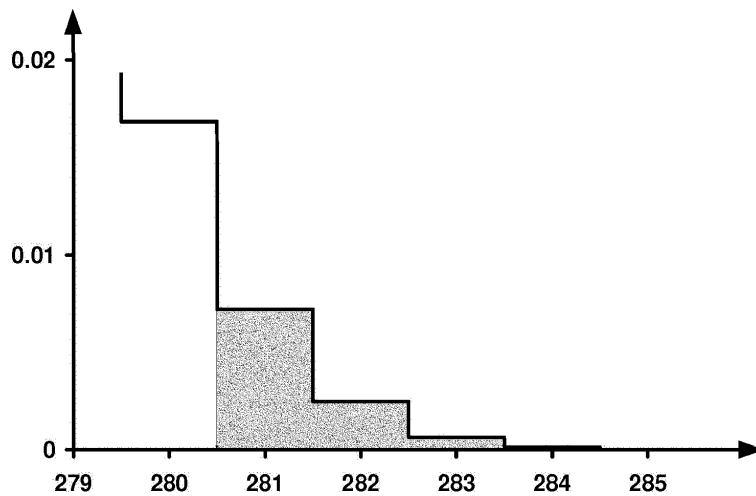
$$k = 280, n = 285, p = 0.96$$

$$\Rightarrow P = \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k} = \frac{285!}{280! \cdot (5)!} \cdot 0.96^{280} \cdot 0.04^5 = 0.017 = 1.7\%.$$

- d) Zu viele Passagiere sind es dann, wenn entweder $k = 281$ auftauchen oder $k = 282$ oder $k = 283$ oder $k = 284$ oder $k = 285$:

$$\begin{aligned} P(k > 280) &= P(k = 281) + P(k = 282) + P(k = 283) + P(k = 284) + P(k = 285) \\ &= \frac{285!}{281! \cdot (4)!} \cdot 0.96^{281} \cdot 0.04^4 + \\ &\quad + \frac{285!}{282! \cdot (3)!} \cdot 0.96^{282} \cdot 0.04^3 + \\ &\quad + \frac{285!}{283! \cdot (2)!} \cdot 0.96^{283} \cdot 0.04^2 + \\ &\quad + \frac{285!}{284! \cdot (1)!} \cdot 0.96^{284} \cdot 0.04^1 + \\ &\quad + \frac{285!}{285! \cdot (0)!} \cdot 0.96^{285} \cdot 0.04^0 = \\ &= \sum_{k=281}^{285} \frac{285!}{k! \cdot (285-k)!} \cdot 0.96^k \cdot 0.04^{285-k}. \end{aligned}$$

Grafisch ist es die Fläche unter der Graphen, ab $k = 281$:



Mit $x_k = 280$, $n = 285$ und $p = 0.96$ kann die Berechnung einfach mittels der R-Funktion

```
1-pbinom(xk,n,p)
```

oder

```
pbinom(xk,n,p,lower.tail=FALSE)
```

erfolgen.

Dies ist eine abkürzende Schreibweise für

```
sum(dbinom(xk+1:n,n,p)).
```

Das Ergebnis lautet $P(k > 280) = 0.01 = 1.0\%$.

Lösung zu Frage 11

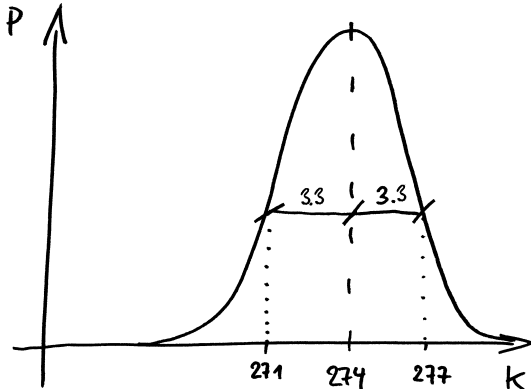
Den Mittelwert haben wir bereits in Frage 10 berechnet:

$$\mu = n \cdot p = 285 \cdot 0.96 = 273.6,$$

die Standardabweichung kommt jetzt noch hinzu:

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)} = \sqrt{285 \cdot 0.96 \cdot 0.04} = 3.3.$$

Diese Verteilung wurde oben bereits aufgetragen, aber mit den jetzigen Informationen hätte man eine ungefähre Skizze auch ohne Berechnung einzelner Werte anfertigen können, ungefähr so:



A.4 Lösungen zu den Fragen in Lektion 5

Lösung zu Frage 12

In der früheren Frage war der Ausschussanteil vorgegeben, und es wurde davon ausgehend eine weitere Fragestellung angegangen. Diesmal ist der Ausschussanteil die eigentlich gesuchte Grösse. Da wir nur eine Stichprobe zur Verfügung haben, lässt sich der Anteil nicht exakt angeben, sondern nur mit einem Konfidenzintervall.

