

Einfache statistische Tests bei medizinischen Fragestellungen

Dr. Harald Heinzl

Medizinische Statistik und Informatik

Medizinische Universität Wien

harald.heinzl@meduniwien.ac.at

Version 2006-10

Zweiter Teil des Skriptums

zum Kurs

Biometrie I für MedizinerInnen

Inhaltsverzeichnis

Kapitel 3 Wahrscheinlichkeitsrechnung	4
3.1. Einführung	4
3.2. Wahrscheinlichkeitsrechnung	7
3.3. Übungen	28
Kapitel 4 Statistische Test I	31
4.1. Das Prinzip von statistischen Tests	31
4.2. t-Test	35
4.3. Wilcoxon Rangsummentest	41
4.4. Übungen	45
Kapitel 5 Statistische Tests II	46
5.1. Mehr zum t-Test bei unabhängigen Stichproben	46
5.2. Chi-Quadrat Test	52
5.3. Gepaarte Tests	57
5.4. Einseitige versus zweiseitige Tests	66
5.5. Übungen	68
Anhänge	72
D. Exakte Tests	72
E. Äquivalenzstudien	79
F. Beschreibung von statistischen Methoden in medizinischen Publikationen	81
G. Literaturverzeichnis	82

Vorwort

Im vorliegenden Skriptum wird für statistische Berechnungen zumeist auf *SPSS 14.0 für Windows, Version 14.0.1 (7 Dec 2005)*, zurückgegriffen (*Copyright © SPSS Inc., 1989-2005*).

Die im Skriptum verwendeten Datensätze können als ZIP-File über das Internet bezogen werden:

<http://www.muw.ac.at/user/harald.heinzl/>

Wenn Sie im Skriptum Fehler oder Ungereimtheiten entdecken, dann kontaktieren Sie mich bitte via E-mail:

harald.heinzl@meduniwien.ac.at

Besonderer Dank gebührt meinen Kollegen Andreas Gleiß und Georg Heinze, die mir nützliche Hinweise und wertvolle Vorschläge für diese Version des Skriptums gegeben haben.

Kapitel 3

Wahrscheinlichkeitsrechnung

3.1. Einführung

Im Seminar *Beschreibung und Visualisierung von medizinischen Daten* haben Sie ein breites Repertoire an Möglichkeiten kennen gelernt, um medizinische Daten darzustellen. Neben diversen graphischen Hilfsmitteln (wie Box-Plot, Histogramm, Streudiagramm, Balkendiagramm, Kreisdiagramm, usw.) wissen Sie auch mit verschiedenen statistischen Maßzahlen (wie Mittelwert, Standardabweichung, Median, Quartil, Quantil, Prozentangaben, usw.) umzugehen.

All diese statistischen Hilfsmittel dienen in der Medizin nur einem Zweck: Die Kommunikation mit Kollegen zu erleichtern. Sei es im direkten Gespräch, bei Vorträgen oder im Rahmen von Publikationen, immer geht es darum, das **Wesentliche** der erhobenen Daten **verständlich** und **korrekt** zu beschreiben.

Beispiel 3.1.1: (...) Von 34 Patienten erhielten 18 die Therapie ABC-NEW und 16 die Therapie XYZ. Mit ABC-NEW wurden 9 Heilerfolge (50 %) verzeichnet. Mit XYZ konnten dagegen nur 4 Heilerfolge (25 %) beobachtet werden. (...)

In Beispiel 3.1.1 wird wesentliches der erhobenen Daten verständlich und korrekt beschrieben.

Doch wann wird diese Information für andere interessant? Dass irgendwo irgendwer bei einer bestimmten Erkrankung unterschiedliche Therapien eingesetzt und dabei unterschiedliche Heilungserfolge erzielt hat, ist für uns vorerst völlig belanglos. Wann also beginnen derartige Informationen auch für uns interessant zu werden?

Wenn die Ergebnisse **verallgemeinerbar** sind! Und damit Vorhersagen für den Behandlungserfolg bei unseren Patienten, die an der selben Krankheit leiden, zulassen.

Doch wann sind auf Beobachtungen beruhende, empirische Ergebnisse verallgemeinerbar? Oder anders formuliert: *Wann können wir von einem Teil auf das Ganze schließen?* Hier sind wir nun an einem wichtigen, aber auch schwierigen Punkt angelangt. Daher sind vorerst ein paar grundlegende Gedanken notwendig.

Grundgesamtheit - Stichprobe

Nehmen wir an, wir hätten einen Behälter (z.B. Urne, Sack, etc.) mit sehr vielen Kugeln. Manche davon wären weiß, die anderen rot. Uns interessieren die Anteile der beiden Farben. Viele praktische Probleme sind diesem Problem ähnlich:

- Lieferung von 5000 Blutkonserven zur Gewinnung von Blutprodukten: Manche sind mit Hepatitis-B-Virus verseucht (rote Kugeln), manche nicht (weiße Kugeln).
- Lieferung von 1 Million Gulaschdosen für das Bundesheer: Manche sind verdorben (rote Kugeln), manche nicht (weiße Kugeln).
- Die Vielzahl an Patienten, die mit der Therapie XYZ behandelt werden könnten: Manche würden geheilt werden (rote Kugeln), manche nicht (weiße Kugeln).
- Münzwurf: Jede Kugel entspricht dem Wurf mit einer Münze, rote Kugeln könnten dann dem Ergebnis "Kopf", weiße Kugeln dem Ergebnis "Zahl" entsprechen. Beachte, dass es hier eine ungeheuer große (bzw. unendlich große) Menge an möglichen Münzwürfen gibt.

Das Gedankenmodell mit der Urne, die zweifarbige Kugeln enthält, entspricht einer sogenannten **binären Zielgröße**. Man kann dieses Modell natürlich sofort auf nominale Zielgrößen erweitern, indem man mehr Farben zulässt. Auch für ordinale oder metrische Zielgrößen kann man sich ähnliche Gedankenmodelle konstruieren.

Wichtig dabei ist die folgende Erkenntnis: Die Überprüfung der Güte von Gulaschdosen oder die Überprüfung der Wirksamkeit von klinischen Therapien oder der Münzwurf ist zwar inhaltlich nicht vergleichbar, formal gesehen besteht aber eine ähnliche Struktur, die durch das Urnenmodell repräsentiert werden kann.

Wie kann man den Anteil der roten Kugeln (und damit auch den der weißen) feststellen:

1. **Alle Überprüfen:** Dies wird bei den Blutkonserven üblicherweise die einzig denkbare Vorgangsweise sein.
2. **Nur ein Teil wird überprüft:** denkbar für Gulaschdosen, Therapie XYZ, Münzwurf
3. **Nichts wird überprüft (Wissen aus anderen Quellen wird übernommen):** denkbar für Blutkonserven, Gulaschdosen, Therapie XYZ, Münzwurf

Wir beschäftigen uns ab jetzt nur mehr mit der Variante 2, die Überprüfung eines Teiles (einer **Stichprobe**), um damit auf das Ganze (die **Grundgesamtheit**) zu schließen. Wie gehen wir dabei vor:

- I.) Entnehmen einer Stichprobe aus der Grundgesamtheit
- II.) Feststellen der benötigten Eigenschaften
- III.) Schließen auf Eigenschaften der Grundgesamtheit von den Ergebnissen der Stichprobe

ad I.) Das Entnehmen der Stichprobe entspricht dem Ziehen von Kugeln aus der Urne bzw. dem Auswählen von Gulaschdosen zur Qualitätsüberprüfung bzw. dem Rekrutieren von Patienten für eine klinische Studie bzw. dem Werfen der Münze. Entscheidend dabei ist die **Repräsentativität** der Stichprobe für die Grundgesamtheit. Repräsentativität wird üblicherweise durch Zufallsauswahl erreicht. Beachte dabei: Nur repräsentative Stichproben erlauben das **Verallgemeinern** der Ergebnisse auf die Grundgesamtheit! Nicht verallgemeinerbare Ergebnisse sind zumeist uninteressant.

ad II.) Das Feststellen der benötigten Eigenschaften ist oftmals schwierig. So wird zwar das Feststellen der Farbe der gezogenen Kugeln kein Problem verursachen. Verdorbenes Gulasch oder einen Behandlungserfolg festzustellen, kann insbesondere bei Grenzfällen zu Problemen führen. Genaue, widerspruchsfreie und verständliche Definitionen sind daher ein wesentlicher Teil eines guten Prüfplanes (bzw. Studienprotokolls).

ad III.) Wie kann man von Ergebnissen der **repräsentativen** Stichprobe auf Eigenschaften der unbekanntes Grundgesamtheit schließen? Dieses schwierige und sehr grundlegende Problem ist der Inhalt unserer Lehrveranstaltung.

Anmerkung: Wir werden das Problem teilweise sogar noch schwieriger gestalten. So werden wir unter anderem zwei jeweils repräsentative Stichproben betrachten, um zu entscheiden, ob sich Eigenschaften der beiden dahinterliegenden Grundgesamtheiten unterscheiden. Dies entspricht einer Situation mit zwei Urnen, aus denen jeweils eine bestimmte Anzahl an Kugeln gezogen wird, um zu entscheiden, ob der Anteil der roten Kugeln in den beiden Urnen verschieden ist. Im klinischen Bereich entspricht dies naturgemäß dem Therapievergleich: Können mit der Therapie ABC-NEW mehr Patienten geheilt werden als mit XYZ?

Zuerst wollen wir aber die Sache umdrehen. Wir wollen annehmen, dass wir die Grundgesamtheit kennen. Was passiert, wenn wir daraus eine zufällige Stichprobe ziehen? Das ist der Inhalt des nächsten Kapitels "Wahrscheinlichkeitsrechnung".

3.2. Wahrscheinlichkeitsrechnung

Ein Experiment mit nicht vorhersagbarem Ergebnis wird **ZUFALLSEXPERIMENT** genannt.

Damit kann, so hart dies auch klingen mag, die Anwendung einer klinischen Therapie als Zufallsexperiment bezeichnet werden. Denn der Ausgang (Z.B.: Kann der Patient dadurch geheilt werden? Wie lange überlebt der Patient? Wie stark wird der Blutdruck abgesenkt? ...) ist bei Therapiebeginn immer ungewiss.

Die Menge aller möglichen Ergebnisse eines Zufallsexperiments bilden den sogenannten **EREIGNISRAUM**. Davon ausgehend kann man die zufälligen **EREIGNISSE** definieren.

- Zufallsexperiment: Würfelwurf

Ereignisraum: $\{1, 2, 3, 4, 5, 6\}$

Ereignis A ... ein 2er wird gewürfelt, abgekürzt $A=\{2\}$

Ereignis B ... eine gerade Zahl wird gewürfelt, $B=\{2,4,6\}$

Ereignis C ... eine Zahl größer als 3 wird gewürfelt, $C=\{4,5,6\}$

- Zufallsexperiment: Meinungsbefragung

Ereignisraum: {Genuss- und Nahrungsmittelpräferenzen aller ÖsterreicherInnen über 14 Jahren}

Ereignis D ... Raucher

Ereignis E ... Person mag lieber Margarine als Butter aufs Brot

Anmerkung: So simpel diese beiden Ereignisse auf den ersten Blick erscheinen, so ungenau sind sie definiert. Was ist beispielsweise ein Raucher? Genügt am Sonntag nach dem Mittagessen eine Zigarette? Ist jemand der seit 3 Tagen das Rauchen aufgegeben hat, als Nichtraucher zu werten?

Noch unklarer ist das Ereignis E? Was ist mit einer Person, die weder Butter noch Margarine auf Brot will? Oder mit einer Person, die zwar Margarine lieber als Butter verzehrt, die aber Brot verabscheut? Was ist überhaupt Margarine, ist damit eine bestimmte Marke oder das Produkt im allgemeinen gemeint?

- Zufallsexperiment: Überlebenszeit von Patienten nach Chemotherapie bei Ewing-Knochensarkom

Ereignisraum: {alle reellen Zahlen im Intervall von 0 bis 130 Jahren}

Ereignis F ... Patient verstirbt innerhalb des ersten Jahres

Ereignis G ... Patient lebt länger als 15 Jahre

- Zufallsexperiment: Dauer des Schlafs innerhalb von 24 h
Ereignisraum: { alle reellen Zahlen im Intervall 0 bis 24 h }
Ereignis H ... weniger als 8 h Schlaf, H wäre dann das Intervall $[0,8)$
Ereignis I abnormale Schlafdauer, das sind weniger als 4 h und mehr als 11 h, I wäre dann aus den Intervallen $[0,4)$ und $(11,24]$ zusammengesetzt

KOMPLEMENTÄRE REIGNIS: Besteht aus den Elementen des Ereignisraumes, die nicht im betrachteten Ereignis liegen.

- Komplementäres Ereignis zu B: eine ungerade Zahl wird gewürfelt, abgekürzt $B^C = \{1,3,5\}$
- Komplementäres Ereignis zu E, abgekürzt E^C : Alle Österreicher über 14 Jahren, die nicht lieber Margarine als Butter aufs Brot mögen(!?)

SICHERES REIGNIS, z.B. eine Zahl kleiner-gleich 6 wird gewürfelt,
 $S = \{1,2,3,4,5,6\}$

UNMÖGLICHES REIGNIS, z.B. eine Zahl zwischen 19 und 25 wird gewürfelt, $U = \{\}$

VEREINIGUNG VON 2 REIGNISSEN: Mindestens eines der beiden Ereignisse muss eintreten

- B vereinigt mit C, d.h. die gewürfelt Zahl ist gerade, oder größer als 3 oder beides
 $B \cup C = \{2,4,5,6\}$
- das Ereignis I ... abnormale Schlafdauer, ist bereits ein zusammengesetztes Ereignis: $I = K \cup L$
Ereignis K ... zu kurzer Schlaf, $K = [0,4)$
Ereignis L ... zu langer Schlaf, $L = (11,24]$

DURCHSCHNITT VON 2 EREIGNISSEN: Beide Ereignisse müssen eintreten

- B geschnitten mit C, d.h. die gewürfelte Zahl ist gerade und gleichzeitig größer als 3
 $B \cap C = \{4, 6\}$
- Was ist der Durchschnitt der Ereignisse K und L? Das **unmögliche Ereignis**, denn man kann nicht gleichzeitig zu kurz und zu lange schlafen. Man sagt die Ereignisse K und L *schließen sich gegenseitig aus* bzw. sie *sind disjunkt*.
- D^c geschnitten mit I^c : Alle Österreicher über 14 Jahren, die nicht rauchen und normale Schlafdauer (zw. 4 h und 11 h) aufweisen.

Nach der Definition der Ereignisse steht dem Berechnen von Wahrscheinlichkeiten nichts mehr im Weg.

Beispiel 3.2.1: Die Wahrscheinlichkeit, mit einem fairen Würfel einen "2er" zu werfen, ist $1/6$.

Beispiel 3.2.2: Die Wahrscheinlichkeit für eine Knabengeburt ist 51.4 %.

Bei beiden Beispielen gibt es eine enge Beziehung zum Begriff der "relativen Häufigkeit".

ad Beispiel 3.2.1: Aus physikalisch-theoretischen Überlegungen weiß man, beim fairen Würfel ist jede Seite gleichwahrscheinlich. Wir nehmen daher an, dass bei häufigem Würfeln die relative Häufigkeit für einen "2er" gegen den mathematisch errechneten Wert von $1/6$ streben wird. Die Berechnung dieser *mathematischen Wahrscheinlichkeit* erfolgte durch die Formel:

$$\frac{\text{Anzahl der günstigen Fälle}}{\text{Anzahl der möglichen Fälle}} = \frac{g}{m} = \frac{1}{6}$$

Diese Formel wird auch Laplace-Wahrscheinlichkeit (nach Pierre Simon Marquis de Laplace, 1749-1827) genannt. Sie ist nur sinnvoll, wenn alle möglichen Fälle gleichwahrscheinlich sind. Daher können wir auch den fairen Münzwurf mit diesem Schema berechnen.

ad Beispiel 3.2.2: Diese Aussage ist durch Beobachtung entstanden. Nachdem man lange Jahre beobachtet hat, dass die relative Häufigkeit für eine Knabengeburt bei 51.4 % liegt, nimmt man an, dass dies auch zukünftig so sein wird. Diese Wahrscheinlichkeit ist demnach eine *statistische Wahrscheinlichkeit*: die relative Häufigkeit in einer sehr, sehr großen Versuchsserie.

