Handouts zum Atelier "Statistik in Maturaarbeiten"

Hansruedi Künsch

TMU Frauenfeld, 11. September 2019

Das Buch "Forschen, aber wie?", Martin Ludwig und Georges Hartmeier (Hrsg.), hep Verlag 2019, bietet eine Einführung in wissenschaftliche Methoden für schriftliche Arbeiten und will insbesondere Gymnasiastinnen und Gymnasiasten bei der Maturaarbeit unterstüzen. Das Kapitel 6 in diesem Buch behandelt die statistische Auswertung von Daten. In diesem Atelier will ich dieses Kapitel kurz vorstellen und Sie insbesondere auch etwas mit dem Statistikprogramm *R* vertraut machen, das für die Auswertungen und die grafischen Darstellungen empfohlen wird. Bitte bringen Sie Ihren Laptop mit und installieren Sie im Voraus *R* und die Benutzeroberfläche *R*-Studio, siehe unten. Wenn Probleme auftauchen, können wir Ihnen am 11.9. in den Pausen weiterhelfen.

1 Das Statistikprogramm *R*

Hier folgt der Text des Abschnitts 6.5 aus dem Buch.

R ist ein Software-System für die statistische Analyse von Daten. Es ist sehr weit verbreitet in Lehre und Forschung an Universitäten und auch in der Wirtschaft und in der Industrie. Es ist sehr flexibel und bietet sehr viele Methoden für fast jede denkbare Fragestellung an. *R* ist ein Open-source-Programm und kann kostenlos von www.R-project.org. heruntergeladen werden. Es wird empfohlen, *R* über die Benutzeroberfläche *R*-Studio zu benutzen. Nach der Installation von *R* können Sie noch *R*-Studio von www.rstudio.com als Gratisversion herunterladen.

In diesem Anhang werden ein paar Hinweise zum Gebrauch von *R* für die hier besprochenen Methoden gegeben. Weitere Informationen finden Sie in der Literatur und auch online auf www.statmethods.net oder www.rstudio.com/online-learning.

R ist befehlsorientiert, d.h. man arbeitet mit Befehlen und mit Funktionen, nicht mit Menus wie in Excel. Funktionen haben meist neben den obligatorischen Argumenten noch optionale Argumente mit einem default-Wert, der bei Bedarf geändert werden kann. Im Reiter "Hilfe" (oder mit `help("function")`) erhalten Sie Informationen, was eine Funktion berechnet und wie die obligatorischen und die optionalen Argumente heissen.

Kleine Datensätze gibt man am besten direkt von Hand in *R* ein. Als Beispiel wird die Eingabe der in 6.2.1 erwähnten Daten über die Körpergrösse gezeigt. Im Folgenden wird rot geschrieben, was Sie eingeben (diese Zeilen beginnen mit `>` oder mit `+` wenn die Eingabe mehr als eine Zeile benötigt), und schwarz, was das *R* zurückgibt.

```
> groesse <- c(161, 165, 167, 168, 171, 171, 171, 173, 175, 178, 179, 179,
+ 179, 180, 181, 183, 185, 185, 187, 191)
> groesse
[1] 161 165 167 168 171 171 171 173 175 178 179 179 179 180
[15] 181 183 185 185 187 191
```

```
> summary(groesse)
```

```
Min.   1st Qu.  Median    Mean   3rd Qu.  Max.
161.0 171.0    178.5    176.4  181.5    191.0
```

Die Funktion `c` (engl. combine = kombinieren) fasst die Zahlen zwischen den Klammern zu einem Objekt, einem sogenannten Vektor, zusammen. Mit `<-` wird das auf der rechten Seite stehende Objekt dem auf der linken Seite angegebenen Namen (Leerschläge oder Sonderzeichen sind dabei nicht erlaubt) zugewiesen und ist nachher unter diesem Namen gespeichert. Gibt man den Namen eines Objekts ohne eine Zuweisung ein, dann werden die Werte angezeigt. Die Funktion `summary` gibt die wichtigsten Kennzahlen an.

In *R* können Sie nicht mit der Backspace-Taste löschen. Wenn Sie eine fehlerhafte Eingabe gemacht haben, meldet Ihnen *R* den Fehler und gibt Ihnen wieder das Zeichen `>`, damit Sie korrekt eingeben können.