Lösung zu Frage 13

Es geht um eine Stichprobe dichotomer Daten (ja/nein-Stimmen) mit $n = 1000$. Wir definieren nun als **einen** Erfolg **eine** ja-Stimme und haben somit $k = 670$ Erfolge und folglich eine Schätzung des Erfolgsanteil in der ganzen Bevölkerung $\hat{p} = k/n = 0.67$. Es handelt sich um eine grosse Stichprobe, wir können somit die Grenzen des Konfidenzintervalls approximativ mit Formel (4.2) berechnen:

$$\text{a) } \gamma = 0.99 \implies \alpha = 1 - \gamma = 0.01 \implies z_{\alpha/2} = 2.58$$

$$\implies \pi_{u/o} = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} = 0.67 \pm 2.58 \cdot \sqrt{\frac{0.67 \cdot 0.33}{1000}} = 0.670 \pm 0.038 = 67.0\% \pm 3.8\%.$$

$$\text{b) } \gamma = 0.90 \implies \alpha = 1 - \gamma = 0.1 \implies z_{\alpha/2} = 1.64$$

$$\implies \pi_{u/o} = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} = 0.67 \pm 1.64 \cdot \sqrt{\frac{0.67 \cdot 0.33}{1000}} = 0.670 \pm 0.024 = 67.0\% \pm 2.4\%.$$

Damit hast du nun selber berechnet, dass die Konfidenzintervalle um so grösser werden, je höher das Signifikanzniveau ist, also wie in der Figur auf S. 41.

c) Wenn man die Vorgabe $67\% \pm 1\%$ vergleicht mit der Formel $\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$,

so erkennt man, dass die 67% dem \hat{p} entsprechen und das 1% dem $z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$.

Man kann somit letzteres auch direkt gleichsetzen:

$$0.01 = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}.$$

Auflösen nach $z_{\alpha/2}$:

$$z_{\alpha/2} = \frac{0.01}{\sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}} = \frac{0.01}{\sqrt{\frac{0.67 \cdot 0.33}{1000}}} = 0.67.$$

Berechnung der Irrtumswahrscheinlichkeit zu gegebenem $z_{\alpha/2}$:

Für die Gesamtfläche unter dem Graphen der Standardnormalverteilung gilt:

$$\Phi(z_{\alpha/2}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = 1$$

Für die Flächeninhalte unter dem Graphen der Standardnormalverteilung bis und einschliesslich zum Wert $z_{\alpha/2}$ gilt:

$$\Phi(z_{\alpha/2}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{\alpha/2}} e^{-\frac{1}{2}x^2} dx.$$

Wir können $\Phi(z_{\alpha/2})$ in **R** mittels des Befehls

`pnorm($z_{\alpha/2}$)=pnorm(0.67)=0.75`

berechnen. Hier handelt es sich um die rosa Fläche in der linken Figur in Figur 4.1 Die rechte blaue Fläche in Figur 4.1 mit Fläche $\alpha/2$ ergibt sich dann aus der Differenz der Fläche unter der gesamten Normalverteilung (=1) und dem Flächeninhalt unter dem Graphen der Standardnormalverteilung bis und einschliesslich zum Wert $z_{\alpha/2}$:

$$\alpha/2 = 1 - \Phi(z_{\alpha/2}) = 0.25$$

Hier handelt es sich um die halbe Irrtumswahrscheinlichkeit und somit $\alpha = \gamma = 50\%$, d.h. die Chance ist 50:50, dass die Vorhersage trifft.

Lösung zu Frage 14

Wir können wieder die Konfidenzintervalle berechnen. Wenn sie sich überlappen, müssen wir damit rechnen, dass eine Partei die andere überholt.

Für die Frage, ob HB noch überholt werden kann, reicht es aus, den nächsten Verfolger (EVA) mit einzubeziehen, denn wenn die es nicht schaffen können, werden es die noch weiter zurückliegenden auch nicht schaffen.

Für $\alpha = 1 - \gamma = 0.05$ erhalten wir $z_{\alpha/2} = 1.96$ (Gleichung 4.3).

Mit $n = 3000$ und Formel (4.2) können wir somit die jeweiligen Konfidenintervalle für HB

und EVA berechnen:

$$\text{HB: } k = 950, \hat{p} = \frac{k}{n} = \frac{950}{3000} = 0.317,$$

$$\pi_{u/o} = 0.317 \pm 1.96 \cdot \sqrt{\frac{0.317 \cdot 0.683}{3000}} = 0.317 \pm 0.017$$

Die untere Grenze ist damit bei ziemlich genau $\hat{p}_u = 30\%$, soviel werden die HBs mit 95%iger Sicherheit mindestens schaffen.

$$\text{EVA: } k = 910, \hat{p} = \frac{910}{3000} = 0.303,$$

$$\pi_{u/o} = 0.303 \pm 1.96 \cdot \sqrt{\frac{0.303 \cdot 0.697}{3000}} = 0.303 \pm 0.016$$

Damit ist die obere Grenze der Möglichkeiten bei $\hat{p}_o = 31.9\%$, was mehr ist als das Minimum bei den HBs. Die EVAs können also durchaus noch überholen.