Unterschied zwischen relativer Häufigkeit und Wahrscheinlichkeit:

- relative Häufigkeit: Zustand in einer Stichprobe
- Wahrscheinlichkeit: Zustand in der Grundgesamtheit, Bezug auf ein zukünftiges *nicht vorhersagbares* Ereignis
- Wahrscheinlichkeit ist *Erwartungswert* einer relativen Häufigkeit

Beachten Sie: Wahrscheinlichkeiten sind immer mit Ereignissen verbunden. Wenn eine Wahrscheinlichkeit kommuniziert wird, dann ist sicherzustellen, dass das entsprechende Ereignis eindeutig und verständlich definiert ist. Die folgenden Beispiele machen das deutlich:

- (Gigerenzer, 2002): Ein Psychiater verschrieb depressiven Patienten regelmäßig das Arzneimittel Prozac. Er informierte die betreffenden Patienten darüber, dass mit einer Wahrscheinlichkeit von 30 bis 50 Prozent sexuelle Probleme auftreten würden. Der Psychiater meinte, dass bei 3 bis 5 von 10 Patienten sexuelle Probleme auftreten würden. Viele Patienten dachten aber, in 30 bis 50 Prozent ihrer sexuellen Aktivitäten würden sich Störungen einstellen.
- (v. Randow, 1992): Nach dem Golfkrieg von 1991 meldete die US-Navy, sie habe ihre Cruise Missiles mit 99-prozentigem Erfolg abgefeuert. Nachfragen ergaben, dass man damit meinte, es seien 99 Prozent der Raketen problemlos gestartet worden. Es handelte sich um keine Angabe über die Trefferquote.
- Im klinischen Bereich ist die Variation von Definitionen über Zeit bzw. zwischen Zentren oftmals ein großes Problem, da "*stage migration*" (auch "*Will Rogers phenomenon*" genannt) auftreten kann. Ein Beispiel dafür wird in Feinstein et al. (1985) beschrieben: Bedingt durch bessere diagnostische Verfahren konnten vormals nicht entdeckbare Metastasen bei Lungenkrebspatienten identifiziert werden. Diese Patienten wurden nun anstatt wie bisher dem Stadium "gut" dem "Stadium "schlecht" zugewiesen. In Folge stiegen die Überlebenschancen in beiden Stadien an, obwohl sich im individuellen Ergebnis nichts geändert hatte. Der Grund lag einfach in der Tatsache, dass die bisher schlechten im Stadium "gut" zu den guten im Stadium "schlecht" wurden.

Mathematisch genügen drei Eigenschaften, um das Rechnen mit **Wahrscheinlichkeiten** eindeutig festzulegen:

- I.) Die Wahrscheinlichkeit eines Ereignisses ist eine Zahl zwischen 0 und 1 (0 und 100 Prozent).
- II.) Das sichere Ereignis hat Wahrscheinlichkeit 1 (100 Prozent).
- III.) Wahrscheinlichkeiten von zwei sich gegenseitig ausschließenden Ereignissen können addiert werden.

Diese Eigenschaften werden auch *Kolmogoroffsche Axiome* genannt, wobei - streng genommen - Kolmogoroff das dritte Axiom mathematisch etwas komplizierter angeschrieben hat.

ad I. und II.) Demnach ist eine Wahrscheinlichkeit nichts anderes als ein Maß, dass das mögliche Eintreten eines Ereignisses quantifiziert. Da das sichere Ereignis immer eintritt, hat es auch die höchstmögliche Wahrscheinlichkeit, nämlich 1. Der Zahlenbereich zwischen 0 und 1 wird übrigens aus Bequemlichkeitsgründen verwendet, wir könnten genausogut irgendeinen völlig beliebigen Bereich verwenden (z.B. zwischen -94.23 und $+2490.08$), wir wären aber sehr ungeschickt, wenn wir so verfahren würden.

ad III.) Dies ist eine einfache, völlig plausible Rechenregel. Dazu als Beispiel das Würfeln mit einem fairen Würfel: Wenn die Wahrscheinlichkeit einen 2er zu würfeln gleich $1/6$ ist, und wenn die Wahrscheinlichkeit einen 3er zu würfeln ebenfalls gleich $1/6$ ist, dann ist die Wahrscheinlichkeit einen 2er oder einen 3er zu würfeln, $1/6$ plus $1/6$ gleich $2/6$. Diese beiden Ereignisse schliessen sich gegenseitig aus, denn wenn wir einen 2er würfeln, dann haben wir keinen 3er gewürfelt, und umgekehrt.

ad I.-III.) Alle Rechenregeln, die Wahrscheinlichkeiten betreffen, lassen sich aus diesen drei Eigenschaften ableiten.

Der Buchstabe "P" (engl. probability, von lat. probare - beglaubigen) wird gerne zur Bezeichnung bzw. Abkürzung von Wahrscheinlichkeiten verwendet. Dazu ein Tipp: Wenn man als Anfänger Wahrscheinlichkeiten ausrechnen will, dann ist es sinnvoll, sich einer Notation in ganzen Sätzen zu bedienen. Zum Beispiel: "Die Wahrscheinlichkeit für eine Knabengeburt ist 0.514 oder 51.4 Prozent." Mit der Zeit wird dies sehr umständlich, und man beginnt ganz automatisch nach einer Abkürzung zu suchen. Diese könnte hier wie folgt aussehen: " $P(K)=0.514$ "

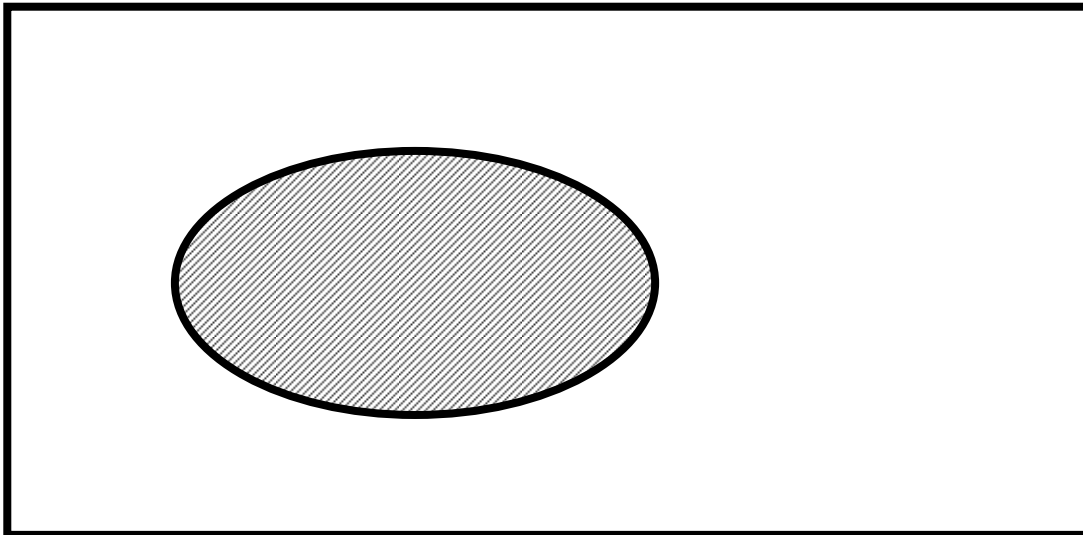
Wir werden im folgenden meist die abgekürzte Notation verwenden.

Wahrscheinlichkeit für das komplementäre Ereignis:

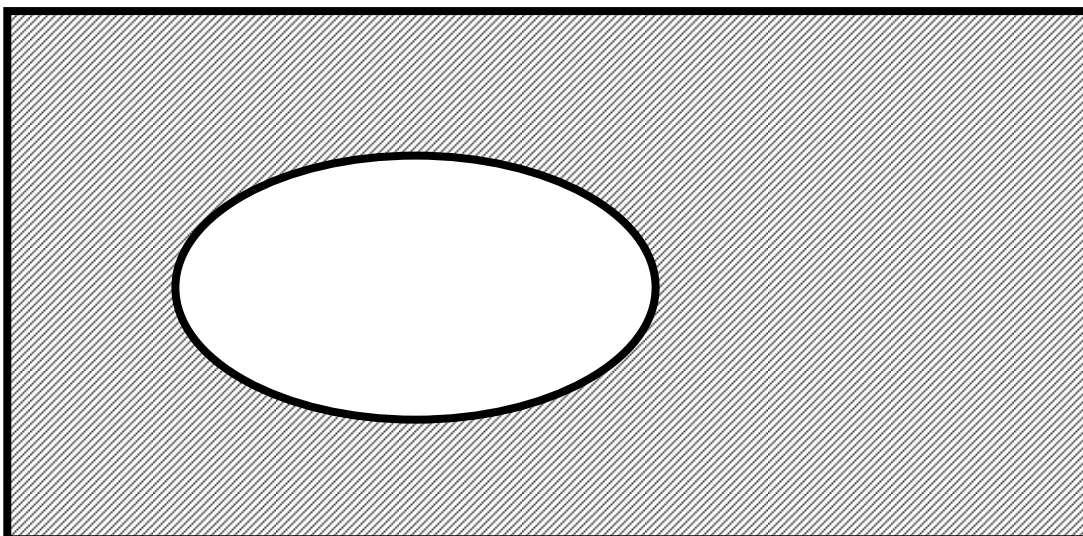
Angenommen, die Wahrscheinlichkeit für ein Ereignis ist bekannt. Eins minus dieser Wahrscheinlichkeit ergibt dann die Wahrscheinlichkeit für das Komplementärereignis:

$$P(X^c) = 1 - P(X)$$

Um diese Wahrscheinlichkeitsberechnung zu veranschaulichen, kann man ein Venn-Diagramm verwenden. Dazu wird das sichere Ereignis als Rechteck gezeichnet. Ereignisse werden als Ellipsen eingezeichnet. In unserem Fall würde also das Ereignis X der Fläche der grau markierten Ellipse entsprechen. Die Wahrscheinlichkeit von X wäre der Anteil der Ellipsenfläche an der Rechteckfläche.



Die Fläche ausserhalb der Ellipse wäre nun das Komplementärerereignis.



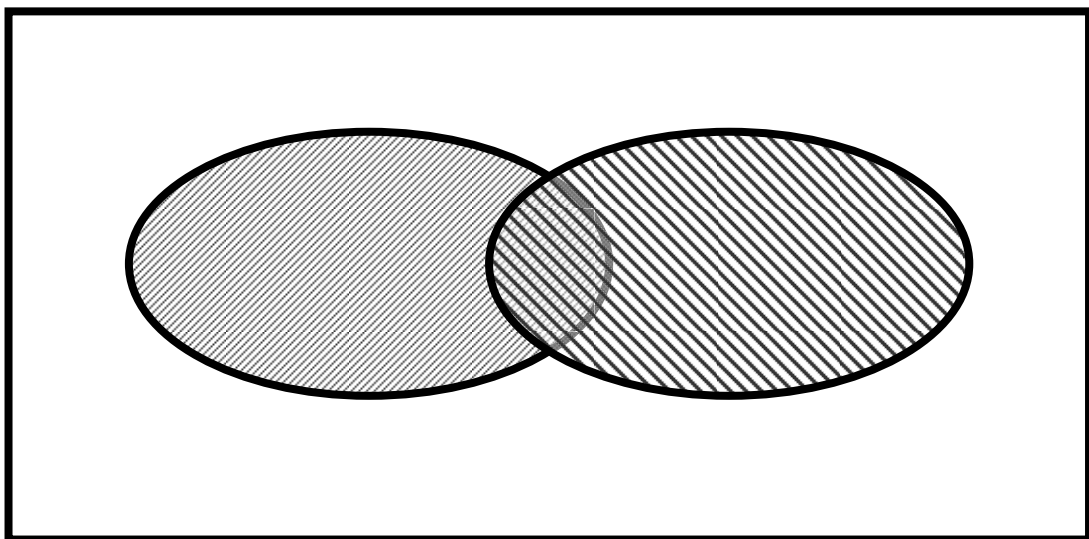
Beispiel: Wahrscheinlichkeit für eine Knabengeburt ist 0.514, d.h. die Wahrscheinlichkeit für eine Mädchengeburt ist daher $1-0.514=0.486$ oder 48.6 %.

Wahrscheinlichkeit für zwei Ereignisse, die sich nicht gegenseitig ausschließen:

Addiere die beiden Wahrscheinlichkeiten, aber ziehe die Wahrscheinlichkeit des Durchschnitts ab, denn dieser würde ansonsten einmal zu oft gezählt werden:

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

Dazu wieder ein Venn-Diagramm:



Beispiel: Wahrscheinlichkeit, mit einem fairen Würfel eine gerade Zahl oder eine Zahl größer als 3 zu würfeln

$$P(B \cup C) = P(B) + P(C) - P(B \cap C) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6}$$

BEDINGTE WAHRSCHEINLICHKEITEN

Das ist ein zentraler Begriff aus unserem täglichen Leben. Er spiegelt die Tatsache wider, dass sich unsere Einschätzung von Sachverhalten ändern kann, sobald wir zusätzliche Informationen erhalten. Zum Beispiel werden wir die Möglichkeit einer Mehrlingsgeburt als relativ selten ansehen. Diese Einschätzung wird sich ändern, wenn wir erfahren, dass die Schwangere eine Hormonbehandlung hinter sich hat.

Eine *bedingte Wahrscheinlichkeit* ist die Wahrscheinlichkeit für ein bestimmtes Ereignis, unter dem Wissen, dass ein anderes Ereignis bereits eingetreten ist. Wir schreiben:

$P(X|Y)$... die bedingte Wahrscheinlichkeit des Ereignisses X, gegeben wir wissen, dass das Ereignis Y bereits eingetreten ist

Mit dem Venn-Diagramm können wir uns verdeutlichen, dass

$$P(X|Y) = P(X \cap Y) / P(Y) \text{ bzw. umgeformt } P(X \cap Y) = P(Y)P(X|Y)$$

(Platz für Zeichnung)

Beispiel: In Diagnosestudien wird untersucht, ob sich klinische, radiologische oder Laborprüfverfahren zur Diagnose von bestimmten Krankheiten eignen. Begriffe wie Sensitivität und Spezifität werden verwendet. Dabei handelt es sich um bedingte Wahrscheinlichkeiten:

Sensitivität = Wahrscheinlichkeit dafür, dass der diagnostische Test positiv ist, gegeben es liegt tatsächlich ein Krankheitsfall vor

Zwei Ereignisse sind dabei im Spiel:
Ereignis1 = { diagnostischer Test positiv }
Ereignis2 = { Krankheit liegt vor }

$$\text{Sensitivität} = P(\text{Ereignis1} | \text{Ereignis2})$$

Spezifität = Wahrscheinlichkeit dafür, dass der diagnostische Test negativ ist, gegeben es liegt keine Erkrankung vor

Hier sind die Komplementärereignisse gefragt:
Ereignis1^c = { diagnostischer Test negativ }
Ereignis2^c = { Krankheit liegt nicht vor }

$$\text{Spezifität} = P(\text{Ereignis1}^c | \text{Ereignis2}^c)$$

In der medizinischen Alltagssituation ist der positive Vorhersagewert von großer Bedeutung, d.h., die Wahrscheinlichkeit dafür, dass eine Erkrankung vorliegt, wenn der diagnostische Test ein positives Ergebnis zeigt. Analog dazu kann man auch den negativen Vorhersagewert definieren.