Mathematische Operation für Vektoren werden elementweise ausgeführt, wenn dies sinnvoll ist. So kann man z. B. mit einem einzigen Befehl die Körpergrösse in Metern berechnen:

```
> groesse.m <- groesse/100
> groesse.m
```

```
[1] 1.61 1.65 1.67 1.68 1.71 1.71 1.71 1.73 1.75 1.78 1.79 1.79 1.79 1.80
[15] 1.81 1.83 1.85 1.85 1.87 1.91
```

Für gewisse Methoden mit kategorialen Daten müssen die Daten nicht als Vektoren, sondern als Matrix eingegeben werden. Dazu verwendet man die Funktion `matrix`. Im Beispiel der Anzahl Mahlzeiten in 6.2.4, ergab sich folgende Tabelle:

	männlich	weiblich	Randhäufigkeit
$\leq$ 2 Mahlzeiten	4	10	14
3 Mahlzeiten	16	12	28
$\geq$ 4 Mahlzeiten	12	6	18
Randhäufigkeit	32	28	6

Die Zahlen dieser Tabelle werden wie folgt als Matrix eingegeben:

```
> meals <- matrix(c(4,16,12,10,12,6),nrow=3)
> meals
```

```
     [,1] [,2]
[1,]    4   10
[2,]   16   12
[3,]   12    6
```

Per default müssen die Zahlenwerte spaltenweise eingegeben werden. Das Argument `nrow=3` bewirkt, dass eine Matrix mit 3 Zeilen gebildet wird. Die Anzahl Spalten bestimmt sich dann daraus, wie viele Zahlen in der Tabelle stehen. Möchte man die Daten lieber zeilenweise eingeben, dann benutzt man das zusätzliche Argument `byrow=TRUE`:

```
> matrix(c(4,10,16,12,12,6),nrow=3,byrow=TRUE)
```

```
     [,1] [,2]
[1,]    4   10
[2,]   16   12
[3,]   12    6
```

`colnames` gibt den Spalten Namen und rownames gibt den Zeilen Namen, die *R* dann beim Erstellen von Diagrammen benutzt:

```
> colnames(meals) <- c("männlich","weiblich")
> rownames(meals) <- c("< 3","3","> 3")
> meals
      männlich    weiblich
<3    4           10
3     16          12
>3    12          6
```

Grössere Datensätze, die als Excel-Datei vorhanden sind, sollten zuerst als CSV-Dokument (comma separated values) gespeichert werden ("Speichern unter "/ "Format"/ "Option(*.csv)" auswählen). Mit dem Befehl `read.csv` können die Daten dann direkt in *R* eingelesen werden. Die Spalten in der Excel-Datei sollten den verschiedenen Merkmalen entsprechen und sie sollten Namen haben. Kategoriale Merkmale müssen nicht als Zahlen kodiert werden, sondern können als Text vorliegen. Sind die Daten als daten.csv gespeichert, dann verwenden Sie den Befehl

```
> daten <- read.csv('daten.csv',header=TRUE)
```

Wenn Sie die Spalten mit "merkmal1" und "merkmal2" bezeichnet haben, dann stehen anschliessend die Werte in *R* unter den Namen `daten$merkmal1` bzw. `daten$merkmal2` zur Verfügung.

2 Konzepte und Methoden

Die in einer Studie untersuchten Objekte oder Subjekte nennt man Merkmalsträger oder statistische Einheiten. In einer Meinungsbefragung sind die befragten Personen die Merkmalsträger, in einem Wachstumsexperiment mit Pflanzen sind es die einzelnen Pflanzen oder die Probeflächen auf dem Versuchsfeld und bei einer wiederholten physikalischen Messung die einzelnen Durchführungen des Messexperiments. Oft möchte man nicht nur über die untersuchten Merkmalsträger Aussagen machen, sondern über eine grössere Menge von Merkmalsträgern, die als Grundgesamtheit oder Population bezeichnet wird. Die schliessende Statistik erlaubt es, mit Hilfe von statistischen Tests und Vertrauensintervallen die Unsicherheit beim Rückschluss von einer Stichprobe auf die Grundgesamtheit anzugeben. Dabei wird vorausgesetzt, dass die vorliegende Stichprobe eine Zufallsstichprobe ist.

In den folgenden Abschnitten sind diese Begriffe kurz erklärt. Dabei werden grösstenteils Formulierungen aus dem Buch verwendet.

2.1 Statistische Tests, Nullhypothese, *P*-Wert

Ein Hypothesentest beantwortet Fragen wie zum Beispiel: "Wie plausibel ist es, dass der Unterschied zwischen Frauen und Männern, der in der Stichprobe vorliegt, auch in der Grundgesamtheit besteht?". Zur Einführung wird im Buch folgendes Beispiel betrachtet: Angenommen, in einer Meinungsbefragung geben 62 von 100 Jugendlichen einer Schule an, dass sie Coca-Cola gegenüber Pepsi bevorzugen. Wie gross ist dann etwa der entsprechende Anteil unter allen Jugendlichen derselben Schule, oder sogar unter allen Jugendlichen in der ganzen Schweiz? Wohl kaum exakt 62 Prozent, aber vermutlich irgendwo in der Nähe. Könnte er zum Beispiel auch 65 Prozent sein, oder sogar 70 Prozent oder 80 Prozent?