Lösung zu Frage 15

Man könnte die Sache so anpacken: Die Zählung 2012 gibt eine Schätzung für die wahre Anzahl, mit Konfidenz-Intervall, und nun schaut man, ob die Zahl von 2013 in diesem Intervall liegt. Wenn ja, kann es sich ohne Weiteres um eine Zufallsschwankung halten, die nichts zu bedeuten hat, einfach, weil eine Stichprobe halt mal etwas grössere Werte liefert und ein ander Mal etwas kleinere.

Da $x = 339 > 100$ ist, können wir Gleichung (4.6) verwenden. Da es um eine erste Abklärung geht, ist $\gamma = 95\%$ keine schlechte Wahl. Also:

$$\lambda_{u/o} = 339 \pm 1.96 \cdot \sqrt{339} = 339 \pm 36.$$

Mit 95%iger Wahrscheinlichkeit liegt der „wahre“ Wert (d.h. hier das, was der langjährige Mittelwert wäre, wenn sich nichts veränderte) zwischen $\lambda_u = 339 - 36 = 303$ und $\lambda_o = 339 + 36 = 375$. Der Wert von 2013 liegt deutlich ausserhalb des Intervalles. Das könnte die Vermutung nähren, dass man einen grossen Schritt vorwärts gemacht hat. Allerdings setzt dies voraus, dass die Bedingungen (z.B. Töff-Wetter) in beiden Jahren identisch waren.

Genau so gut hätte man die Zahl von 2013 als Ausgangspunkt für ein Konfidenz-Intervall nehmen können. Überhaupt ist unser Vorgehen hier nicht wirklich perfekt. Es handelt sich hier nämlich um die Situation *zweier* zu vergleichender Stichproben (von denen jede dem Zufall unterworfen ist), während unsere Formeln eigentlich für die Analyse nur einer Stichprobe gemacht sind. Aber unser Vorgehen ist erlaubt im Sinne einer Abklärung der ungefähren Verhältnisse. Eine exakte Analyse müsste mit modifizierten Methoden vorgehen, die wir aber erst später besprechen werden.

A.5 Lösungen zu den Fragen in Lektion 6

Lösung zu Frage 16

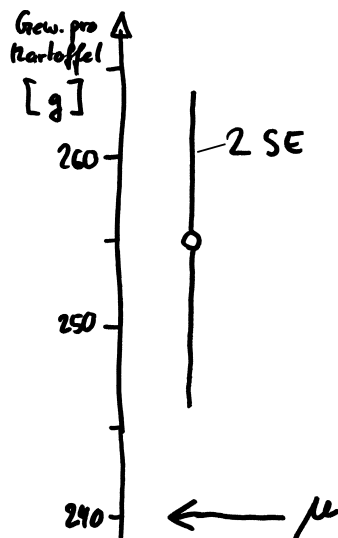
- a) Es geht um 1 Verteilung, nämlich die Häufigkeitsverteilung des Gewichtes der Kartoffeln (metrischer Datentyp).

b) Um den Mittelwert.

A.6 Lösungen zu den Fragen in Lektion 7

Lösung zu Frage 17

Eine Darstellung der Situation könnte z.B. so aussehen:



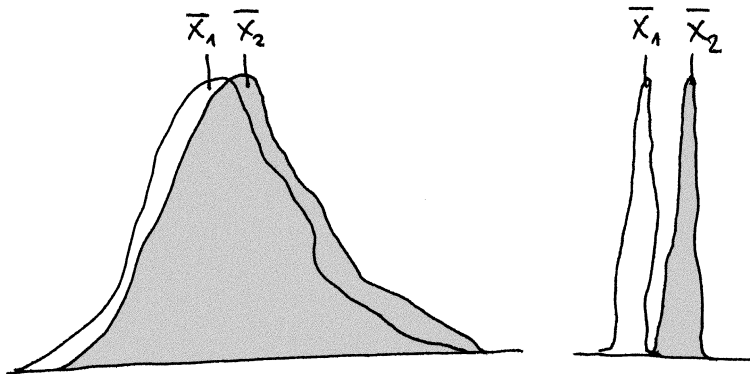
Es zeichnet sich ein signifikanter Unterschied ab, der Referenzwert liegt deutlich ausserhalb des Fehlerbalkens.