UNABHÄNGIGKEIT VON EREIGNISSEN

Dies ist ein weiterer zentraler Begriff aus unserem täglichen Leben. Etwas salopp formuliert bedeutet Unabhängigkeit von zwei Ereignissen, dass diese beiden Ereignisse sich nicht gegenseitig beeinflussen. Das bedeutet damit auch, dass das Eintreten des einen Ereignisses die Wahrscheinlichkeit für das andere Ereignis nicht ändert.

So ist beispielsweise das Würfeln einer 2 unabhängig davon, ob davor eine Zahl größer als oder kleiner-gleich 3 gewürfelt wurde.

Oder wenn Sie in einem Fachjournal lesen, dass bei einer bestimmten Erkrankung die Verabreichung von Kräuter-Tropfen und der darauffolgende Heilungserfolg statistisch unabhängige Ereignisse wären, dann ist dies nur eine noble Umschreibung für den Umstand, dass die Kräuter-Tropfen bei der betreffenden Erkrankung nicht wirken.

Ereignis X und Y sind unabhängig,

- wenn $P(X|Y)=P(X)$
- äquivalent dazu: $P(X|Y)=P(X|Y^c)$
- äquivalent dazu: $P(X \cap Y)=P(X)P(Y)$... Multiplikationssatz

Das Gegenteil von Unabhängigkeit ist Abhängigkeit. So ist z.B. das Ereignis "Mehrlingsgeburt" vom Ereignis "Hormonbehandlung" abhängig.

Anmerkung: Bei Abhängigkeit von Ereignissen können wir zwischen "stochastischer" und "kausaler" Abhängigkeit unterscheiden. Stochastische Abhängigkeit ist symmetrisch, kausale Abhängigkeit geht stets in eine Richtung. So sind Mehrlingsgeburten offenbar von der Hormonbehandlung kausal abhängig, denn eine Umkehrung wäre sinnlos! Der Begriff kausale Abhängigkeit bedeutet hier, dass das Auftreten der Ursache die Wahrscheinlichkeit für die Wirkung ändert. Es bedeutet aber nicht notwendigerweise, dass die Wirkung zwingend aus der Ursache folgt.

Viele beobachtete Abhängigkeiten bleiben vorerst stochastisch, einfach deshalb, weil das aktuell vorhandene substanzwissenschaftliche Wissen nicht ausreicht. Durch spätere Einsichten gelingt es aber oft, daraus kausale Abhängigkeiten zu entwickeln.

Beim Berechnen von bedingten Wahrscheinlichkeiten müssen Sie auf kausale Abhängigkeiten keine Rücksicht nehmen. Es ist sowohl das Berechnen von $P(\text{Wirkung}|\text{Ursache})$ als auch $P(\text{Ursache}|\text{Wirkung})$ "erlaubt". Letzteres, also die Frage nach der Wahrscheinlichkeit für eine Ursache, gegeben man hat eine bestimmte Wirkung beobachtet, ist übrigens ganz typisch für Detektive, Historiker und Gerichtsmediziner.

Beispiel 3.2.3: Wie groß ist die *Wahrscheinlichkeit mit einem fairen Würfel beim Mensch-ärgere-Dich-nicht dreimal hintereinander keinen "6er" zu würfeln (nicht "ansetzen" zu dürfen)?*

Vorerst definieren wir drei Ereignisse:

E1 ... keinen "6er" beim ersten Mal würfeln

E2 ... keinen "6er" beim zweiten Mal würfeln

E3 ... keinen "6er" beim dritten Mal würfeln

Wir wissen, die einzelnen Würfe sind unabhängig. Daher können wir den Multiplikationssatz (vorige Seite) verwenden:

$$P(\text{dreimal hintereinander keinen "6er" zu würfeln}) \\ = P(E1 \cap E2 \cap E3) = P(E1)P(E2)P(E3)$$

Jetzt brauchen wir nur noch die Wahrscheinlichkeit für beim einmaligen Würfeln keinen "6er" zu würfeln. Das ist $\frac{5}{6}$.

$$P(\text{dreimal hintereinander keinen "6er" zu würfeln}) = \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = \frac{125}{216} = 0.5787.$$

Beispiel 3.2.4: *Infektionskrankheiten können auch "stumm" verlaufen. Angenommen, die Wahrscheinlichkeit für den stummen Verlauf einer bestimmten Infektion liegt bei 40 %. Zwei Personen sind infiziert. Wie groß ist die Wahrscheinlichkeit für*

- a.) *zwei stumme Verläufe,*
- b.) *zwei offene Verläufe,*
- c.) *genau einen stummen Verlauf?*

- a.) $P(\text{beide stumm}) = 0.4 \times 0.4 = 0.16$
- b.) $P(\text{beide offen}) = 0.6 \times 0.6 = 0.36$
- c.) $P(\text{genau einer stumm}) = 0.4 \times 0.6 + 0.6 \times 0.4 = \underline{0.48}$
 $\underline{1.00}$

ad c.) Das Ereignis "genau einer stumm" besteht aus zwei sich gegenseitig ausschliessenden Ereignissen. Und zwar aus "erste Person stumm und zweite Person offen" und "erste Person offen und zweite Person stumm".

Beispiel 3.2.5: Jedes Individuum hat zwei Ausgaben von jedem Gen (je eins von Mutter und Vater). Jedes Gen kann in unterschiedlichen Formen (Allelen) auftreten. Zum Beispiel gibt es drei Hauptallele (A,B und O) des ABO-Gens zur Blutgruppenbestimmung. Die relativen Populationshäufigkeiten von unterschiedlichen Allelen eines Gens werden Allelhäufigkeiten (Genhäufigkeiten) genannt, wobei jedes Individuum zwei Allele zum Populationsgenpool beiträgt. In einer kaukasischen Bevölkerung betragen die Genhäufigkeiten für die drei Blutgruppenallele:

$$P(\text{Allel A})=0.28, P(\text{Allel B})=0.06, P(\text{Allel O})=0.66$$

Das Allel-Paar eines Individuums bildet den Genotyp des Individuums. Bei den Blutgruppen gibt es 6 davon, nämlich AA, AB, AO, BB, BO, OO.

Beachte: BA, OA, OB sind dasselbe wie AB, AO bzw. BO.

Die Blutgruppen-Phänotypen (Typ A, Typ B, Typ AB, Typ O) entstehen aus den Genotypen, wobei bestimmte Penetranz-Relationen gelten. In unserem Fall ist A und B dominant über O, während A und B beide kodominant sind. Somit entsteht Phänotyp A aus den Genotypen AA und AO, Phänotyp B aus den Genotypen BB und BO, Phänotyp AB aus dem Genotyp AB, und Phänotyp O aus dem Genotyp OO.

Durch die Bildung von Gameten überträgt jeder Elternteil eines seiner zwei Allele mit Wahrscheinlichkeit 1/2. Dadurch wird der Genotyp des Kindes konstituiert.

In einer großen Population

- mit "random mating" (jede/jeder hat gleiche Chance auf Paarung mit jedem/jeder),
- ohne Migration
- ohne Mutation
- und ohne Selektion (Einfluss auf Fertilität des Individuums und Lebensfähigkeit seines Nachwuchs),

sind die Allelhäufigkeiten und die Genotyphäufigkeiten konstant über die Generationen. So eine Population befindet sich im sogenannten Hardy-Weinberg-Gleichgewicht, und die Genotyphäufigkeiten lassen sich dann sehr einfach aus den Allelhäufigkeiten errechnen. (Frage: Warum wohl?)

Angenommen, die Voraussetzungen für das Hardy-Weinberg-Gleichgewicht gelten, wie sind die einzelnen Genotypwahrscheinlichkeiten bei den Blutgruppen in einer kaukasischen Bevölkerung?

$$P(\text{Genotyp AA}) = P(\text{Allel A}) \times P(\text{Allel A}) = 0.0784$$

$$P(\text{Genotyp AB}) = 2 \times P(\text{Allel A}) \times P(\text{Allel B}) = 0.0336$$

$$P(\text{Genotyp AO}) = 2 \times P(\text{Allel A}) \times P(\text{Allel O}) = 0.3696$$

$$P(\text{Genotyp BB}) = P(\text{Allel B}) \times P(\text{Allel B}) = 0.0036$$

$$P(\text{Genotyp BO}) = 2 \times P(\text{Allel B}) \times P(\text{Allel O}) = 0.0792$$

$$P(\text{Genotyp OO}) = P(\text{Allel O}) \times P(\text{Allel O}) = \underline{0.4356}$$

$$\underline{1.0000}$$

ZIEHEN MIT UND OHNE ZURÜCKLEGEN

Wie groß ist die Wahrscheinlichkeit, beim Lotto 6 aus 45 tatsächlich 6 Richtige zu tippen?

Dazu eine Denkanleitung: Im Trichter liegen 45 Kugeln, 6 davon sind rot (diese entsprechen den von uns getippten Zahlen), die restlichen 39 Kugeln sind weiß. Wie groß ist die Wahrscheinlichkeit beim 6-maligen Ziehen 6-mal eine rote Kugel zu ziehen? Entscheidend dabei ist, dass die Kugeln nach dem Ziehen **nicht** mehr **zurückgelegt** werden.

Beim ersten Zug ist die Wahrscheinlichkeit für rot, $P(R_1) = 6/45 = 0.1333$.

Beim zweiten Zug hängt die Wahrscheinlichkeit für rot davon ab, ob beim ersten Ziehen rot oder weiß gezogen wurde, $P(R_2|R_1) = 5/44 = 0.1136$ und $P(R_2|R_1^c) = 6/44 = 0.1364$. Das heißt aber, Unabhängigkeit ist nicht mehr gegeben!

Die gleiche Argumentation gilt auch für die Wahrscheinlichkeit beim dritten, vierten, fünften und sechsten Ziehen. Die Wahrscheinlichkeit für einen Lottosechser, also 6 rote Kugeln hintereinander aus dem Trichter zu ziehen, ist daher gar nicht mehr so einfach berechenbar. Es ergibt sich schließlich:

$$\begin{aligned} P(\text{Lottosechser}) &= P(R_1) \times P(R_2|R_1) \times P(R_3|R_1R_2) \times P(R_4|R_1R_2R_3) \\ &\times P(R_5|R_1R_2R_3R_4) \times P(R_6|R_1R_2R_3R_4R_5) \\ &= \frac{6}{45} \times \frac{5}{44} \times \frac{4}{43} \times \frac{3}{42} \times \frac{2}{41} \times \frac{1}{40} = \underline{\underline{0.000000123}} \end{aligned}$$

Anmerkung: Wenn bei einer großen Grundgesamtheit ohne Zurücklegen gezogen wird, dann sind die entstehenden Auswirkungen auf die Wahrscheinlichkeiten für das nächste Experiment üblicherweise so gering, dass dies oftmals für Berechnungen ignoriert wird. Zum Beispiel: 5000 Kugeln, 1000 davon rot. Beim ersten Zug ist die Wahrscheinlichkeit für eine rote Kugel, $P(R) = 1000/5000 = 0.2$. Die Wahrscheinlichkeit für eine rote Kugel beim zweiten Zug ohne Zurücklegen der ersten Kugel ist nun entweder $P(R|R) = 999/4999 = 0.19984$ oder $P(R|W) = 1000/4999 = 0.20004$. Hier könnte man also getrost mit dem gerundeten Wert von 0.2 arbeiten, was der Wahrscheinlichkeit des Ziehens **mit** Zurücklegen entspricht.

WAHRSCHEINLICHKEITSVERTEILUNGEN

Es gibt eine Reihe von inhaltlich völlig unterschiedlichen Problemen, die aber bei der Berechnung von Wahrscheinlichkeiten einem einheitlichen Schema entsprechen. Hier sind drei Beispiele dazu:

- Eine Anzahl n von nicht verwandten Personen erleidet eine Infektion. Die Wahrscheinlichkeit p für den stummen Verlauf dieser Infektion ist bekannt. Wie groß ist die Wahrscheinlichkeit, dass bei genau x (bzw. bei höchstens x) Personen diese Infektion stumm verläuft?
- Eine erfahrene Chirurgin führt einen Routineeingriff bei n nicht verwandten Patienten an jeweils verschiedenen Tagen durch. Die Wahrscheinlichkeit p für eine Komplikation während des Routineeingriffs ist bekannt. Wie groß ist die Wahrscheinlichkeit, dass bei genau x (bzw. bei höchstens x) Patienten eine Komplikation auftritt?
- Insgesamt n nicht verwandte Patienten mit Knollenblätterpilzvergiftung werden ins AKH eingeliefert. Die Wahrscheinlichkeit p , dass ein Patient eine derartig schwere Vergiftung überlebt, ist bekannt. Wie groß ist die Wahrscheinlichkeit, dass genau x (bzw. mindestens x) Patienten die Vergiftung überleben?

Um die Wahrscheinlichkeiten zu berechnen, muss man das Rad nicht jedesmal neu erfinden. Obige Fragen können alle mittels der sogenannten **Binomialverteilung** beantwortet werden. Die Ergebnisse für $n=18$ Patienten bei einer Erfolgswahrscheinlichkeit von $p=0.35$ wurden berechnet (siehe nächste Seite). Erfolg bedeutet in unseren Fällen entweder "Infektion verläuft stumm" oder "Komplikation" oder "Überleben nach Vergiftung".

Frage: Warum wurde bei den obigen Beispielen so viel Wert auf Formulierungen wie "nicht verwandte Personen/Patienten", "erfahrene Chirurgin" und "an jeweils verschiedenen Tagen" gelegt?

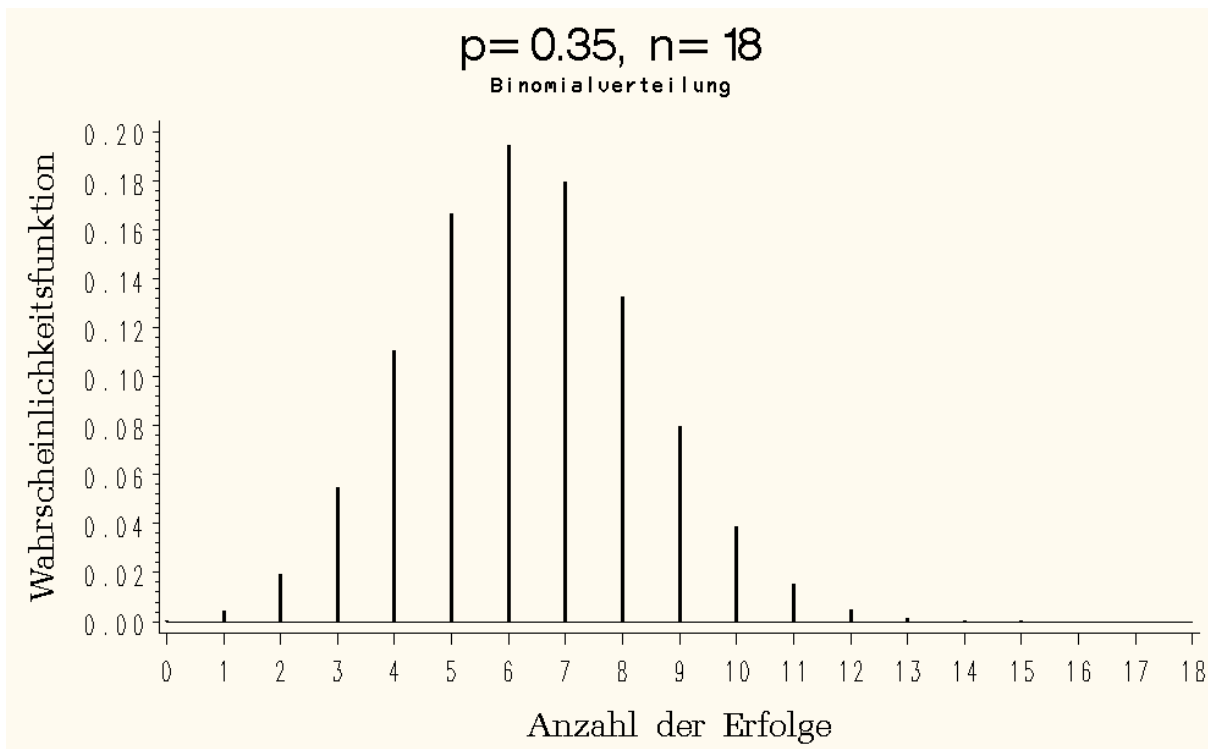
Binomialverteilung mit $n=18$, $p=0.35$

x ... Anzahl der Erfolge	Wahrscheinlichkeit	kumulierte Wahrscheinlichkeit
0	0.00043	0.00043
1	0.00416	0.00459
2	0.01903	0.02362
3	0.05465	0.07827
4	0.11035	0.18862
5	0.16638	0.35500
6	0.19411	0.54910
7	0.17918	0.72828
8	0.13266	0.86094
9	0.07937	0.94031
10	0.03846	0.97877
11	0.01506	0.99383
12	0.00473	0.99856
13	0.00118	0.99974
14	0.00023	0.99996
15	0.00003	0.99999...
16	0.00000...	0.99999...
17	0.00000...	0.99999...
18	0.00000...	1.00000

Ablesebeispiele:

- Wenn die Erfolgswahrscheinlichkeit 35 % beträgt, dann liegt die Wahrscheinlichkeit für genau 7 Erfolge bei 18 Versuchen bei 17.9 %.
- Die Wahrscheinlichkeit, dass höchstens 3 Erfolge erzielt werden, liegt bei 7.83 %.
- Die Wahrscheinlichkeit, dass mindestens 10 Erfolge erzielt werden, liegt bei 5.97 %. (Anmerkung: Die Berechnung erfolgte hier über die Gegenwahrscheinlichkeit. Höchstens 9 Erfolge zu haben weist eine Wahrscheinlichkeit von 94.03 % auf, und damit ergibt sich die Wahrscheinlichkeit für mindestens 10 Erfolge als $100\% - 94.03\% = 5.97\%$.)