Um diese Fragen zu beantworten, wird der sogenannte Binomialtest verwendet. In *R* steht dafür die Funktion `binom.test` zur Verfügung. Der Befehl und die Ausgabe sehen wie folgt aus:

```
> binom.test(x=62,n=100)

Exact binomial test
data: 62 and 100
number of successes = 62, number of trials = 100, p-value = 0.02098
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.5174607 0.7152325
sample estimates:
probability of success
0.62
```

R verwendet den Begriff "Erfolg" (success) für eine befragte Person, die Coca-Cola bevorzugt. In der Stichprobe gibt es somit 62 "Erfolge". Wenn die befragten Personen zufällig ausgewählt wurden, ist die Wahrscheinlichkeit für einen "Erfolg" gleich dem Anteil der Personen in der Grundgesamtheit, die Coca-Cola bevorzugen. Den wahren Wert dieser Wahrscheinlichkeit kennt man nicht. Man hat die Befragung durchgeführt, um diese Wahrscheinlichkeit genauer einzugrenzen.

In der Stichprobe ist Coca-Cola mit 62 Prozent beliebter als Pepsi. Der Binomialtest entscheidet, ob die Nullhypothese "In der Grundgesamtheit sind die beiden Getränke gleich beliebt" verworfen werden kann. Wenn die Nullhypothese stimmt, dann wäre der Unterschied von 12% (62%-50%) zwischen Stichprobe und Grundgesamtheit einfach ein Zufallsergebnis. Muss man vernünftigerweise damit rechnen, dass eine so grosse Differenz zwischen der Stichprobe und der Grundgesamtheit rein zufällig auftritt? Diese Frage beantwortet der P-Wert. Er gibt die Wahrscheinlichkeit an, dass in einer Stichprobe von 100 Jugendlichen aus einer Grundgesamtheit, in der die beiden Getränke gleich beliebt sind, 62 oder mehr Jugendliche das eine Getränk bevorzugen. Da im Beispiel der P-Wert mit rund zwei Prozent klein ist, ist der Zufall keine plausible Erklärung für die grössere Beliebtheit von Coca-Cola in der Stichprobe und man zieht daraus den Schluss, dass die beiden Getränke nicht gleich beliebt sind.

Die Alternativhypothese (alternative hypothesis) ist das Gegenteil der Nullhypothese. Sie lautet daher "Der Anteil der Personen in der Grundgesamtheit, die Coca-Cola bevorzugen, ist nicht gleich 50 Prozent". Wenn der P-Wert klein ist, dann hat man entweder durch einen unglücklichen Zufall eine "atypische" (nicht repräsentative) Stichprobe erwischt, oder die Nullhypothese ist falsch. Mit solch einem unglücklichen Zufall rechnen wir nicht: Wenn der P-Wert klein ist, wird üblicherweise die Nullhypothese verworfen und die Alternativhypothese angenommen. Was man noch als klein betrachtet, ist zu einem grossen Grad willkürlich. Diese Grenze wird als Signifikanzniveau bezeichnet. Üblich ist dafür der Wert von 5%.

Die Annahme, dass die Nullhypothese stimmt, ist eine Hilfskonstruktion. In einer Situation, in der man die Wahrheit nicht kennt, soll man aus Gründen der wissenschaftlichen Redlichkeit verschiedene mögliche Hypothesen in Betracht ziehen und für jede prüfen, ob sie mit den bekannten Fakten übereinstimmt oder im Widerspruch dazu steht. Die Person, welche die Befragung durchführt, vermutet meist, dass nicht die Null-, sondern die Alternativhypothese zutrifft. Fällt der P-Wert dann klein aus, ist diese Vermutung durch die Daten bestätigt.

Wenn Ihnen die Interpretation von P-Wert und Signifikanzniveau etwas Mühe macht, dann sind Sie nicht allein. Diese beiden Begriffe werden immer wieder falsch interpretiert. Oft liest oder hört man "Der P-Wert ist die Wahrscheinlichkeit, dass die Nullhypothese stimmt". Ob die Nullhypothese richtig oder falsch ist, hängt jedoch nicht von einem Zufallsmechanismus ab. Deshalb sollte man auch nicht von der Wahrscheinlichkeit, dass die Nullhypothese stimmt, sprechen. In der korrekten Definition des P-Werts respektive des Signifikanzniveaus wird davon ausgegangen, dass die Nullhypothese stimmt. Zufällig sind nur die Stichprobe und damit die beobachtete Abweichung von der Nullhypothese.