Hier und in vielen folgenden Beispielen präsentiere ich ganz bewusst eine Handzeichnung. Es geht jeweils um eine schnelle Skizze, die das Wesentliche zeigt, und nicht um etwas Hochpräzises oder Ästhetisches.

Lösung zu Frage 18

Methode (2) wird das genauere Resultat liefern (bzw. man kann einen vorhandenen Unterschied schon mit einer kleineren Stichprobe nachweisen). Das hat mit der Streuung zu tun. Die Streuung der einzelnen Werte ist ja das Grundproblem von allem, was mit Stichproben zu tun hat. Je kleiner die Streuung, desto signifikanter ist eine gegebene Mittelwert-Differenz zwischen zwei Stichproben. In der folgenden

Figur ist die Differenz zwischen \bar{x}_1 und \bar{x}_2 im Fall rechts (Streuung klein) bestimmt hoch signifikant, im Fall links mit der grossen Streuung aber vielleicht nicht:



Entscheidend ist in diesem Zusammenhang, woher die Ursache der Streuung rührt. Bei den Fluglärmwahrnehmungen dürfte die Hauptursache für die Streuung sein, dass verschiedene Menschen den Lärm unterschiedlich wahrnehmen (der eine ist vom psychischen Typ her sehr empfindlich auf Störungen, schläft bei offenem Fenster, möchte morgens zur Zeit der ersten Anflüge noch schlafen – der andere arbeitet zu dieser Zeit vielleicht schon, ist an Lärm gewöhnt, usw.). Diese Empfindlichkeit wird sich von der ersten zur nächsten Befragung bei den meisten Leuten nicht gravierend ändern. Das absolute Urteil einer Stichprobe ist unsicher, weil es Zufall ist, ob eher viele Empfindliche dabei sind oder eher wenige. Aber wenn man bei beiden Stichproben die gleichen Leute nimmt, liegen sozusagen beide Stichproben gleich falsch. Die Differenz ist dann wieder aussagekräftig. Tatsächlich arbeitet man in der Situation (2) mit der *Differenz* der zusammengehörigen Werte.

Lösung zu Frage 19

Eine Problemstellung mit gepaarten Stichproben müsste die gleichen Personen mehrfach befragen (und dabei jeweils die Noten einer Person vergleichen.) Man könnte z.B. untersuchen, ob es einen signifikanten Unterschied in den Mathematiknoten vom 1. Semester zum 2. Semester gibt. Dann weiss man aber immer noch nicht, woher dieser Unterschied kommt. Darauf geben die statistischen Tests nie eine Antwort, Kausalzusammenhänge werden nicht aufgedeckt.

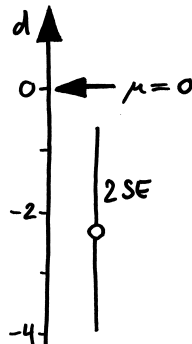
Lösung zu Frage 20

- a) Benzinsorten: 2 abhängige Stichproben
- b) Einkaufszentrum: 1-Stichproben-Fall
- c) Glühbirnen: 2 unabhängige Stichproben
- d) Leuchtstoffröhren: 1-Stichproben-Fall

Lösung zu Frage 21

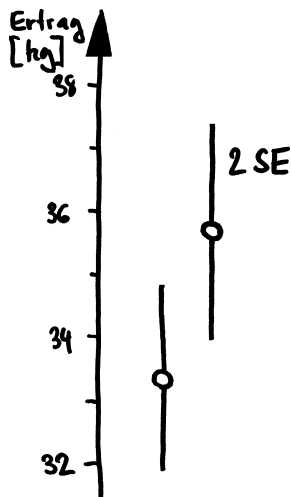
- a) Ausgangspunkt sind hier die Differenzwerte $d_i = x_{1i} - x_{2i}$

Der Mittelwert des Differenzvektors $\bar{d} = -2.3$ und Standardfehler des Differenzvektors $SE(d) = \sigma/\sqrt{n} = 0.8$ führen zu folgender Skizze (Mittelwert $\pm 2 SE$):



Die Paardifferenzen $d_i = x_{1i} - x_{2i}$ scheinen sich deutlich von $\mu = 0$ zu unterscheiden.

- b) Hier werden die beiden Jahrgänge je als Stichprobe aufgetragen wie beim Fall zweier unabhängiger Stichproben geübt. Mit $\bar{x}_1 = 33.3$ und $SE(x_1) = 0.7$ sowie $\bar{x}_2 = 35.7$ und $SE(x_2) = 0.9$ erhalten wir folgende Skizze:



Der Unterschied zwischen den Jahrgängen scheint nicht signifikant zu sein (was die genaue Berechnung bestätigt).