Die zu umseitiger Tabelle passende graphische Darstellung wird als **Wahrscheinlichkeitsfunktion** bezeichnet:



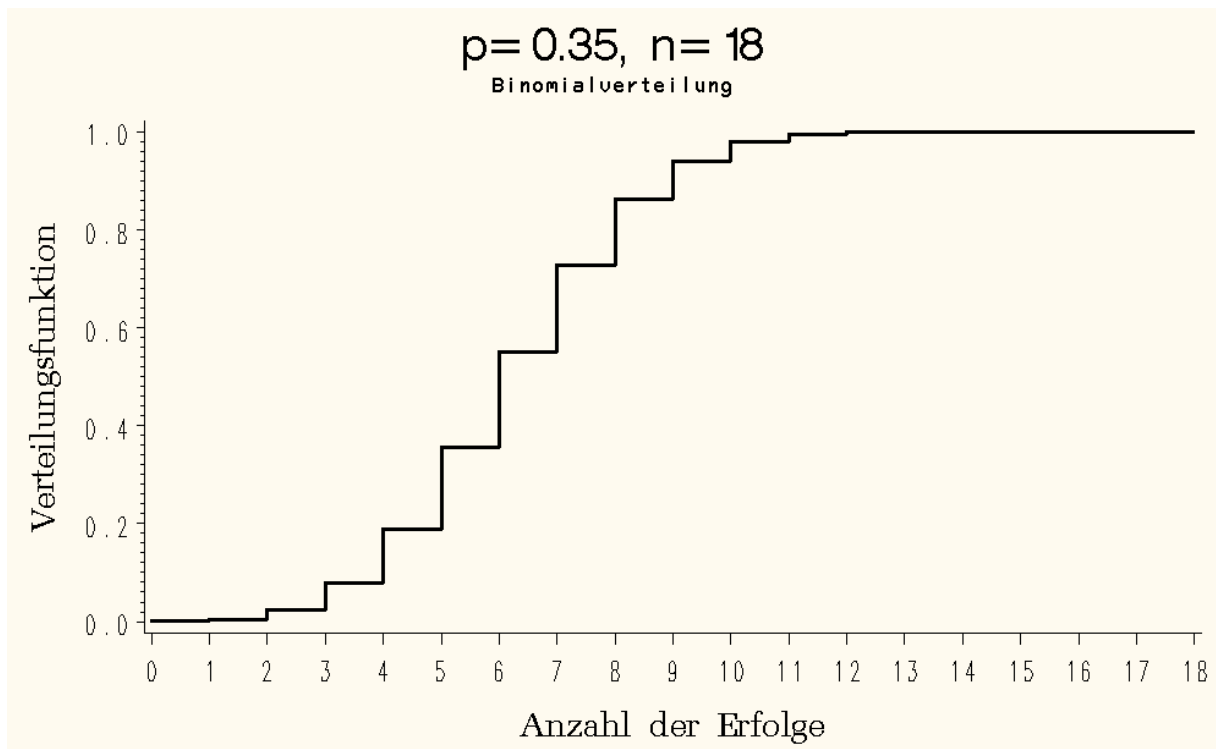
Beachten Sie hier die Tatsache, dass es unter der gegebenen Konstellation zwar **möglich** ist, 15 oder mehr der 18 Erfolge zu erzielen, dass dies aber ein **höchst unwahrscheinliches** Ereignis darstellt.

Wenn wir nun z.B. 17 Erfolge in 18 Versuchen erzielt hätten, dann gibt es grundsätzlich drei mögliche Erklärungen dafür:

- (a) dies ist wirklich ein "glücklicher" Zufall
- (b) die angegebene Erfolgswahrscheinlichkeit von 35 % pro Patient ist falsch
- (c) andere Voraussetzungen unserer Berechnungen sind nicht korrekt, wie z.B. die angenommene Unabhängigkeit der einzelnen Versuche zueinander

Übrigens, die soeben angestellte Überlegung, (a) oder (b) oder (c), wird uns in ähnlicher Form noch öfters begegnen. Sie ist eine ganz zentrale Grundüberlegung, auf der das Prinzip des statistischen Testens letztendlich beruht.

Die kumulierten Wahrscheinlichkeiten werden auch Verteilungsfunktion genannt. (Anmerkung für Leute, die sich schon mit dem Kaplan-Meier-Plot beschäftigt haben, auch dieser ist nichts anderes als eine Art "umgedrehte" Verteilungsfunktion.)



Die soeben behandelte Binomialverteilung ist eine **diskrete** Verteilung. Im folgenden sind andere bekannte diskrete Verteilungen und Beispiele für ihre Anwendung angeführt:

Binomialverteilung: wie wahrscheinlich sind x Erfolge bei n Versuchen, Erfolgswahrscheinlichkeit p bleibt dabei konstant

Hypergeometrische Verteilung: wie wahrscheinlich sind x Erfolge bei n Versuchen, wenn ohne Zurücklegen gezogen wird (z.B. beim Lotto)

Multinomialverteilung: Ähnlich zur Binomialverteilung, bei mehr als zwei Zuständen eingesetzt

Poissonverteilung: Zur Beschreibung seltener Ereignisse (z.B. in der Epidemiologie)

geometrische Verteilung: Wie viele Versuche n werden benötigt, bis zum ersten Erfolg, Trefferwahrscheinlichkeit p bleibt konstant

negative Binomialverteilung (verallgemeinert die geometrische Verteilung): Wie viele Versuche n werden benötigt, um x Erfolge zu haben, Trefferwahrscheinlichkeit p bleibt konstant

diskrete Gleichverteilung: z.B. Wurf mit fairem Würfel

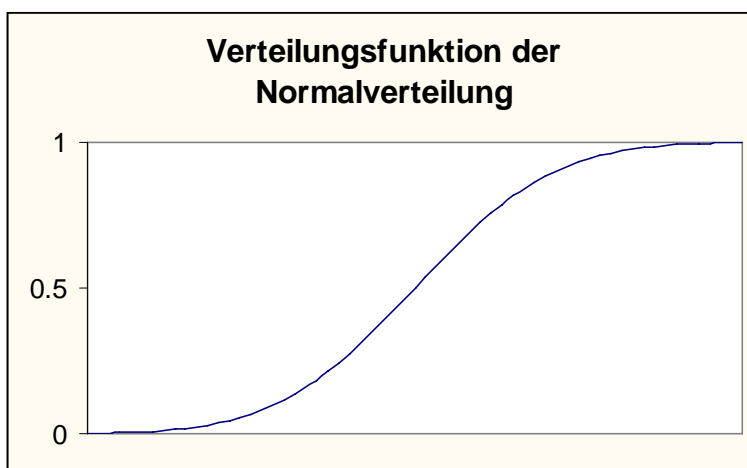
Neben diskreten Verteilungen gibt es auch **stetige** Verteilungen. Beim Schlafdauerbeispiel würden wir so eine stetige Verteilung verwenden. Es ist nämlich die einzelne Punktwahrscheinlichkeit völlig uninteressant. So interessiert sich niemand dafür, wie groß die Wahrscheinlichkeit ist, dass jemand genau 8 h 34 min 17.2348234230980901 sec schläft. Es wird immer nur nach Wahrscheinlichkeiten für Intervalle gefragt. Z.B. wie groß ist die Wahrscheinlichkeit, dass jemand

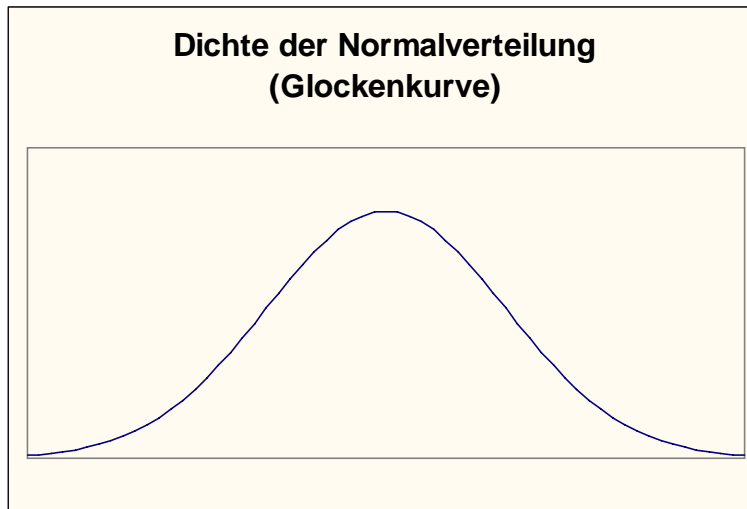
- länger als 11 h schläft?
- zwischen 8 h und 9 h 30 min schläft?

Bei stetigen Verteilungen gibt es keine Wahrscheinlichkeitsfunktion, da es einerseits so viele einzelne Punkte gibt, und andererseits jeder dieser Punkte eine unendlich kleine Wahrscheinlichkeit ("praktisch gleich null") besitzt. Außerdem interessieren sowieso nur Wahrscheinlichkeiten für Intervalle (siehe oben).

Anmerkung: Aus Sicht der Mathematik (und ihrer Teildisziplin "Maßtheorie") stellt der Übergang von diskreten zu stetigen Verteilungen eine enorme Verkomplizierung dar. Für unsere praktischen Anwendungen ist dies aber ohne größere Bedeutung.

Anstatt der Wahrscheinlichkeitsfunktion gibt es bei den stetigen Verteilungen die sogenannte **Dichtefunktion**. Wahrscheinlichkeiten für Intervalle werden durch Berechnen von Flächen unter der Dichtefunktion ermittelt. Die "kumulierte" Dichtefunktion heißt wieder **Verteilungsfunktion**. Im folgenden ist die Verteilungsfunktion und die Dichtefunktion für die wohl wichtigste stetige Verteilung, der **Normalverteilung**, schematisch abgebildet.





Aufgrund ihrer charakteristischen Form ist die Dichtefunktion der Normalverteilung unter dem Namen "Glockenkurve" bekannt. Man findet aber auch die Bezeichnung "Gauß-Kurve" (nach Carl Friedrich Gauß, 1777-1855) bzw. die Kombination "Gauß'sche Glockenkurve".

Neben der Normalverteilung gibt es noch viele andere stetige Verteilungen. Die bekanntesten sind unter anderem die t-Verteilung, die Chi-Quadrat-Verteilung, die F-Verteilung, die Gammaverteilung, die Lognormalverteilung, die Exponentialverteilung, die Weibullverteilung, die stetige Gleichverteilung und die Betaverteilung. Für jede davon gibt es (ähnlich wie bei den diskreten Verteilungen) klassische Einsatzsituationen. So werden wir später die t-, die Chi-Quadrat- und die F-Verteilung verwenden.

Warum aber ist die Normalverteilung so bedeutend?

- Die Ergebnisse vieler Experimente im biologischen Bereich sind - zumindest annähernd - normalverteilt.
- **Die Verteilung des Mittelwerts (der Summe) von Zufallsgrößen nähert mit sich steigender Stichprobenzahl der Normalverteilung an (zentraler Grenzwertsatz).**
- Andere bekannte Verteilungen "entstammen" der Normalverteilung. So kann z.B. die Chi-Quadrat-Verteilung als Summe von quadrierten normalverteilten Zufallsgrößen aufgefasst werden.

Anmerkung: Die Bezeichnung Normalverteilung bedeutet **nicht**, dass diese Verteilung der "Normalfall", also die am häufigsten vorkommende Verteilung ist, oder eine Art von Standard darstellt.

IN WELCHEM VERHÄLTNIS STEHT DIE WAHRSCHEINLICHKEITSRECHNUNG ZUR DESKRIPTIVEN STATISTIK?

Die Dichtefunktion und die Wahrscheinlichkeitsfunktion sind eng mit dem Histogramm und dem Balkendiagramm verwandt. Der Unterschied liegt dabei darin, dass erstere theoretische Einsichten bzw. die Grundgesamtheit beschreiben, während letztere zur Darstellung von Stichprobenergebnissen verwendet werden.

Unabhängig davon, ob wir es mit stetigen oder diskreten Verteilungen zu tun haben, folgende aus der Datenbeschreibung bekannte Konzepte können für alle theoretischen Verteilungen verwendet werden:

- Erwartungswert: dies ist nichts anderes als der theoretische Mittelwert, der Mittelwert der Grundgesamtheit, der "wahre" Mittelwert
- theoretische Varianz: die erwartete mittlere quadratische Abweichung vom Erwartungswert, die "wahre" Varianz
- gilt auch für Median, die sonstigen Quantile oder die Schiefe, usw.

Der wesentliche Unterschied zwischen einem Erwartungswert und einem empirischen Mittelwert ist folgender: Der Erwartungswert ist der wahre Wert in der Grundgesamtheit, er ist fix und unverrückbar. Der empirische Mittelwert ist ein aus zufälligen Ergebnissen (aus der Stichprobe) errechneter Wert, damit ist er aber selbst ein Produkt des Zufalls und besitzt auch eine Wahrscheinlichkeitsverteilung, die man studieren kann.

Dasselbe gilt natürlich auch für den Unterschied zwischen allen anderen theoretischen und empirischen Maßen.

DAS GESETZ DER GROßEN ZAHLEN UND DER ZENTRALE GRENZWERTSATZ

Zwei wichtige Grundlagen der Wahrscheinlichkeitsrechnung, die den aus einer Stichprobe errechneten Mittelwert betreffen, wollen wir uns jetzt noch näher ansehen. Der Einfachheit halber wollen wir annehmen, die Stichprobe besteht aus unabhängigen, identisch verteilten Zufallsgrößen.

Das Gesetz der großen Zahlen: Der empirische Mittelwert aus der Stichprobe nähert sich bei wachsender Fallzahl immer mehr dem Erwartungswert (Mittelwert der Grundgesamtheit) an.

In diesem Zusammenhang soll auch erwähnt werden, dass die theoretische Streuung des empirischen Mittelwerts mit wachsender Fallzahl n immer kleiner wird (exakt um den Faktor \sqrt{n}).

Hinter dem Gesetz der großen Zahlen steckt eine Binsenweisheit: Je mehr wir messen, d.h. je größer unsere Stichprobe wird, desto genauer und sicherer wird unser Ergebnis.