Im Beispiel "Coca-Cola oder Pepsi" ist oft nicht die Nullhypothese "Beide Getränke sind gleich beliebt" von Interesse, sondern vielmehr "Coca-Cola ist höchstens gleich beliebt wie Pepsi". Dann lautet die Alternativhypothese nämlich "Coca-Cola ist belieber als Pepsi", und das ist, was z.B. die Marketing-Abteilungvon Coca-Cola nachweisen möchte. Dann wird ein einseitiger Test (und entsprechend ein einseitiges Vertrauensintervall) benötigt. In *R* gibt man dafür den Befehl `binom.test(x=62,n=100,alternative="greater")` ein.

2.2 Vertrauensintervalle

Das von *R* berechnete Vertrauensintervall im Beispiel Coca-Cola gegen Pespi gibt an, dass der Anteil der Jugendlichen in der Grundgesamtheit, die Coca-Cola bevorzugen, mit 95-Prozent Sicherheit zwischen 51.7 und 71.5 Prozent liegt. Der Anteil von 62 Prozent in der Stichprobe liegt ungefähr in der Mitte des Intervalls. Der Wert von 50 Prozent liegt nicht im Vertrauensintervall, was mit der Testentscheidung übereinstimmt. Das Vertrauensintervall ist aber informativer als der Test: Es besagt nicht nur, dass der Wert von 50% vermutlich nicht zutrifft, sondern sagt auch, welche Werte von der Alternativhypothese plausibel sind. Es besteht eine grosse Unsicherheit von $62\% \pm 10\%$, und der Vorsprung von Coca-Cola könnte auch nur marginal sein.

Die Bedeutung der 95 Prozent Sicherheit ist die folgende: Wenn man wiederholt Zufallsstichproben vom Umfang 100 aus derselben Grundgesamtheit ziehen würde, dann würden sowohl der Anteil von Befragten, die Coca-Cola bevorzugen, als auch das Vertrauensintervall je nach Stichprobe unterschiedlich ausfallen. Im Fall von vielen Wiederholungen läge der Anteil von Coca-Cola-Liebhaberinnen und -Liebhabern in der Grundgesamtheit jedoch bei 19 von 20 Stichproben im Vertrauensintervall. Tatsächlich hat man nur eine einzige Stichprobe gezogen, aber man geht davon aus, dass man nicht durch einen unglücklichen Zufall eine "atypische" (nicht repräsentative) Stichprobe erwischt hat.

Beim Vertrauensintervall kann man für die Sicherheit auch einen anderen Wert als 95 Prozent wählen. Dazu verwendet man das optionale Argument `conf.level`. Für das 99-Prozent-Vertrauensintervall gibt man z.B. `binom.test(x=62,n=100,conf.level=0.99)` ein und erhält $[48.7\%, 74.2\%]$. Der *P*-Wert bleibt unverändert gleich 0.021, wenn das Argument `conf.level` verändert wird. Die Nullhypothese, dass beide Getränke gleich beliebt sind, wird somit nicht verworfen, wenn man als Bedingung für das Verwerfen einen *P*-Wert unter einem Prozent verlangt. Entsprechend liegt der Anteil von 50 Prozent Coca-Cola-Liebhaberinnen und -Liebhabern im 99%-Vertrauensintervall.

2.3 Zufallsstichproben

In einer Zufallsstichprobe muss jeder Merkmalsträger die gleiche Chance haben, in die Stichprobe aufgenommen zu werden, und zwar unabhängig davon, welche Merkmalsträger bereits ausgewählt wurden. Eine solche Zufallsstichprobe ist in der Praxis oft schwierig zu erhalten. Man benötigt dazu eine nummerierte Liste aller Merkmalsträger in der Population. Dann werden Zettel mit diesen Nummern gut gemischt und nacheinander blind gezogen. In *R* gibt es dafür die Funktion `sample`: Jeder Befehl von `sample(x=n,size=N)` ergibt eine neue Zufallsstichprobe vom Umfang n aus einer Grundgesamtheit mit N Merkmalsträgern.

Der Vorteil von Zufallsstichproben ist, dass sie systematische Verzerrungen der Ergebnisse vermeiden. In biologischen Untersuchungen sollte sichergestellt sein, dass sich die in der Stichprobe ausgewählten Tiere oder Pflanzen nicht gegenseitig beeinflussen oder konkurrenzieren. Bei Daten, die in wiederholten Messvorgängen gewonnen werden, ist die Grundgesamtheit ein abstraktes Konzept, nämlich die Menge aller möglichen Durchführungen des Messvorgangs. Hier ist es wichtig zu vermeiden, dass das Resultat einer Messung die nächste Messung beeinflusst. Bei einem chemischen Experiment können zum Beispiel Rückstände in den Reagenzgläsern vorhanden

sein, oder in einem Wachstumsexperiment können sich zwei Pflanzen im selben Topf konkurrenzieren. Ferner ist es wichtig dafür zu sorgen, dass die Bedingungen bei allen Wiederholungen gleich sind.