Lösung zu Frage 22

Mit Schätzungen macht man ausgehend von einer einzigen Stichprobe direkt eine Aussage über die Population, aus der die Stichprobe gezogen ist. [Den 2-Stichproben-Fall gibt es bei Schätzungen ebenfalls, aber die grundlegende Situation ist die gleiche, einfach mit Fokus auf der Differenz zwischen den Werten der beiden Stichproben, und man macht eine Aussage darüber, wie diese Differenz in der Population aussieht.]

In einem statistischen Test geht es immer um einen *Vergleich* zwischen *zwei* (oder manchmal

noch mehr) Dingen. Das können zwei (oder mehr) Stichproben sein, oder eine Stichprobe und ein fixer Vergleichswert.

Man muss aber auch noch anfügen, dass es Situationen gibt, wo Schätzung und Test äusserst nahe beieinanderliegen.

A.7 Lösungen zu den Fragen in Lektion 8

Lösung zu Frage 23

Der zentrale Unterschied besteht darin, dass der t -Test die Lage (Mittelwert) verschiedener Stichproben vergleicht, der F -Test hingegen die Streuung (Standardabweichung oder Varianz). Das sind zwei völlig verschiedene Fragestellungen.

Lösung zu Frage 24

- a) In der Aufgabenstellung wird nur eine Stichprobe erwähnt, bestehend aus 20 Personen. Falsch wäre es, die 11 richtig Tippenden und die 9 falsch Tippenden je als eine unabhängige Stichprobe zu betrachten.
- b) Es wäre doch zu erwarten, dass 50% der Biertester auf Bier A tippt und die anderen 50% auf Bier B. Also liegt gerade die Hälfte der Leute richtig. Das ist der Vergleichswert: $\pi_0 = 0.5$.

Lösung zu Frage 25

Hier lautet die Nullhypothese, dass es *keinen* signifikanten Unterschied zwischen $\pi_0 = 0.5$ (d.h. ununterscheidbaren Bieren) und dem Wert der Degustatoren $\pi = 0.55$ gibt. Unsere Testerei hat ergeben, dass man die Nullhypothese, also ununterscheidbare Biere nicht verwerfen kann und somit behalten muss. Soweit wäre das in deinem Sinne als Brauer. Man muss aber im Kopf behalten, dass ein Test nie die Nullhypothese beweisen kann. Wir haben lediglich auch keinen Beweis für die Unterscheidbarkeit gefunden.

Lösung zu Frage 26

Im Datentyp. Einen Anteilstest kann man auf dichotome Daten anwenden, t - und F -Test auf metrische.

A.8 Lösungen zu den Fragen in Lektion 9

Lösung zu Frage 27

Ganz wichtig ist, dass man keine Informationen wegwirft. Mit der Einteilung der Auberginen-Pflanzen in nur zwei Kategorien (normal/überdurchschnittlich) hat man dies getan. Viel

aussagekräftiger wäre ein metrisches Mass für den Ertrag gewesen (z.B. Ertrag in kg). Es kann sehr gut sein, dass man die Verbesserung dann mit einem t -Test hätte beweisen können.

Lösung zu Frage 28

- a) Beobachtete Stichprobe: Die 15, 15 und 20 Stück. Die beobachteten Häufigkeiten sind immer Zahlen, die man wirklich in dieser Form gezählt hat.
- b) Erwartete Verteilung: Die prozentualen Anteile an den Stückzahlen der drei Produktkategorien (45%, 30% und 25%) ist die Vergleichsverteilung, mit der man die beobachtete vergleicht (\rightarrow Erwartungswerte).

Lösung zu Frage 29

	A	B	C	D
a)	Streuung	Form/Verteilung als Ganzes	Anteil	Lage
b)	F -Test	χ^2 -Test	Anteils-Test	t -Test, Wilcoxon-Test
c)	metrisch	nominal oder ordinal. Beliebige Datentypen können durch das Zusammenfassen von Klassen zu nominalen Daten gemacht werden.	dichotom	metrisch; für Wilcoxon-Test auch ordinal
d)	Die beiden Länder sind im Durchschnitt gleich heiss, aber in einem Land variieren die Temperaturen viel stärker als im anderen.	Im Land mit der ausgezogenen Verteilung hat man hin und wieder kaltes Wetter, aber häufiger warmes und nie heisses. Im Land mit der gestrichelten Verteilung hat man nie kaltes, aber häufig kühles Wetter und seltener heisses.	Beispiel für etwas Dichotomes wäre hier: Frosttag ja/nein. Ein Land hat einen grösseren Anteil Frosttage als das andere.	Die Schwankungen sind in beiden Ländern gleich, aber die Durchschnittstemperatur ist einem Land höher als im anderen.