Beispiel: Auf einer wissenschaftlichen Fachtagung werden zwei qualitativ hochwertige Studien zum selben Thema präsentiert. Der einzige Unterschied liegt darin, dass Studie A auf 30 Patienten und Studie B auf 200 Patienten basiert. Naturgemäß werden wir die Ergebnisse von Studie B als plausibler erachten. Einfach deshalb, weil es wahrscheinlicher ist, dass aufgrund der höheren Fallzahl Studie B näher bei der Wahrheit liegt als Studie A.

Der zentrale Grenzwertsatz: Die Verteilung des empirischen Mittelwerts konvergiert mit steigender Fallzahl gegen die Normalverteilung.

Das ist eine sehr praktische Aussage, die für viele statistische Verfahren ausgenutzt wird. Leider ist sie intuitiv nur schwer zugänglich. Für Interessierte empfiehlt es sich daher, diese Aussage durch Würfelexperimente oder Computersimulationen selbst nachzuprüfen. Auch im WWW findet man diverse Java-Applets, die den zentralen Grenzwertsatz demonstrieren, unter anderem:

<http://medweb.uni-muenster.de/institute/imib/lehre/skripte/biomathe/bio/grza2.html>
<http://statistik.wu-wien.ac.at/mathstat/hatz/vo/applets/cenlimit/cenlim.html>

RESÜMEE

Mit den Mitteln der Wahrscheinlichkeitsrechnung können wir nun Aussagen über Stichproben machen, wenn wir die Wahrheit (die Grundgesamtheit) kennen. Und wir können auch "was wäre wenn"-Szenarien durchspielen! Auf diese Möglichkeit werden wir später noch oft zurückgreifen.

UNSER NÄCHSTES ZIEL

In der empirischen klinischen Forschung kennen wir die Wahrheit (die Grundgesamtheit) nicht, sonst müssten wir ja nicht forschen. Wir haben üblicherweise nur eine Stichprobe zur Verfügung, anhand der wir auf Verhältnisse in der Grundgesamtheit schließen wollen. Wir können zwar diese Schlussfolgerungen nicht mit völliger Sicherheit ziehen, wir können aber die Wahrscheinlichkeitsrechnung und damit "was wäre wenn"-Szenarien einsetzen, um die unvermeidlich vorhandene Unsicherheit zu quantifizieren.

3.3. Übungen

- 3.3.1. (a) Warum meinen Sie, ist eine vollständige Überprüfung aller Gulaschdosen nicht sinnvoll?
- (b) Nennen Sie Beispiele aus dem medizinischen Bereich, wo vollständige Qualitätsprüfungen nicht sinnvoll sind.
- 3.3.2. Nehmen wir an, unter den vielen Gulaschdosen gibt es wirklich nur 5 mit verdorbenem Inhalt. Eine Stichprobe von 20 wird gezogen und die 5 verdorbenen Gulaschdosen sind dabei. Kann es sein, dass diese Stichprobe trotzdem als repräsentativ für die Grundgesamtheit aller Gulaschdosen angesehen werden könnte?
- 3.3.3. Nennen Sie Voraussetzungen, damit die Patientenauswahl bei einer klinischen Studie zu einer repräsentativen Stichprobe führt.
- 3.3.4. Besonders im angloamerikanischen Raum werden Wahrscheinlichkeiten gerne in Form von Wettchancen (engl. odds) angegeben. Im Gegensatz zur klassischen Wahrscheinlichkeitsdefinition, wo man "günstige" zu "möglichen" Ereignissen in Beziehung setzt, werden Odds als "günstige" zu "komplementären/ungünstigen" Ereignissen definiert. Odds sind Zahlenwerte größer-gleich 0, die nach oben hin unbeschränkt sind. Zum Beispiel: Die Odds für eine Knabengeburt wären $0.514:0.486=1.0576$.
- (a) Berechnen Sie die Odds für eine Mädchengeburt.
- (b) Angenommen, vor Beginn der Fußballeuropameisterschaft 2008 in Österreich und der Schweiz stehen die Odds, dass die Schweiz Europameister wird, bei 2:7. Wie groß ist demnach die Wahrscheinlichkeit für einen Fußballeuropameister Schweiz?
- 3.3.5. Ein neu entwickelter HIV-Test entdeckt 98 % der tatsächlich HIV-Positiven. Allerdings schlägt er auch bei 5 % der HIV-Negativen an.
- (a) Berechnen Sie Sensitivität und Spezifität des neuen HIV-Tests.
- (b) Berechnen Sie den positiven und den negativen Vorhersagewert des neuen HIV-Tests.

3.3. Übungen

- 3.3.6. Sind zwei Ereignisse, die sich gegenseitig ausschließen, im allgemeinen unabhängig? (Versuchen Sie sich den Sachverhalt an Hand eines selbstgewählten Beispiels zu verdeutlichen)
- 3.3.7. Während einer wissenschaftlichen Tagung wird im Rahmenprogramm auch ein Casinobesuch angeboten. Sie gehen hin und treffen beim Roulettetisch einen alten Bekannten, der sich schon länger dort aufhält. Er berichtet aufgeregt, dass die Kugel zuletzt 8-mal hintereinander auf "Rot" gefallen sei. Der Bekannte meint daher, es wäre jetzt ratsam auf "Schwarz" zu setzen, denn dies wäre bereits sehr wahrscheinlich. Er tut es und tatsächlich kommt beim nächsten Mal "Schwarz". Was meinen Sie dazu?
- 3.3.8. (aus Gigerenzer, 2002) Ein bayerischer Innenminister äußerte sich einmal über die Gefahren des Drogenmissbrauchs und erklärte, weil die meisten Heroinabhängigen Marihuana geraucht hätten, würden die meisten Marihuanaraucher auch zu Heroinsüchtigen. Beurteilen Sie die Schlussfolgerung des bayerischen Innenministers aus Sicht der Wahrscheinlichkeitsrechnung. (**Anleitung:** Definieren Sie die entsprechenden Ereignisse, und identifizieren Sie die in der Schlussfolgerung verwendeten bedingten Wahrscheinlichkeiten. Auch Venn-Diagramme können nützlich sein.)
- 3.3.9. Infektionskrankheiten können auch "stumm" verlaufen. Angenommen, die Wahrscheinlichkeit für den stummen Verlauf einer bestimmten Infektion liegt bei 40 %. Drei Personen sind infiziert. Wie groß ist die Wahrscheinlichkeit für
- (a) drei stumme Verläufe,
 - (b) zwei stumme und einen offenen Verlauf,
 - (c) einen stummen und zwei offene Verläufe,
 - (d) drei offene Verläufe?
- (e) Angenommen, Sie erfahren, alle drei Personen sind miteinander eng verwandt. Könnte dies Auswirkungen auf die Gültigkeit der Berechnungen von (a)-(d) haben?
- 3.3.10. Angenommen, die Voraussetzungen für das Hardy-Weinberg-Gleichgewicht gelten.
- (a) Berechnen Sie die einzelnen Phänotypwahrscheinlichkeiten bei den Blutgruppen in einer kaukasischen Bevölkerung.

3.3. Übungen

(b) Wie groß ist die Wahrscheinlichkeit, dass in dieser Bevölkerung zwei zufällig ausgewählte Menschen die gleiche Blutgruppe haben?

(c) Wie groß ist die Wahrscheinlichkeit, dass in dieser Bevölkerung ein biologisches Elternpaar (beide mit Blutgruppe A) ein Kind mit Blutgruppe O bekommt?

(d) Wie groß ist die Wahrscheinlichkeit, dass in dieser Bevölkerung ein biologisches Elternpaar (beide mit Blutgruppe O) ein Kind mit Blutgruppe A bekommt?

3.3.11. Wie groß sind die Wahrscheinlichkeiten im Lotto 6 aus 45 für 0, 1, 2, 3, 4, 5 und 6 Richtige? Berechnen Sie auch die Wahrscheinlichkeiten für 5 Richtige ohne und mit Zusatzzahl.

3.3.12. Angenommen, zur Behandlung einer bestimmten Erkältungskrankheit gibt es zwei Therapien, die Standardtherapie A und die neue Therapie B. Mit A beträgt die Erfolgsquote $p_A=24\%$, und mit B beträgt sie $p_B=26\%$. Die NNT (*Number Needed to Treat*), um den Vorteil von Therapie B gegenüber Therapie A zu beschreiben, ist wie folgt definiert:

$$NNT = 1/(p_B - p_A) = 1/(0.26 - 0.24) = 1/0.02 = 50$$

(Achtung: Prozentsätze müssen auf Anteile umgeformt werden!)

(a) Können Sie das Ergebnis interpretieren? Was wäre eigentlich im Falle von $p_A=26\%$ und $p_B=24\%$?

(b) Berechnen Sie die NNT für $p_A=3.5\%$ und $p_B=16.1\%$. Interpretieren Sie das Ergebnis.

(c) Nehmen Sie an, Therapie A würde immer und B nie wirken. Berechnen und interpretieren Sie die NNT.

(d) Nehmen Sie an, beide Therapien würden gleich gut wirken. Berechnen und interpretieren Sie die NNT.

(e) Eine Kollegin erzählt Ihnen, die NNT von Aspirin bei Kopfschmerzen wäre 4.2 Patienten. Geben Sie sich mit dieser Aussage zufrieden?

Kapitel 4

Statistische Test I

4.1. Das Prinzip von statistischen Tests

Beispiel 4.1.1.: (metrische Zielgröße)

Tierversuch von Dr. X: Zwei Gruppen weiblicher Ratten erhalten stark bzw. schwach proteinhaltiges Futter

Forschungsfrage: Gibt es Unterschiede bei Gewichtszunahme (vom 28. zum 84. Lebensstag)?

Ergebnis (Gewichtszunahme in Gramm):

Gruppe 1 (viel Protein):

134, 146, 104, 119, 124, 161, 107, 83, 113, 129, 97, 123

Gruppe 2 (wenig Protein):

70, 118, 101, 85, 107, 132, 94

Auswertung von Dr. X:

mittlere Gewichtszunahme in Gruppe 1 (viel Protein)	120 g
--	-------

mittlere Gewichtszunahme in Gruppe 2 (wenig Protein)	101 g
---	-------

Konklusion von Dr. X:

Viel Protein im Futter bringt bei jungen weiblichen Ratten mehr Gewichtszunahme als bei schwach proteinhaltiges Futter

(Verallgemeinerung von Stichprobe auf Grundgesamtheit!)

Behauptung des Kollegen Y:

Unterschiede sind ZUFALL! In Wirklichkeit verursachen beide Diäten identische Gewichtsveränderungen. Obige Konklusion ist wertlos.

Einerseits:

Kollege Y könnte recht haben! Sind also die Ergebnisse reiner Zufall?

Andererseits:

Mit dem Argument „alles Zufall“ kann jede Schlussfolgerung aus Beobachtungen oder Experimenten in Zweifel gezogen werden!

Was sollen wir nun tun?

Erkenntnis:

Wollen wir zu einer Entscheidung kommen, dann müssen wir die Möglichkeit einer **falsch positiven** Antwort akzeptieren! Die Statistiker sprechen in so einem Fall vom Fehler 1. Art bzw. α -Fehler.

Konkret am Beispiel: Wenn wir den Schluss ziehen, dass verschieden proteinhaltiges Futter zu unterschiedlichen Gewichtszunahmen führt, dann könnte dies auch ein **Fehlschluss** sein. Um diese Tatsache kommen wir nicht herum. Anders formuliert: Wenn wir bei einer empirischen Studie zu einer positiven Antwort gelangen, dann müssen wir mit der Ungewissheit leben, dass es auch eine falsch positive Antwort sein könnte!

Dies ist allerdings auch nicht die letzte Weisheit. Findige KollegInnen könnten somit jedes noch so unbedeutende Ergebnis als empirischen Beweis für ihre Forschungshypothese deklarieren und KritikerInnen mit der Replik: "Ich akzeptiere immer die Möglichkeit einer falsch positiven Antwort!" ins Leere laufen lassen. Konsequenterweise folgt daraus, dass eine mögliche falsch-positive Entscheidung *nicht beliebig oft* und auch *nicht nach individueller Willkür* getroffen werden darf.

Wenn die Möglichkeit für eine falsch-positive Antwort besteht, dann wollen wir so eine Antwort

- **nur selten geben**
- **nur in Situationen geben, wo das "alles Zufall"-Argument aufgrund der beobachteten Daten als inplausibel erscheint**

Man kann obige Überlegungen formalisieren,
und erhält dann einen **statistischen Test**

Das Prinzip eines statistischen Tests:

- **Festlegung der Nullhypothese, zumeist:**
„Es gibt keinen Unterschied!“
Beim Beispiel 4.1.1. entspricht die Nullhypothese der Ansicht des skeptischen Kollegen Y.

Anmerkung: Die Nullhypothese ist oft die Negation der Forschungsfrage.

- **Alternativhypothese: „Es gibt einen Unterschied!“**
Das ist natürlich eine sehr breite Aussage. Vorerst aber wollen wir uns damit zufrieden geben.

- **Nimm an, die Nullhypothese ist wahr.**
Berechne die Wahrscheinlichkeit, dann das aktuelle Ergebnis (oder ein noch Extremes) zu beobachten.
Diese Wahrscheinlichkeit heißt p-Wert. Die Berechnung erfolgt heutzutage fast ausschließlich mittels Computerprogrammen.

Beachte: der p-Wert entspricht nicht der Wahrscheinlichkeit, dass die Nullhypothese gilt. Der p-Wert kann eher als eine Art bedingter Wahrscheinlichkeit aufgefasst werden, die Bedingung ist dabei die angenommene Gültigkeit der Nullhypothese.

Anmerkung: Einem mathematischen Statistiker wird allerdings die Aussage "der p-Wert ist eine bedingte Wahrscheinlichkeit" genauso gegen den Strich gehen, wie einem Mediziner die Beschreibung eines Wadenbeinbruchs als "gebrochenem Fuß". Umgangssprachlich sind beide Aussagen akzeptabel, eng fachlich aber inkorrekt.

Nur für Interessierte (und wirklich nur für diese): Es handelt sich beim p-Wert um eine Einschränkung des Parameterraums und nicht - wie für die Definition einer bedingten Wahrscheinlichkeit notwendig - um eine Einschränkung des Ereignisraums.

- **Ist der errechnete p-Wert kleiner oder gleich 0.05, dann wird die Nullhypothese verworfen, und das Ergebnis wird als STATISTISCH SIGNIFIKANT bezeichnet.**

Die Grenze von 5 % nennt man das Signifikanzniveau, oft wird der griechische Buchstabe α als Abkürzung dafür verwendet.

Anmerkung: Theoretisch wäre jede Zahl zwischen 0 und 1 als Signifikanzniveau verwendbar, praktisch sind aber nur kleine Werte wie 0.01, 0.05 oder 0.10 wirklich sinnvoll. In der Medizin hat sich der Wert von $\alpha = 0.05$ als Standard etabliert.

Ein statistischer Test entspricht also einem „was-wäre-wenn“ Szenario:

1. Was-wäre-wenn die Nullhypothese wahr wäre?
2. Sind unsere Daten damit plausibel erklärbar?
3. Der p-Wert misst Grad der Plausibilität
4. Wenn der p-Wert "klein" ist, dann bietet Nullhypothese keine plausible Erklärung für unsere Daten
5. Dann "muss" es wohl einen andere Erklärung geben (Alternativhypothese!), wir „verwerfen“ die Nullhypothese

4.2. t-Test

Zurück zum Rattendiäts-Beispiel 4.1.1.:

Wenn der p-Wert kleiner als 5% wäre, dann könnte Dr. X behaupten, dass eine Variation in der Proteindiät zu Unterschieden in der Gewichtszunahme bei Ratten führt. Und der Einwand des skeptischen Kollegen Y wäre damit entkräftet.

Beachte, Kollege Y könnte in Wirklichkeit trotzdem recht haben, durch ein signifikantes Ergebnis erwürbe Dr. X aber den „Anspruch“ auf Respektierung seines Ergebnisses.