3 Im Buch besprochene Tests und Vertrauensintervalle

Je nach der Fragestellung und der Art, wie die Stichprobe ausgewählt wurde, kommt ein anderes Testverfahren zum Einsatz. Der Entscheidungsbaum in Abbildung 1 gibt Ihnen eine Übersicht über die Fragestellungen, die Arten von Stichproben und die passenden Testverfahren, die im Buch besprochen werden. Hier folgen ein paar Erläuterungen zu den Fragestellungen und den betrachteten Stichprobenarten, sowie kurze Erklärungen zur Durchführung der Tests in *R*.

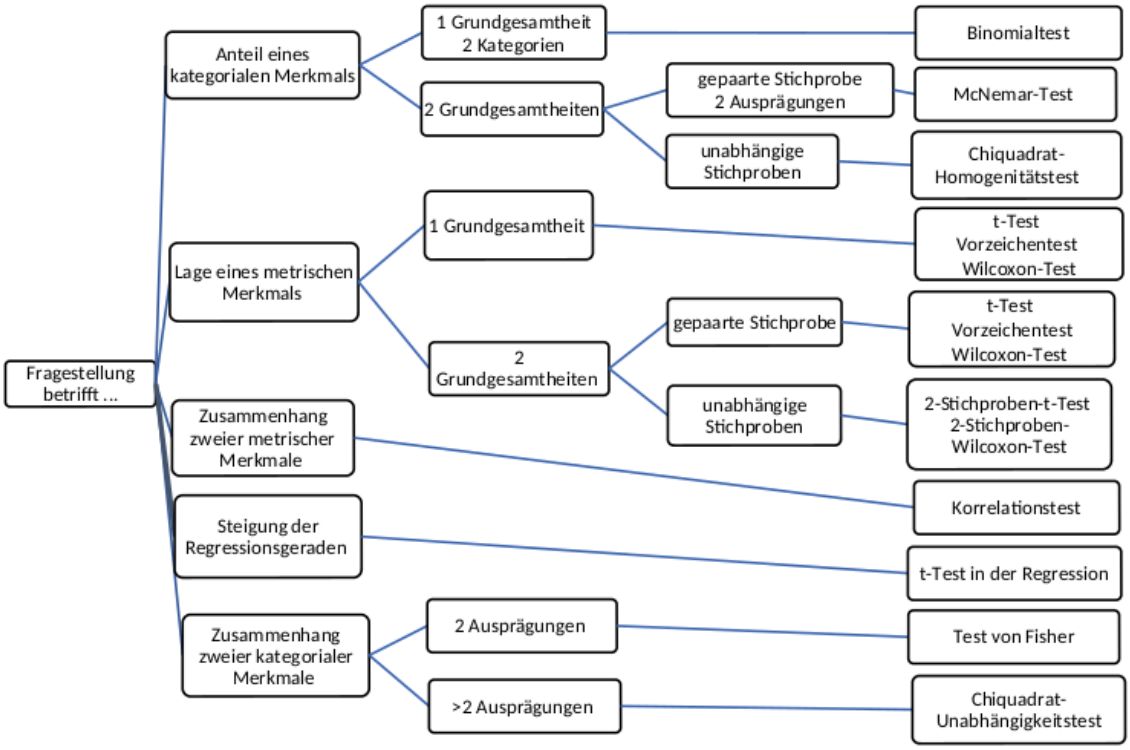

Abbildung 1: Entscheidungsbaum: Testverfahren on Abhängigkeit von Fragestellung, Merkmalsart und Stichprobenart

3.1 *t*-, Vorzeichen- und Wilcoxon-Test

Mit diesen Tests kann man Hypothesen über Lagemasse von metrischen Merkmalen testen, bzw. Vertrauensintervalle für ein solches Lagemass berechnen.

Hypothesen über den Erwartungswert μ werden mit dem ***t*-Test** getestet. In *R* wird dafür die Funktion `t.test` benützt. Die Nullhypothese wird mit dem Argument `mu` (englische Schreibweise für μ) festgelegt. Als Beispiel betrachten wir die oben eingegebenen 20 Körpergrössen und nehmen an, dass die Nullhypothese lautet "Die erwartete Grösse in der Grundgesamtheit ist gleich 170 cm". Dann sehen Befehl und Ausgabe wie folgt aus:

```
> t.test(x=groesse,mu=170)
One Sample t-test
data: groesse
t = 3.628, df = 19, p-value = 0.00179
alternative hypothesis: true mean is not equal to 170
95 percent confidence interval:
172.7291 180.1709
sample estimates: mean of x
176.45
```

Das 95-Prozent-Vertrauensintervall und der P-Wert sind direkt ablesbar: Der Mittelwert in der Grundgesamtheit liegt mit der Sicherheit von 95 Prozent zwischen 172.7 und 180.2 Zentimeter. Weil der P-Wert mit 0.18 Prozent sehr klein ist, ist es nicht plausibel, dass der Mittelwert der Grundgesamtheit gleich 170 Zentimeter ist. Die Nullhypothese wird daher verworfen. Der Output von t.test enthält zusätzlich die Angaben `t= 3.628` und `df=19`. Dabei steht `df` für Freiheitsgrade (engl. degrees of freedom) und ist gleich $n-1$. Auf die Erklärung der Grösse `t`, die dem Test den Namen gegeben hat, verzichten wir.