Lösung zu Frage 30

- a) Mittelwertschätzung, t -Test, Wilcoxon-Test, F -Test, χ^2 -Test;

- b) Schätzung für nominale und ordinale Stichproben, Wilcoxon-Test, χ^2 -Test;
- c) Schätzung für nominale und ordinale Stichproben, χ^2 -Test;
- d) Anteilschätzung, Anteils-Test, χ^2 -Test;
- e) Anzahlschätzungen;
- f) entsprechende Variante des t -Testes oder des Wilcoxon-Testes;
- g) mehr-Stichproben- χ^2 -Test;
- h) t -Test, Wilcoxon-Test, χ^2 -Test;
- i) F -Test;
- j) Anteils-Test, χ^2 -Test;
- k) χ^2 -Test.

Lösung zu Frage 31

c)

A.9 Lösungen zu den Fragen in Lektion 10

Lösung zu Frage 32

- a) abhängige Variable: Nährwert (metrisch), abhängige Variable: Käsesorte (nominal)
- b) Z.B. die Zahlen zu Tilsiter sind eine Stichprobe. Die Nährwertangaben der anderen Käsesorten stellen je auch eine Stichprobe dar.
- c) Weil es z.B. beim Tilsiter noch viel mehr Produkte gibt als nur die 5 ausgewählten.
- d) cZ.B. die 1120, die beim Tilsiter als erste Zahl genannt ist.

Lösung zu Frage 33

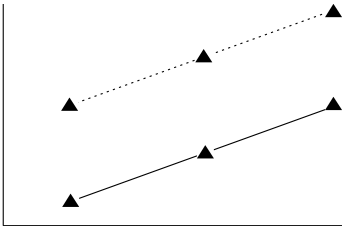
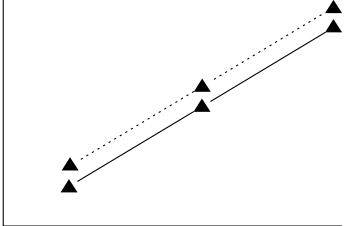
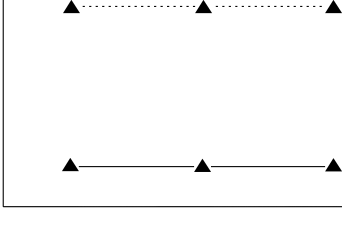
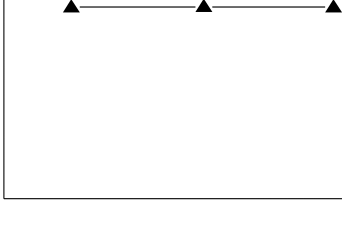
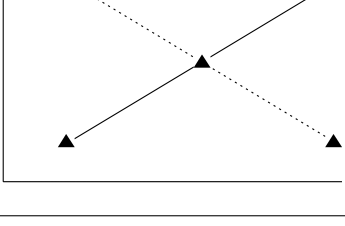
Nein. Der χ^2 -Test arbeitet mit Häufigkeiten. Hier hingegen sind einzelne Zahlenwerte eines metrischen Merkmals aufgeführt.

A.10 Lösungen zu den Fragen in Lektion 11

Lösung zu Frage 34

Der p -Wert 0.002837 ist die Wahrscheinlichkeit der Nullhypothese: Es liegt kein signifikanter Unterschied zwischen den Mittelwerten der Faktor-Stufen vor. Nehmen wir nun an, dass $\alpha = 0.05$, dann gilt $p < \alpha = 0.05$. Wir können somit die Nullhypothese ablehnen. Es liegt somit ein signifikanter Unterschied zwischen den Faktor-Stufen vor.

Lösung zu Frage 35

	Abhängigkeit von der Düngung	Abhängigkeit von der Sorte	Wechselwirkung Sorte \longleftrightarrow Düngung
	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Anhang B

Ergänzungen ANOVA (fakultativ)

B.1 Grundidee

Die Varianzanalyse ist daran interessiert, die **Unterschiede zwischen den beobachteten Werten und dem Gesamtmittelwert der Stichprobe zu erklären**. Jede der Faktorstufen stellt eine Verteilung dar (Abbildung B.1). Zum Verständnis der ANOVA führen wir