Obige Aussagen stehen im Konjunktiv, denn sie sind nur gültig bei einem signifikantem Ergebnis. Wie können wir nun einen etwaigen signifikanten Unterschied beim Rattendiät-Beispiel feststellen? Eine Möglichkeit dafür bietet der t-Test für unabhängige Stichproben (auch ungepaarter t-Test genannt).

Die Nullhypothese steht fest:

„Kein Unterschied in der Gewichtszunahme zwischen den beiden Proteindiäten.“

Berechnung:

Nimm an, die Nullhypothese wäre wahr. Berechne dann die Wahrscheinlichkeit, das aktuelle Ergebnis oder noch extremere Ergebnisse zu beobachten.

Was ist aber ein "extremes Ergebnis?"

Jedes Ergebnis, das weiter von der Nullhypothese entfernt ist, als unser aktuell beobachtetes Ergebnis.

Notwendig:

Wir brauchen ein intuitives Maß, welches Abweichungen von der Nullhypothese feststellt.

Dafür bietet sich naturgemäß die Differenz der Mittelwerte zwischen den beiden Gruppen an. Denn je größer der Unterschied zwischen den beiden Mittelwerten ist, ein desto extremes Ergebnis liegt vor. Anders formuliert: Desto weiter sind wir von der Nullhypothese weg.

Konkret beim Beispiel 4.1.1.:

Nimm an, die beiden Diäten verursachen in Wahrheit keinen Gewichts-

unterschied. Berechne die Wahrscheinlichkeit, dann zufällig einen absoluten Gewichtsunterschied von 19 Gramm oder größer zu beobachten (geht in beide Richtungen - zweiseitiger Test).

Wenn

- **annähernd Normalverteilung vorliegt,**
- **und wenn die Streuungen in beiden Gruppen in etwa gleich groß sind,**

dann ist unter Gültigkeit der Nullhypothese der beobachtete Mittelwertsunterschied dividiert durch seine geschätzte Streuung („standard error“) t-verteilt. Diese Zahl wird übrigens auch **Teststatistik** genannt.

Warum wird der beobachtete Mittelwertsunterschied durch seine Streuung dividiert? Einerseits um ein dimensionsloses allgemeines Maß zu erhalten, und andererseits um den Umstand zu berücksichtigen, dass die relative Bedeutung eines Unterschieds von z.B. 19 Gramm sehr davon abhängt, wie stark die Populationen streuen.

Wenn die oben genannten Voraussetzungen erfüllt sind, dann können wir die t-Verteilung verwenden, um den p-Wert zu ermitteln. Dies wird heutzutage fast ausschließlich mit Hilfe von Computerprogrammen durchgeführt.

Verwendung von SPSS

Jetzt spricht eigentlich nichts mehr dagegen, das Beispiel auch tatsächlich mit SPSS anzugehen.

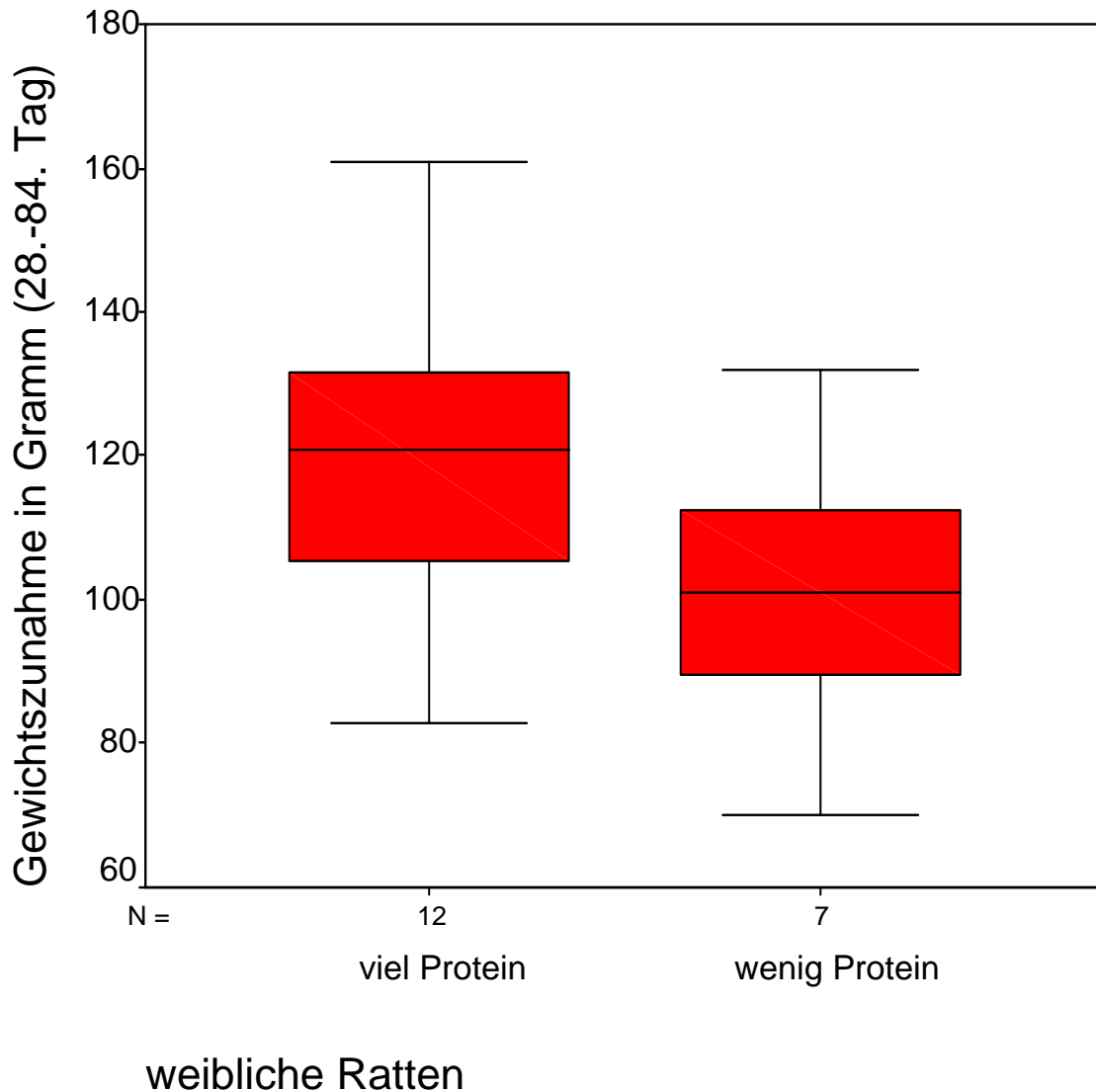
- I.) Zuerst müssen die Daten eingegeben werden
- II.) Dann veranschaulichen wir uns die Situation durch graphische und/oder sonstige deskriptive Hilfsmittel
- III.) Anhand der Ergebnisse von II überprüfen wir, ob die Voraussetzungen zur Durchführung des von uns geplanten Testverfahrens überhaupt erfüllt sind
- IV.) Erst dann Durchführung des Tests
- V.) Interpretation der Ergebnisse

ad I.) Für das Rattendiätbeispiel 4.1.1. legen wir zuerst zwei Variablen an, GRUPPE und ZUNAHME:

GRUPPE ist eine binäre Variable zur Unterscheidung des Proteingehalts im Futter (1=viel Protein, 2=wenig Protein)

ZUNAHME ist eine metrische Variable, die die gemessene Gewichtszunahme in Gramm enthält

ad II.) Zur graphischen Darstellung der Daten wurden Boxplots verwendet.



ad III.) Die Gewichtszunahme zeigt in beiden Gruppen eine symmetrische Verteilung ohne Ausreißer, auch streuen die Daten in beiden Gruppen ungefähr gleich stark. Der geplanten Verwendung des t-Tests steht also nichts im Wege.

ad IV.) Zur Durchführung des t-Tests klicken wir auf

Analysieren

Mittelwerte vergleichen

T-Test bei unabhängigen Stichproben

Wir verschieben nun die Variable ZUNAHME ins Feld

Testvariable(n)

und die Variable GRUPPE ins Feld

Gruppenvariable

Neben letzterer scheinen zwei Fragezeichen auf. Wir klicken daher auf den Button

Gruppen definieren

und ordnen dem Feld

Gruppe 1

den Wert 1 und dem Feld

Gruppe 2

den Wert 2 zu, denn dies waren die von uns gewählten Gruppencodes, um zwischen der Gruppe mit viel Protein und der mit wenig Protein zu unterscheiden.

Wir klicken dann auf

Weiter

und

Ok

und der gewünschte t-Test wird berechnet. Zuerst erhalten wir deskriptive Maßzahlen pro Gruppe.

Gruppenstatistiken

		N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Gewichtszunahme in Gramm (28.-84. Tag)	weibliche Ratten viel Protein	12	120	21.388	6.174
	wenig Protein	7	101	20.624	7.795

Das Testergebnis wird in einer sehr breiten und damit unübersichtlichen Tabelle ausgedruckt. Diese wurde im folgenden in 3 Teile zerlegt. Die für uns vorerst wichtigen Teile wurden grau unterlegt.

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit	
		F	Signifikanz
Gewichtszunahme in Gramm (28.-84. Tag)	Varianzen sind gleich	0.015	0.905
	Varianzen sind nicht gleich		

Test bei unabhängigen Stichproben

		T-Test für die Mittelwertgleichheit			
		T	df	Sig. (2-seitig)	Mittlere Differenz
Gewichtszunahme in Gramm (28.-84. Tag)	Varianzen sind gleich	1.891	17	0.076	19.000
	Varianzen sind nicht gleich	1.911	13.082	0.078	19.000

Test bei unabhängigen Stichproben

		T-Test für die Mittelwertgleichheit		
		Standardfehler der Differenz	95% Konfidenzintervall der Differenz	
			Untere	Obere
Gewichtszunahme in Gramm (28.-84. Tag)	Varianzen sind gleich	10.045	-2.194	40.194
	Varianzen sind nicht gleich	9.944	-2.469	40.469

Die "Mittlere Differenz" von 19 dividiert durch ihre Streuung ("Standardfehler der Differenz") von 10.045 ergibt den beobachteten Teststatistik-Wert "T" von 1.891. Da unter Gültigkeit der Nullhypothese diese Teststatistik eine t-Verteilung mit 17 Freiheitsgraden ("df") besitzt, errechnet sich ein p-Wert ("Sig. (2-seitig)") von 0.076.

ad V.) Beim üblichen **Signifikanzniveau** von 5 % folgt nun: Die Nullhypothese der Mittelwertgleichheit kann nicht verworfen werden, womit auch Dr. X den „Alles Zufall“-Einwand des Kollegen Y nicht entkräften kann!!

4.3. Wilcoxon Rangsummentest

Der Wilcoxon Rangsummentest ist ein sogenannter nicht-parametrischer Test. Er ist äquivalent zum Mann-Whitney U-Test. Man liest daher des öfteren auch Wilcoxon-Mann-Whitney U-Test. Wir wollen in anhand des Beispiels 4.1.1. kennen lernen.

Nullhypothese steht fest: „Kein Unterschied in der Gewichtszunahme zwischen den Protein-Diäten.“

Berechnung: Nimm an, die Nullhypothese ist wahr. Berechne die Wahrscheinlichkeit, dann das aktuelle Ergebnis (oder noch extremere Ergebnisse als das Aktuelle) zu beobachten.

Notwendig: Ein Maß, welches „extreme Ergebnisse“ feststellt. Zum Beispiel Rangsumme der kleineren Gruppe (das ist beim Rattenbeispiel die Gruppe 2, da nur 7 Ratten).

Exkurs: Berechnung der Rangsummen

Ränge	Gewichtszunahme	Ränge Gruppe 1	Ränge Gruppe 2
(kleinster Wert) 1	70		1
2	83	2	
3	85		3
4	94		4
5	97	5	
6	101		6
7	104	7	
8.5	107	8.5	
8.5	107		8.5
10	113	10	
11	118		11
12	119	12	
13	123	13	
14	124	14	
15	129	15	
16	132		16
17	134	17	
18	146	18	
(größter Wert) 19	161	19	
Σ 190		Σ 140.5	Σ 49.5

Was ist ein extremes Ergebnis?

Gesamte Rangsumme ist 190.

Durchschnittlicher Rang pro Ratte ist 10.

Durchschnittliche Rangsumme von 7 Ratten sollte daher 70 sein.

Je größer Abweichung von 70, desto extremer.

Konkret am Beispiel:

Nimm an, es bestünde kein Diät-Unterschied. Berechne die Wahrscheinlichkeit, dann bei 7 Ratten eine Rangsumme von 49.5 oder kleiner bzw. 90.5 oder größer zu beobachten (wieder zweiseitiger Test).

Um mittels SPSS den Wilcoxon-Mann-Whitney U-Test durchzuführen, klicken wir auf

Analysieren

Nichtparametrische Tests

Zwei unabhängige Stichproben

Wir verschieben nun die Variable ZUNAHME ins Feld

Testvariable(n)

und die Variable GRUPPE ins Feld

Gruppenvariable

Neben letzterer scheinen zwei Fragezeichen auf. Wir klicken daher auf den Button

Gruppen definieren

und ordnen dem Feld

Gruppe 1

den Wert 1 und dem Feld

Gruppe 2

den Wert 2 zu, denn dies waren die von uns gewählten Gruppencodes, um zwischen der Gruppe mit viel Protein und der mit wenig Protein zu unterscheiden. Wir klicken dann auf

Weiter

Bei der Rubrik

Welche Tests durchführen

wählen wir den Mann-Whitney U-Test aus. Schließlich klicken wir auf

Ok

und der gewünschte Wilcoxon-Mann-Whitney U-Test wird berechnet.

Zuerst erhalten wir deskriptive Maßzahlen über die Ränge pro Gruppe.

Ränge				
	weibliche Ratten	N	Mittlerer Rang	Rangsumme
Gewichtszunahme in Gramm (28.-84. Tag)	viel Protein	12	11.71	140.50
	wenig Protein	7	7.07	49.50
	Gesamt	19		

Die nächste Tabelle enthält die Testergebnisse. Die für uns wichtigen Teile wurden grau unterlegt.

Statistik für Test ^b	
	Gewichtszunahme in Gramm (28.-84. Tag)
Mann-Whitney-U	21.500
Wilcoxon-W	49.500
Z	-1.733
Asymptotische Signifikanz (2-seitig)	0.083
Exakte Signifikanz [2*(1-seitig Sig.)]	0.083 ^a

- a. Nicht für Bindungen korrigiert.
- b. Gruppenvariable: weibliche Ratten

Der errechnete p-Wert ("Asymptotische Signifikanz (2-seitig)") von 0.083 zeigt wie schon der t-Test keinen statistisch signifikanten Unterschied.

Der Vorteil des Wilcoxon-Mann-Whitney U-Tests liegt darin, dass er auch dann eingesetzt werden kann, wenn die Voraussetzungen für den t-Test nicht gegeben sind. Andererseits, wenn die Voraussetzungen für den t-Test gegeben sind (so wie beim Rattendiätbeispiel), dann sollte dieser dem Wilcoxon-Mann-Whitney U-Test vorgezogen werden. Eine einfache **Entscheidungshilfe** bietet die folgende Regel: Immer dann, wenn man mit "gutem Gewissen" Mittelwerte zur Beschreibung der Daten in den beiden Gruppen einsetzen kann, dann ist auch der Einsatz des t-Tests gerechtfertigt, denn er basiert ja auf dem Mittelwertsvergleich.

Eine **Faustregel** warnt vor Problemen beim Wilcoxon-Mann-Whitney U-Test. Wenn in einer der beiden Gruppen weniger als 10 Werte vorhanden sind, dann kann der asymptotisch ermittelte p-Wert ("Asymptotische Signifikanz (2-seitig)") relativ ungenau werden. Es empfiehlt sich dann, den p-Wert "exakt" über die Permutationsverteilung zu ermitteln. Diese Faustregel trifft bei unserem Beispiel zu, denn die kleinere Gruppe umfaßt nur 7 Ratten.