Um die Nullhypothese "Der Median in der Grundgesamtheit ist gleich μ" zu testen, verwendet man den **Vorzeichentest**. Dazu zählt man die Anzahl positiver Vorzeichen der Differenzen "Beobachteter Wert minus μ", d.h. die Anzahl Werte in der Stichprobe, die grösser als μ sind. Wenn die Nullhypothese stimmt, dann ist der Anteil der Werte in der Grundgesamtheit, die grösser als μ sind, gleich 50%. Damit liegt die gleiche Situation vor wie beim Binomialtest: Ein positives Vorzeichen wird als Erfolg bewertet, und man berechnet den P-Wert genau gleich wie beim Binomialtest. Es braucht also keine neue Funktion für den Vorzeichentest.

Im Beispiel der Körpergrössen sind 16 von 20 Werten grösser als 170, und `binom.test(16,20)` ergibt einen P-Wert von 1.2 Prozent. Also wird die Nullhypothese "Der Median der Körpergrössen in der Grundgesamtheit ist gleich 170 Zentimeter" auf dem 5%-Niveau verworfen. Allerdings erhält man so kein Vertrauensintervall für den Median in der Grundgesamtheit (Das Vertrauensintervall, das `binom.test` liefert, bezieht sich auf den Anteil der Werte in der Grundgesamtheit, die grösser als μ sind). Um mit R auch ein Vertrauensintervall für den Median zu berechnen, muss man das Zusatzpackage `DescTools` installieren und laden. Danach kann man die Funktion `SignTest` verwenden (sign, engl. für Vorzeichen):

```
> library("DescTools")
> SignTest(x=groesse,mu=170)
One-sample Sign-Test
data: groesse
S = 16, number of differences = 20, p-value = 0.01182
alternative hypothesis: true median is not equal to 170
95.9 percent confidence interval:
171 181
sample estimates:
median of the differences
178.5
```

Der P-Wert stimmt mit dem Wert überein, der mit binom.test berechnet wurde.

Ein dritter Test für ein Lagemass ist der **Wilcoxon-Test**. Mit ihm werden Hypothesen über den Pseudomedian getestet. Dieser liegt im Allgemeinen zwischen dem Median und dem Mittelwert. Eine Definition ersparen wir Ihnen hier. Wenn die Verteilung des Merkmals in der Grundgesamtheit symmetrisch bezüglich eines Werts ist, dann sind alle drei Lagemasse gleich und damit testen die drei Tests die gleiche Nullhypothese. Am Beispiel der Körpergrössen sieht der Wilcoxon-Test wie folgt aus:

```
> wilcox.test(x=groesse,mu=170,conf.int=TRUE)

Wilcoxon signed rank test with continuity correction
data: groesse
V = 181.5, p-value = 0.00449
alternative hypothesis: true location is not equal to 170
95 percent confidence interval:
173.0 180.5
sample estimates:
(pseudo)median
176.5
Warning messages:
1: In wilcox.test.default(groesse, mu = 170, conf.int = TRUE) :
cannot compute exact p-value with ties
2: In wilcox.test.default(groesse, mu = 170, conf.int = TRUE) :
cannot compute exact confidence interval with ties
```

Damit auch das Vertrauensintervall berechnet wird, muss man das optionale Argument `conf.int` auf `TRUE` setzen. Wenn in der Stichprobe Werte mehrfach auftreten (sogenannte Bindungen, engl. ties), dann können P-Wert und Vertrauensintervall nur näherungsweise berechnet werden und R liefert eine Warnung.

Wenn die Verteilung des Merkmals in der Grundgesamtheit symmetrisch ist, prüfen alle drei Tests dieselbe Nullhypothese. Welchen Test soll man dann verwenden? Beim t-Test ist der berechnete P-Wert nur dann exakt, wenn das Merkmal in der Grundgesamtheit normalverteilt ist. Dieser Nachteil fällt jedoch nicht gross ins Gewicht, weil der Fehler beim P-Wert meist klein ist. Ein grösserer Nachteil des t-Tests ist, dass sein Ergebnis sehr stark durch einzelne extreme Werte beeinflusst ist. Der Vorzeichentest hat den Nachteil, dass er die Grösse der Abweichungen von μ nicht berücksichtigt. Der Wilcoxon-Test stellt einen guten Kompromiss dar zwischen der Berücksichtigung von möglichst vielen Informationen in den Daten und der Unempfindlichkeit auf wenige extreme Werte.