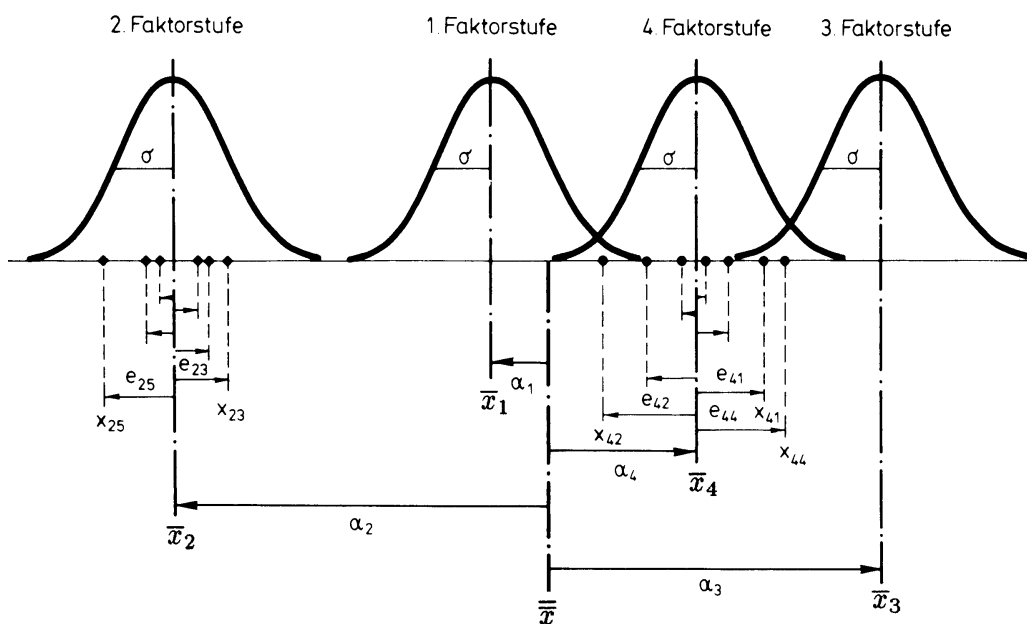


Abbildung B.1: Verteilung der Faktorstufen

folgende Bezeichnungen ein (siehe Abbildung B.1)

- k : Anzahl der Faktorstufen
- n_j : Anzahl der Wiederholungen innerhalb Faktorstufe j
- $N = \sum_{j=1}^k n_j$: Anzahl aller Messungen (ungeachtet der Faktorstufenzugehörigkeit)
- x_{ij} : i . Wiederholung in der j . Faktorstufe
- $T_j = \sum_{i=1}^{n_j} x_{ij}$ Summe aller Wiederholungen innerhalb der Faktorstufe j ,
- $T = \sum_{j=1}^k (\sum_{i=1}^{n_j} x_{ij}) = \sum_{j=1}^k T_j$: Summe aller Messwerte.
- $\bar{\bar{x}}$: Gesamtmittelwert aller Messungen (ungeachtet der Faktorstufenzugehörigkeit).

$$\bar{\bar{x}} = \frac{\sum_{j=1}^k (\sum_{i=1}^{n_j} x_{ij})}{N} = \frac{T}{N}$$

- \bar{x}_j : Mittelwert der Faktorstufe j ¹:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

- e_{ij} : Restfehler, Versuchsfehler oder Residuum:
Der Name rührt daher, dass die Differenz zwischen Wiederholung und Faktorstufenmittelwert

$$e_{ij} = x_{ij} - \bar{x}_j$$

in vielen Fällen mit der Ungenauigkeit einer Messung zu tun hat. Die Wiederholungen einer Faktorstufe scharen sich also um den Faktorstufenmittelwert, streuen aber mehr oder weniger.

- a_j : Der fester Effekt bezeichnet die Abweichung der Stufenmittelwerte \bar{x}_j vom Gesamtmittelwert $\bar{\bar{x}}$

$$a_j = \bar{x}_j - \bar{\bar{x}}$$

Die Grundidee der ANOVA ist nun jeden Messwert x_{ij} als Summe von Gesamtmittelwert, festem Effekt der Faktorstufen und Versuchsfehler zu verstehen

$$x_{ij} = \bar{\bar{x}} + a_j + e_{ij}$$

und die Abweichungen der individuellen Werte vom Gesamtmittelwert zu erklären.

¹Faktorstufen werden im Englischen "groups" genannt, häufig wird in der deutschsprachigen Literatur der Faktorstufenmittelwert deshalb auch Gruppenmittelwert genannt.