SPSS errechnet den exakten p-Wert für den Wilcoxon-Mann-Whitney U-Test automatisch ("Exakte Signifikanz [2*(1-seitig Sig.)]"), wobei sich beim Beispiel ein Wert von 0.083 ergibt, der sich hier überhaupt nicht vom asymptotischen p-Wert unterscheidet.

Allerdings ist das Ermitteln von exakten p-Werten eine rechentechnisch relativ schwierige Sache, was dazu führt, dass SPSS in seiner Standardversion nur eine Sparvariante dieser Technik anbietet. Konkret, beim Vorliegen von Bindungen in den Daten ist der ermittelte exakte p-Wert nicht völlig korrekt, und SPSS weist auch in der Fußnote "a" ganz explizit darauf hin.

Da bei unserem Beispiel eine Bindung vorkommt (der Wert von 107 Gramm Gewichtszunahme scheint in beiden Gruppen auf), stellt sich die Frage, wie man nun zu einem völlig exakten p-Wert kommt. Hier besteht die Möglichkeit, das SPSS-Modul "Exact Tests" zusätzlich zu installieren. Damit ergibt sich dann ein exakter p-Wert ("Exakte Signifikanz (2-seitig)") von 0.087. Auch dieser Wert unterscheidet sich nur marginal vom asymptotisch ermittelten p-Wert.

Anmerkung: Bei geringen Fallzahlen bietet ein exakt ermittelter p-Wert oftmals die einzige Alternative. Leider ist aber der Begriff "exakt" sehr unglücklich gewählt, da der Eindruck von genau und damit besser entsteht. Dabei wird ein Nachteil von exakten Tests übersehen: Sie sind nämlich zumeist konservativ, das heißt, die zugestandene Fehlerwahrscheinlichkeit (von z.B. 5 Prozent) wird nicht vollständig ausgenützt.

4.4. Übungen

- 4.4.1. Bei 13 mageren und bei 9 schwer übergewichtigen Frauen wurde der 24-Stunden-Total-Energie-Verbrauch (Angabe in MJ/Tag) bestimmt.

Werte für die Mageren:

6.13, 7.05, 7.48, 7.48, 7.53, 7.58, 7.90, 8.08, 8.09, 8.11, 8.40, 10.15, 10.88

Werte für die Übergewichtigen:

8.79, 9.19, 9.21, 9.68, 9.69, 9.97, 11.51, 11.85, 12.79

a.) Gibt es Unterschiede im Energieverbrauch zwischen den beiden Gruppen?

b.) In der Datei b4_4_1.sav finden Sie die Studiendaten. Beim Anlegen der Datei ist ein Fehler passiert. Welcher?

- 4.4.2. In der Datei b4_1_1.sav finden Sie die Rattendiätdateien. Ändern Sie den größten Wert in der Gruppe 1 auf 450 Gramm.

(a) Führen Sie einen t-Test durch

(b) Führen Sie einen Wilcoxon-Mann-Whitney U-Test durch

(c) Vergleichen und kommentieren Sie die Ergebnisse

- 4.4.3. Wäre es sinnvoll, bei den Teilnehmern dieses Seminars verschiedene Merkmale (wie Körpergröße in cm, Alter in Jahren, bereits graduiert ja/nein, etc.) zu erheben, und diese dann auf Unterschiede zwischen Männern und Frauen zu testen?

Kapitel 5

Statistische Tests II

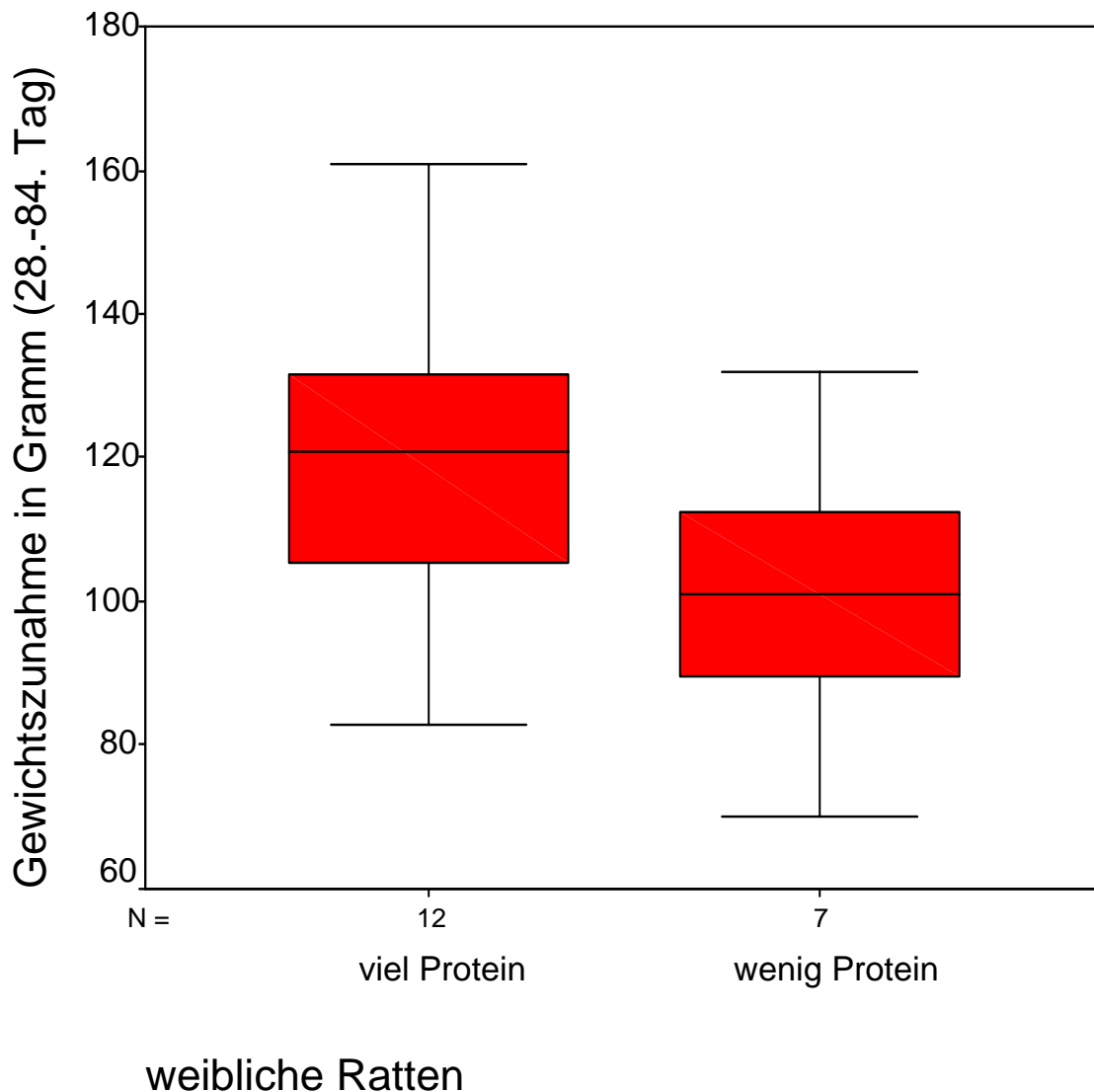
5.1. Mehr zum t-Test bei unabhängigen Stichproben

Wiederholung von Beispiel 4.1.1.: (stetige Zielgröße)

Tierversuch von Dr. X: Zwei Gruppen weiblicher Ratten erhalten stark bzw. schwach proteinhaltiges Futter

Forschungsfrage: Gibt es Unterschiede bei Gewichtszunahme (vom 28. zum 84. Lebenstag)?

Boxplot zeigt sehr ähnliche Streuungen in beiden Gruppen:



5.1. Mehr zum t-Test bei unabhängigen Stichproben

Wenn wir dafür den t-Test mit SPSS berechnen, dann erhalten wir zuerst eine Tabelle mit deskriptiven Maßzahlen. Auch daran können wir sehen, dass beide Gruppen sehr ähnliche Streuungen aufweisen (Standardabweichungen 21.4 und 20.6).

Gruppenstatistiken

	weibliche Ratten	N	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes
Gewichtszunahme in Gramm (28.-84. Tag)	viel Protein	12	120	21.388	6.174
	wenig Protein	7	101	20.624	7.795

Das t-Test Ergebnis besteht interessanterweise aus zwei Zeilen. Die erste Zeile enthält das Ergebnis "Varianzen sind gleich", die zweite "Varianzen sind nicht gleich". Da in unserem Beispiel die Standardabweichungen (und damit auch die Varianzen) sehr ähnlich sind, entscheiden wir uns für das Ergebnis in der ersten Zeile.

Test bei unabhängigen Stichproben

		T-Test für die Mittelwertgleichheit			
		T	df	Sig. (2-seitig)	Mittlere Differenz
Gewichtszunahme in Gramm (28.-84. Tag)	Varianzen sind gleich	1.891	17	0.076	19.000
	Varianzen sind nicht gleich	1.911	13.082	0.078	19.000

Test bei unabhängigen Stichproben

		T-Test für die Mittelwertgleichheit		
		Standardfehler der Differenz	95% Konfidenzintervall der Differenz	
			Untere	Obere
Gewichtszunahme in Gramm (28.-84. Tag)	Varianzen sind gleich	10.045	-2.194	40.194
	Varianzen sind nicht gleich	9.944	-2.469	40.469

Warum wird zwischen einem t-Test für gleiche und einem t-Test für ungleiche Varianzen unterschieden?

Wir wissen bereits: Die "Mittlere Differenz" von 19 dividiert durch ihre Streuung ("Standardfehler der Differenz") von 10.045 ergibt den beobachteten Teststatistik-Wert "T" von 1.891. Unter Gültigkeit der Nullhypothese besitzt diese Teststatistik eine t-Verteilung mit 17 Freiheitsgraden ("df"). Damit können wir den p-Wert errechnen.

Dies ist allerdings nur bei gleicher Varianz in den zugrundeliegenden Grundgesamtheiten gültig.

Bei **ungleichen Varianzen** ergibt sich ein Problem. Dieses ist bereits sehr lange bekannt und heißt auch **Behrens-Fisher-Problem**. Kurz, die Streuung der mittleren Differenz ("Standardfehler der Differenz") und die Freiheitsgrade ("df") für die t-Verteilung müssen korrigiert werden. Wir erhalten daher auch einen etwas anderen p-Wert.

Wann sollen wir jetzt den t-Test für gleiche und wann den für ungleiche Varianzen verwenden?

Grundsätzlich gilt: Der t-Test ist **robust** gegenüber kleinen bis moderaten Abweichungen von den Voraussetzungen. Als sehr grobe Faustregel zu obiger Frage könnte man daher vielleicht angeben, dass bei einem Unterschied um den Faktor 3 in den Standardabweichungen die Variante des t-Tests mit den ungleichen Varianzen verwendet werden sollte. Allerdings ist das Problem gleiche/ungleiche Varianz sehr von der Stichprobengröße abhängig. Bei großen Fallzahlen wird es immer weniger bedeutsam.

Neben diesen etwas schwammigen Faustregeln gibt es aber noch eine andere Möglichkeit. Manche von Ihnen werden sich bereits gedacht haben: Für eine Frage wie *"gleiche Varianz versus ungleiche Varianz"* zwischen den beiden Gruppen wird es doch wohl auch einen statistischen Test geben?

Gibt es auch!

SPSS verwendet den sogenannten **Levene-Test der Varianzgleichheit**.

Nullhypothese: Die beiden Stichproben kommen aus Grundgesamtheiten mit gleichen Varianzen

Alternativhypothese: Die Varianzen sind ungleich

5.1. Mehr zum t-Test bei unabhängigen Stichproben

Wir wissen bereits:

Ein einigermaßen vernünftiges Maß zur Bestimmung der Abweichung von der Nullhypothese wird benötigt. SPSS verwendet dazu die sogenannte Levene-Statistik (wir gehen aber nicht näher darauf ein).

Unter der Bedingung, dass die Nullhypothese (Varianzgleichheit) gilt, kann man dann die Wahrscheinlichkeit ermitteln, den aus den aktuellen Daten errechneten Wert oder noch extremere Werte für die Levene-Statistik zu beobachten. Diese bedingte Wahrscheinlichkeit ist der p-Wert.

SPSS führt den Levene-Test beim Aufruf des t-Tests automatisch durch.

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit	
		F	Signifikanz
Gewichtszunahme in Gramm (28.-84. Tag)	Varianzen sind gleich	0.015	0.905
	Varianzen sind nicht gleich		

Für den Levene-Test ergibt sich also beim Rattendiätbeispiel ein p-Wert von 0.9. Die Nullhypothese der Varianzgleichheit kann damit nicht verworfen werden, was aber nicht unbedingt bedeuten muss, dass die Varianzen in den dahinterliegenden Grundgesamtheiten wirklich gleich sind.

Damit sind wir beim großen Nachteil dieses Tests:

Bei kleinen Fallzahlen, wenn sein Ergebnis wirklich wichtig für uns wäre, führen selbst große Varianzunterschiede zu insignifikanten Resultaten. Bei großen Fallzahlen hingegen, wo unterschiedliche Varianzen kein wirkliches Problem mehr darstellen, werden kleinste und unbedeutendste Varianzunterschiede statistisch signifikant gesichert.

Daher: Daten immer graphisch ansehen.

Wenn man sich dann nicht ganz sicher ist, ob wirklich gleiche Varianzen vorliegen, dann verwendet man den t-Test mit ungleichen Varianzen.

Anmerkung: Andere Programmpakete (früher auch SPSS) verwenden oft anstelle des Levene-Tests einen F-Test um Varianzgleichheit zu prüfen. Dieser würde beim Rattendiäts-Beispiel 4.1.1. einen p-Wert von 0.9788 ergeben. Auch für den F-Test auf Varianzgleichheit gilt natürlich der oben erwähnte Nachteil.

Was tun, wenn keine Normalverteilung in den Gruppen vorliegt?

Wir wissen, in so einem Fall ist eine wichtige Voraussetzung für den t-Test nicht erfüllt.

Der Wilcoxon-Mann-Whitney U-Test bietet eine sinnvolle Alternative.

Wir sollten den t-Test aber nicht immer gleich aufgeben. Vor allem dann nicht, wenn wir durch simple Transformationen annähernde Normalverteilung erreichen können. Dies gilt insbesondere für Logarithmieren von rechtsschiefen Verteilungen.

Doch Vorsicht!

- Datentransformationen ändern die Interpretation der Ergebnisse.
- Nicht immer sind Transformationen erfolgreich.

Zusammenhang zwischen t-Test bei unabhängigen Stichproben und der Varianzanalyse

Dies ist ein Vorgriff auf später:

In vielen Fällen wollen wir nicht nur zwei, sondern mehr Gruppen miteinander vergleichen. Dies geschieht durch eine Varianzanalyse (englisch: *analysis of variance*, daher auch mit ANOVA abgekürzt).

Der t-Test ist nun die einfachste Form der einfaktoriellen Varianzanalyse.

- Ein "Faktor" ist ein auf einer nominalen Skala gemessenes Merkmal.
- "Einfaktoriell" bedeutet dabei, dass wir uns für die Bedeutung eines einzigen Faktors interessieren.

Zur Veranschaulichung nehmen wir wieder das Rattendiätbeispiel 4.1.1.

Um die einfaktorielle ANOVA durchzuführen, klicken wir auf

Analysieren

Mittelwerte vergleichen

Einfaktorielle ANOVA

5.1. Mehr zum t-Test bei unabhängigen Stichproben

Wir verschieben nun die Variable ZUNAHME ins Feld

Abhängige Variablen

und die Variable GRUPPE ins Feld

Faktor

Beachte: Neben letzterer scheinen diesmal keine Fragezeichen auf. Diese Methode ist nämlich nicht nur auf zwei Gruppen beschränkt.

Wir klicken auf

Ok

und die gewünschte einfaktorielle ANOVA wird berechnet.