Diese drei Tests können auch benutzt werden, um Hypopthesen über den Unterschied eines Lagemasses in zwei Grundgesamtheiten (z.B. Männer und Frauen, oder Patienten, welche zwei unterschiedliche medizinische Behandlungen erhalten haben) zu testen. Üblicherweise testet man dann die Nullhypothese, dass die beiden Lagemasse gleich sind, d.h. dass kein Unterschied besteht.

Wenn bei einer Untersuchung zwei Grundgesamtheiten verglichen werden sollen, muss man zunächst entscheiden, ob man zwei unabhängige Stichproben oder eine gepaarte Stichprobe erheben will. Bei zwei unabhängigen Stichproben darf die Auswahl der Merkmalsträger in der einen Grundgesamtheit keinen Bezug zur Auswahl in der anderen Grundgesamtheit haben. Bei einer gepaarten Stichprobe werden zuerst Paare von möglichst ähnlichen Merkmalsträgern aus den beiden Grundgesamtheiten gebildet und danach Paare gezogen, die anschliessend verglichen werden. Manchmal ist es möglich, Merkmalsträger mit sich selbst zu paaren, z.B. wenn man zwei medizinische Behandlungen bei der gleichen Person in einem zeitlichen Abstand durchführt.

In R können die Funktionen `t.test` und `wilcox.test` auch für den Vergleich von Lagemassen zweier Grundgesamtheiten benutzt werden: Die Eingabe hat dann die Form `t.test(x=..,y=..)` im ungepaarten und `t.test(x=...,y=..., paired =TRUE)` im gepaarten Fall, bzw. analog bei `wilcox.test`. Die Argumente `x` und `y` enthalten die Werte aus den beiden Grundgesamtheiten, und per default wird die Nullhypothese "Die Lagemasse in den beiden Grundgesamtheiten sind gleich" getestet. Der Vorzeichentest ist nur im gepaarten Fall definiert und wird mit dem `SignTest(x=...,y=...)` durchgeführt.

3.2 Korrelationstest, t-Test in der Regression

Die gebräuchlichste Masszahl für die Stärke und die Richtung des Zusammenhangs zweier metrischer Merkmale ist die Korrelation von Pearson, oft auch kurz Korrelationskoeffizient oder Korrelation genannt. In R kann man mit Hilfe der Funktion `cor.test` Vertrauensintervalle für die Korrelation berechnen und die Nullhypothese testen, dass diese in der Grundgesamtheit gleich Null ist. Die gleiche Funktion deckt auch die Rangkorrelation von Spearmann ab, welche weniger empfindlich ist auf einzelne extreme Werte.

Die Korrelation misst nur die Stärke und Richtung des linearen Zusammenhangs. Die Regressionsgerade beschreibt ebenfalls einen linearen Zusammenhang. Allerdings ist sie nicht symmetrisch in den beiden Merkmalen, sondern sie gibt an, wie man mit Hilfe einer Eingangsgrösse eine Zielgrösse vorhersagen kann. Die Funktion `lm` (für linear model) in R berechnet die Steigung und den Achsenabschnitt der Regressionsgeraden. Ausserdem wird der P-Wert für die Nullhypothese, dass die Steigung in der Grundgesamtheit gleich Null ist, sowie noch viele andere Resultate für weiterführende Analysen berechnet. Mit der Funktion `summary` werden die wichtigsten dieser Resultate ausgegeben. Im Buch wird als Beispiel eine Messreihe des schottischen Physikers Forbes verwendet mit dem Luftdruck als Eingangsgrösse und dem Siedepunkt von Wasser als Zielgrösse. Wenn die Werte in R eingelesen und als `forbes$Press`, bzw. `forbes$Temp` gespeichert sind, dann sieht der Aufruf und die Ausgabe wie folgt aus:

```
> summary(lm(forbes$Temp ~ forbes$Press))

Call:
lm(formula = forbes$Temp ~ forbes$Press)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6816 -0.1232  0.0429  0.1094  0.2833

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.498046   0.515187  132.96   <2e-16 ***
forbes$Press  0.031200   0.000603   51.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2467 on 15 degrees of freedom
Multiple R-squared:  0.9944,Adjusted R-squared:  0.9941
F-statistic:  2677 on 1 and 15 DF,  p-value: < 2.2e-16
```

Unter `Estimate` finden Sie in der Zeile `(Intercept)` den Achsenabschnitt und in der Zeile `forbes$Press` die Steigung der angepassten Geraden. Unter `Pr(>|t|)` ist der P-Wert für die Nullhypothese, dass die Steigung in der Grundgesamtheit gleich Null ist, angegeben. Der Wert `2e-16` bedeutet $2 \cdot 10^{-16}$ (`e` steht für Exponent). Somit gibt es keinen Zweifel daran, dass ein Zusammenhang zwischen den beiden Variablen besteht. Der Test ergibt immer den gleichen P-Wert wie der oben erwähnte Test `cor.test` für die Hypothese, dass die (Pearson-)Korrelation 0 ist, denn die Steigung der Regressionsgerade ist genau dann 0, wenn die Korrelation 0 ist. Daher sind diese beiden Tests äquivalent.