B.2 Quadratsummen

Eine Varianzanalyse versucht, die Abweichungen der individuellen Werte vom Gesamtmittelwert zu erklären. Hierzu werden die Abweichungen jeden einzelnen Messwerts vom Gesamtmittelwert quadriert und anschliessend aufsummiert. Dies führt zur Gesamtquadratsumme SQT (da engl. "Sum of Squares, total"):

$$\begin{aligned} SQT &= \sum_{i=1}^{n_j} \sum_{j=1}^k (x_{ij} - \bar{x})^2 \\ &= \sum_{i,j} x_{ij}^2 - \frac{T^2}{N} \end{aligned} \quad (\text{A.1})$$

Diese Quadratsumme wird im Rahmen der Varianzanalyse in zwei Teile zerlegt: Das, was durch die Stufenzugehörigkeit erklärt werden kann, und das, was nicht erklärt werden kann. Das erstere nennt sich die Quadratsumme zwischen der Gruppen (SQZ) und basiert auf den Abweichungsquadraten zwischen den Faktorstufenmittelwerten und dem Gesamtmittelwert:

$$\begin{aligned} SQZ &= \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^k n_j a_j^2 \\ &= \sum_{j=1}^k \frac{T_j^2}{n_j} - \frac{T^2}{N} \end{aligned} \quad (\text{A.2})$$

Der zweite Anteil ist die Quadratsumme innerhalb der Gruppen (Faktorstufen) SQI und basiert auf den Abweichungsquadraten zwischen jeder individuellen Messung vom jeweiligen Gruppenmittelwert:

$$\begin{aligned} SQI &= \sum_{i=1}^{n_j} \sum_{j=1}^k (x_{ij} - \bar{x}_j)^2 = \sum_{i=1}^{n_j} \sum_{j=1}^k e_{ij}^2 \\ &= \sum_{i=1}^{n_j} \sum_{j=1}^k x_{ij}^2 - \sum_{j=1}^k \frac{T_j^2}{n_j} \end{aligned} \quad (\text{A.3})$$

Es gilt die folgende Beziehung zwischen den Quadratsummen

$$SQT = SQZ + SQI.$$

B.3 Berechnung der Teststatistik: die Prüfgrösse F

Zur Berechnung der Teststatistik F werden die mittleren Quadratsummen MQZ und MQI benötigt ("MS" für engl. "mean squares"). Dazu werden die Quadratsummen durch ihre jeweiligen Freiheitsgrade dividiert. Mit $df_1 = k - 1$ folgt

$$MQZ = \frac{SQZ}{df_1} = \frac{SQZ}{k - 1}.$$

Mit $df_2 = N - k$ folgt

$$MQI = \frac{SQI}{df_2} = \frac{SQI}{N - k}.$$

Anschliessend wird die Teststatistik F folgendermassen berechnet:

$$F_b = \frac{MQZ}{MQI}$$

B.4 Rezept für einfaktorielle ANOVA ohne R

Frage: Unterscheiden sich die Mittelwerte der Faktorstufen signifikant voneinander?

1. Formuliere H_0 und H_1 .
2. Lege α fest.
Überprüfe die Voraussetzungen: Normalverteilung in den einzelnen Stichproben (Überprüfung in der Regel nur bei kleinen Stichproben notwendig), Varianzhomogenität über alle Stichproben.
3. Berechnung des P-Wertes: Die Varianzanalyse ist ein F -Test mit der Prüfgrösse

$$F_b = \frac{MQZ}{MQI}$$

Nun berechnen wir die Wahrscheinlichkeit der Nullhypothese mit den mit den Freiheitsgraden

$$df_1 = k - 1 \quad \text{und} \quad df_2 = N - k$$

mittels

```
pf(F_b, df_1, df_2, lower.tail=TRUE)
```

4. Falls $P < \alpha$ ist, wird H_0 verworfen.

B.5 Beispiel: Lebenserwartung

Führen wir nun eine ANOVA anhand des Lebenserwartungsbeispiels durch.

1. H_0 : Die Mittelwerte der Faktorstufen unterscheiden sich nicht signifikant.
 H_1 : Die Mittelwerte der Faktorstufen unterscheiden sich signifikant.
2. $\alpha = 0.01$. Voraussetzungen: Bei kleinen Fallzahlen, wie hier, ist eine Abweichung von der Normalverteilung nur bei eklatanten Ausreissern wahrscheinlich. Dies ist hier nicht der Fall. Die Varianzhomogenität ist mit geeignetem Test (z.B. `bartlett.test`) überprüft worden und ist gegeben.

3. Berechnung der Teststatistik:

(a) $k = 5, N = 31,$

$T_1 = 428, T_2 = 353, T_3 = 318, T_4 = 461, T_5 = 465, T = 2025.$

(b) $SQZ = 1277.7, MQZ = 1277.7/4 = 319.4$

(c) $SQT = 1982.8, SQI = 1982.8 - 1277.7 = 705.1, MQI = 705.1/26 = 27.1$

(d)

$$F_b = \frac{319.4}{27.1} = 11.79$$

4. Mit $df_1 = k-1 = 4, df_2 = N-k = 26$. erhalten wir mit `pf(11.78,4,26,lower.tail=FALSE)` einen $P - Wert$ von $1.4 \cdot 10^{-5}$. Die Nullhypothese kann somit verworfen werden. Die Mittelwerte der Faktorstufen unterscheiden sich demnach signifikant.