Interessanter als der Test der Nullhypothese ist die Frage, wie genau die Steigung bestimmt ist. Diese Information wird vom 95-Prozent-Vertrauensintervall geliefert, das mit der Funktion `confint` berechnet werden kann.

```
> confint(lm(forbes$Temp ~ forbes$Press))
                2.5 %     97.5 %
(Intercept)  67.39995097 69.59614176
forbes$Press  0.02991453 0.03248506
```

3.3 Chi-Quadrat-Test und Test von Fisher

Diese beiden Tests werden benutzt, um den Zusammenhang von zwei kategorialen Merkmalen zu untersuchen. Als Beispiel betrachten wir die oben erwähnten Daten mit dem Namen meals zu Essgewohnheiten von Schülerinnen und Schülern. Der Aufruf und die Ausgabe in *R* für den Chi-Quadrat-Test sehen wie folgt aus

```
> chisq.test(meals)
Pearson's Chi-squared test

data:  meals
X-squared = 4.898, df = 2, p-value = 0.08638
```

Da der *P*-Wert grösser als 0.05 ist, lässt sich mit diesen Daten kein Zusammenhang zwischen Geschlecht und Anzahl Mahlzeiten auf dem 5-Prozent-Signifikanzniveau nachweisen: Der Unterschied zwischen den beiden Geschlechtern kann noch durch Zufall erklärt werden. Die zum Chi-Quadrat-Test zugehörige Masszahl der Abhängigkeit ist der sogenannte Kontingenzkoeffizient, der allerdings nicht sehr verbreitet ist. Darum gibt *R* auch kein Vertrauensintervall an.

Wenn beide Merkmale nur zwei Ausprägungen haben, kann alternativ der Test von Fisher verwendet werden. Dieser liefert gleichzeitig ein Vertrauensintervall für eine andere Masszahl der Abhängigkeit, die sogenannte Odds ratio.

Im obigen Beispiel sind wir davon ausgegangen, dass die Sitchprobe aus einer aus Frauen und Männern bestehenden Grundgesamtheit ohne Berücksichtigung des Geschlechts gezogen wurde. Die Merkmale "Geschlecht" und "Anzahl Mahlzeiten" sind dann zufällig. Alternativ könnte man auch zwei unabhängige Stichproben aus zwei Grundgesamtheiten bestehend aus Männern, bzw. aus Frauen ziehen. Die Anzahl Frauen und Männer sind dann nicht zufällig, sondern von der Studienleitung festgelegt. Der Chi-Quadrat-Test kann in beiden Fällen verwendet werden. Will man zwischen den beiden Situationen differenzieren, dann spricht man vom Chi-Quadrat-Unabhängigkeitstest, bzw. vom Chi-Quadrat-Homogenitätstest.

3.4 McNemar-Test

Auch bei kategorialen Merkmalen kann man eine gepaarte Stichprobe verwenden, um Unterschiede zwischen zwei Grundgesamtheiten zu untersuchen. Im Unterschied zu metrischen Daten können bei kategorialen Daten jedoch keine Differenzen gebildet werden. Daher gibt es nur im Fall von zwei Ausprägungen einen einfachen Test für eine gepaarte Stichprobe, den McNemar-Test. Als Beispiel betrachten wir den Vergleich zweier Filme: Angenommen, 50 Schülerinnen und Schüler haben diese beiden Filme gesehen und als "gut" respektive "schlecht" eingestuft, und wir nehmen an, dass die Ergebnisse wie folgt aussehen:

	Film 1 gut	Film 1 schlecht
Film 2 gut	12	11
Film 2 schlecht	19	8

12+8=20 Schülerinnen und Schüler stufen also die beiden Filme als gleichwertig ein, 19 halten Film 1 für besser und elf halten Film 2 für besser. Der Test in *R* ergibt folgendes Resultat:

```
> mcnemar.test(x=film)

McNemar's Chi-squared test with continuity correction
data: film
McNemar's chi-squared = 1.6333, df = 1, p-value = 0.2012
```

Der P-Wert von 20 Prozent bedeutet, dass die Nullhypothese "Es gibt keine Präferenz für einen der beiden Filme" beibehalten wird.