ONEWAY ANOVA

Gewichtszunahme in Gramm (28.-84. Tag)

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	1596.000	1	1596.000	3.578	.076
Innerhalb der Gruppen	7584.000	17	446.118		
Gesamt	9180.000	18			

Neben für uns Unverständlichem finden wir Altbekanntes. Der p-Wert ("Signifikanz") von 0.076 ist der gleiche wie beim t-Test. Auch bei den Freiheitsgraden ("df") kommt uns der Wert 17 bekannt vor. Wenn wir die Wurzel aus dem "F"-Wert von 3.578 ziehen, dann erhalten wir den Wert der Teststatistik "T" beim t-Test, nämlich 1.891.

Eine ANOVA mit einem Zwei-Gruppen-Faktor entspricht also dem t-Test bei unabhängigen Stichproben und gleicher Varianz.

Wir können daher die Varianzanalyse gewissermaßen als Verallgemeinerung des t-Tests ansehen.

Damit sind aber auch die beim t-Test aufgetretenen Probleme bei der ANOVA vorhanden! (Anmerkung: Dazu kommen noch andere...)

5.2. Chi-Quadrat Test

Beispiel 5.2.1.: (binäre Zielgröße)

Therapievergleich von Dr. X: Standardtherapie wird mit neuer Therapie verglichen, Zielgröße ist binär (geheilt versus nicht geheilt)

Forschungsfrage: Gibt es Unterschiede in den Heilungsraten zwischen den beiden Therapien?

	Geheilt		
	Ja	Nein	
Standard Th.	4	12	16
Neue Therapie	9	9	18
	13	21	34

Standard Th. 25 % Erfolge

Neue Therapie 50 % Erfolge

Schlussfolgerung von Dr. X:

Die neue Therapie ist besser als die bisherige Standardtherapie!

Behauptung des Kollegen Y (offensichtlich ein unangenehmer, neidischer Mensch):

Diese Ergebnisse sind rein **zufällig** entstanden, in Wirklichkeit sind die Erfolgsraten bei beiden Therapien gleich!

Wieder taucht die Frage auf: **Was nun?** Offensichtlich müssen wir hier wieder einen statistischen Test anwenden. Allerdings ist der t-Test und der Wilcoxon Rangsummentest hier ungeeignet, da die Zielgröße binär ist.

Für den Vergleich von zwei Gruppen bei binären Zielgrößen kann man den sogenannten **Chi-Quadrat Test** verwenden.

Nullhypothese steht fest:

„Kein Unterschied im Heilungserfolg zwischen den beiden Therapien“

Berechnung:

Nimm an, Nullhypothese ist wahr. Berechne Wahrscheinlichkeit, dann das aktuelle Ergebnis (oder noch extremere Ergebnisse als das Aktuelle) zu beobachten.

Notwendig:

Intuitives Maß, welches „extremere Ergebnisse“ feststellt. Zum Beispiel das Pearson'sche Chi-Quadrat Kriterium. (Grundidee: Verwende quadrierte Unterschiede zwischen dem tatsächlich beobachteten Ergebnis und dem erwarteten Ergebnis, wenn die Nullhypothese wahr wäre.)

Wenn Nullhypothese wahr wäre:

Insgesamt gibt es 13 Heilungen von 34 Fällen, das sind 38.2 %

daher (wenn Nullhypothese wahr):

- 38.2 % erwartete Heilungen von 16 Standardtherapie-Patienten, das sind 6.1
- 38.2 % erwartete Heilungen von 18 Neue-Therapie-Patienten, das sind 6.9

Erwartet, wenn Nullhypothese gilt	Geheilt		
	Ja	Nein	
Standard Th.	6.1	9.9	16
Neue Therapie	6.9	11.1	18
	13	21	34

Beim Pearson'sches Chi-Quadrat Kriterium werden nun die tatsächlich beobachteten Werte den unter der Nullhypothese erwarteten Werten gegenübergestellt.

5.2. Chi-Quadrat Test

	Geheilt	
	Ja	Nein
Standard Th.	4 6.1	12 9.9
Neue Therapie	9 6.9	9 11.1

Tabelle: In den Zellen stehen links oben die tatsächlich beobachteten Werte, und rechts unten die erwarteten Werte, wenn die Nullhypothese wahr wäre.

Das **Pearson'sche Chi-Quadrat Kriterium** errechnet sich nun wie folgt:

$$\frac{(4-6.1)^2}{6.1} + \frac{(12-9.9)^2}{9.9} + \frac{(9-6.9)^2}{6.9} + \frac{(9-11.1)^2}{11.1} = 2.2$$

Idee dahinter:

- 1.) Beobachtet minus Erwartet (je größer dieser Differenzbetrag ist, eine desto schlechtere Übereinstimmung mit der Nullhypothese liegt vor).
- 2.) Davon das Quadrat (damit werden große Abweichungen stärker „bestraft“).
- 3.) Skalieren mit Erwartet (damit werden gleich große Abweichungen auf hohem Niveau nicht „so arg bestraft“ wie auf niedrigem Niveau).
Ein Beispiel um die dahinterliegende Idee zu verdeutlichen: Eine Abweichung von 7 bei erwarteten 1000 ist relativ gering im Vergleich zu einer Abweichung von 7 bei erwarteten 20.
- 4.) Summiere über alle vier Zellen.

Was ist ein extremes Ergebnis? Je größer der Wert für das Chi-Quadrat Kriterium ist, ein desto extremeres Ergebnis liegt vor.

Konkret:

Angenommen, es bestünde kein Therapie-Unterschied. Berechne die Wahrscheinlichkeit, dann einen Pearson Chi-Quadrat Wert von 2.2 oder größer zu beobachten.

Um das Therapievergleichsbeispiel 5.2.1. mit SPSS zu rechnen, legen wir zuerst zwei binäre Variablen an, THERAPIE und GEHEILT:

- THERAPIE dient zur Unterscheidung der Behandlungsgruppen (0=Standardtherapie, 1=neue Therapie)
- GEHEILT enthält als Zielgröße den Heilungserfolg (0=nein, 1=ja)

Zur Durchführung des Chi-Quadrat-Tests klicken wir auf

Analysieren

Deskriptive Statistiken

Kreuztabellen

Wir verschieben nun die Variable THERAPIE ins Feld

Zeilen

und die Variable GEHEILT ins Feld

Spalten

Wir klicken auf den Button

Statistik

und wählen den Chi-Quadrat Test aus. Dann klicken wir auf

Weiter

und

Ok

und der gewünschte Chi-Quadrat Test wird berechnet. Als erstes erscheinen Informationen über die Anzahl der tatsächlich verarbeiteten Fälle.

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
Therapie * Heilungserfolg	34	100.0 %	0	.0 %	34	100.0 %

Dann erscheint die 2x2-Tabelle, in der Therapie und Heilungserfolg gekreuzt werden.

Therapie * Heilungserfolg Kreuztabelle

Anzahl

		Heilungserfolg		Gesamt
		Nein	Ja	
Therapie	Standard	12	4	16
	Neu	9	9	18
Gesamt		21	13	34

5.2. Chi-Quadrat Test

Die letzte Tabelle enthält die verlangten Ergebnisse des Chi-Quadrat Tests. Die für uns wichtigen Teile wurden grau unterlegt.

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	2.242 ^b	1	0.134		
Kontinuitätskorrektur ^a	1.308	1	0.253		
Likelihood-Quotient	2.286	1	0.131		
Exakter Test nach Fisher				0.172	0.126
Zusammenhang linear mit-linear	2.176	1	0.140		
Anzahl der gültigen Fälle	34				

- a. Wird nur für eine 2x2-Tabelle berechnet
- b. 0 Zellen (0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 6.12.

Das Pearson'sche Chi-Quadrat Kriterium ist eine Teststatistik, die bei der 2x2-Kreuztabelle unter Gültigkeit der Nullhypothese asymptotisch eine Chi-Quadratverteilung mit einem Freiheitsgrad ("df") besitzt. Für den beobachteten Teststatistik-Wert "Wert" von 2.242 errechnet sich ein p-Wert ("Asymptotische Signifikanz (2-seitig)") von 0.134.

Das bedeutet: Wieder kann der Einwand des Kollegen Y nicht entkräftet werden!

Eine **Faustregel** warnt vor Problemen beim Chi-Quadrat-Test. Wenn eine der vier erwarteten Häufigkeiten unter der Nullhypothese kleiner als 5 ist, dann sollte auf den sogenannten Fisher's exact test ausgewichen werden. Diese Faustregel trifft bei unserem Beispiel aber nicht zu. SPSS unterstützt uns übrigens bei der Einhaltung dieser Faustregel, siehe Fußnote "b".

Mehr zum Fisher's exact test findet man im Anhang A.

5.3. Gepaarte Tests

Beispiel 5.3.1.: Die tägliche Nahrungsaufnahme von 11 jungen, gesunden Frauen wurde über einen längeren Zeitraum gemessen. Um jegliche willentliche Beeinflussung des Studienergebnisses zu vermeiden, wurde den Frauen vor Beginn der Studie nicht gesagt, dass der Zweck der Studie der Vergleich von prae- und post-menstrueller Nahrungsaufnahme wäre. Die durchschnittliche Nahrungsaufnahme (in kJ) über 10 prae- (PREMENS) und 10 post-menstruelle Tage (POSTMENS) einer jeden Probandin finden Sie in der folgenden Tabelle:

PROBANDIN	PREMENS	POSTMENS
1	5260	3910
2	5470	4220
3	5640	3885
4	6180	5160
5	6390	5645
6	6515	4680
7	6805	5265
8	7515	5975
9	7515	6790
10	8230	6900
11	8770	7335

Die Forschungsfrage lautet: Unterscheidet sich die praemenstruale von der postmenstrualen Nahrungsaufnahme?

Erste Idee: Verwende t-Test oder Wilcoxon-Mann-Whitney U-Test

Zweite Idee (eher eine Frage): Werden wir damit der Situation gerecht? Im Gegensatz zu den bisherigen 2-Gruppen-Vergleichen liegen diesmal 2 Messungen pro Individuum vor. Wir stehen also vor einer Situation mit Abhängigkeiten. Wir nennen dies auch eine **gepaarte** Situation. Zwei-Gruppen-Vergleiche, wo pro Individuum nur eine einzige Messung vorliegt, werden als **ungepaart** bezeichnet.

Grundsätzlich ist es erlaubt, auch bei gepaarten Situationen die bisher bekannte *ungepaarte* Variante des t-Tests oder den Wilcoxon-Mann-Whitney U-Test zu verwenden. Da aber Messungen innerhalb eines Individuums üblicherweise ähnlicher als Messungen zwischen Individuen sind, führt die Berücksichtigung der gepaarten Situation in der Analyse zu größerer Mächtigkeit.

Die Frage „Unterscheidet sich die praemenstruale von der postmenstrualen Nahrungsaufnahme?“ ist äquivalent zur Frage „Ist die mittlere Differenz zwischen prae- und postmenstrueller Nahrungsaufnahme ungleich Null?“

Der dafür adäquate Test ist der gepaarte t-Test. Wir gehen vor, wie gehabt (nach dem Prinzip des statistischen Testens!)

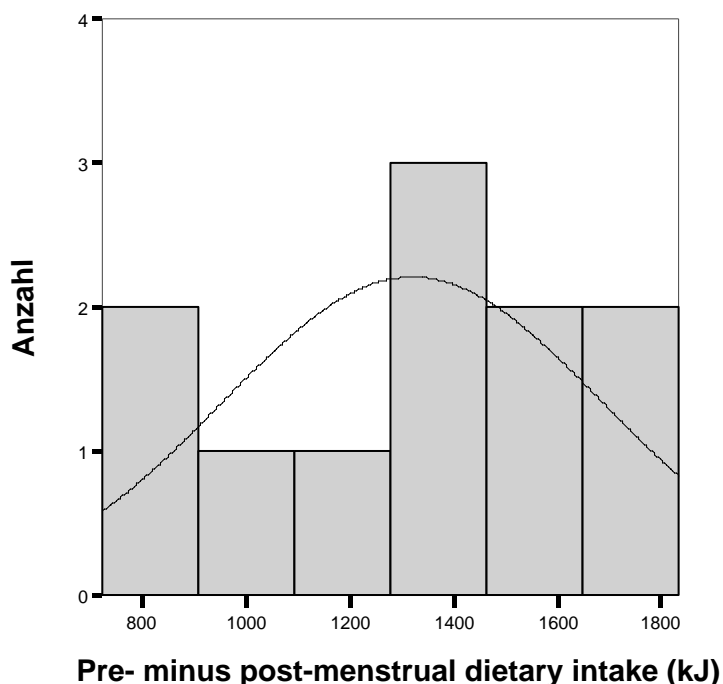
Konkret:

- **Nullhypothese:** „Die mittlere Differenz zwischen prae- und postmenstrueller Nahrungsaufnahme ist gleich Null!“
(also: kein Effekt)
- **Zweiseitige Alternativhypothese:** „Die mittlere Differenz ist ungleich Null!“
- **Intuitives Maß für Abstand zur Nullhypothese:** Absolut-Betrag der beobachteten mittleren Differenz
- **Wann ist ein Ergebnis extremer als das Beobachtete?** Offenbar wenn der Absolut-Betrag größer als das beobachtete Ergebnis ist

Wenn die Verteilung der einzelnen Differenzen approximative Normalverteilung zeigt, dann ist der Einsatz des gepaarten t-Tests angebracht. Die mittlere Differenz muss noch durch ihre geschätzte Streuung dividiert werden, um eine t-verteilte Größe zu erhalten. Damit kann dann der p-Wert berechnet werden.

Vorgehen in SPSS:

Zuerst prüfen wir graphisch, ob die Verteilung der Differenzen zwischen den beiden Zeitpunkten annähernd normalverteilt ist. (Um es klar herauszustreichen: Die Differenzen PREMENS-POSTMENS sollten annähernd normalverteilt sein! Die Verteilungen der Ausgangswerte PREMENS und POSTMENS sind hier nicht von Interesse.)



5.3. Gepaarte Tests

Wir akzeptieren dies als nicht allzu große Abweichung von der Normalverteilung. Damit können wir nun den gepaarten t-Test durchführen. Dazu klicken wir auf

Analysieren

Mittelwerte vergleichen

T-Test bei gepaarten Stichproben

Wir klicken dann die Variablen PREMENS und POSTMENS an. Diese scheinen daraufhin im Feld

Aktuelle Auswahl

als

Variable 1 und Variable 2

auf. Dann verschieben wir diese ins Feld

Gepaarte Variablen

und es erscheint die Differenz PREMENS--POSTMENS.

Wir klicken dann auf

Ok

und der gewünschte gepaarte t-Test wird berechnet. Nach einer deskriptiven Tabelle und einer Tabelle mit einem Korrelationskoeffizient erhalten wir schließlich das Ergebnis.

Test bei gepaarten Stichproben

		Gepaarte Differenzen				
		Mittelwert	Standardabweichung	Standardfehler des Mittelwertes	95 % Konfidenzintervall der Differenz	
					Untere	Obere
Paaren 1	Pre-menstrual dietary intake (kJ) - Post-menstrual dietary intake (kJ)	1320.45	366.746	110.578	1074.07	1566.84

Der Mittelwert und die Standardabweichung der gepaarten Differenzen kann hier abgelesen werden. Anhand des 95 % Konfidenzintervalls kann ersehen werden, dass die Nahrungsaufnahme im weiblichen Zyklus statistisch signifikant schwankt (Null wird nicht überdeckt).

Test bei gepaarten Stichproben

		T	df	Sig. (2-seitig)
Paaren 1	Pre-menstrual dietary intake (kJ) - Post-menstrual dietary intake (kJ)	11.941	10	.000

Der p-Wert ist <0.001 . Er basiert auf einem zweiseitigen Test. Er bestätigt damit nur, was wir schon vom zweiseitigen Konfidenzintervall abgelesen haben.

Gibt es auch eine nichtparametrische Alternative zum gepaarten t-Test? Es gibt sogar zwei, den **Wilcoxon Vorzeichen-Rangtest** und den **Vorzeichen-Test**.

Das Problem mit dem Wilcoxon Vorzeichen-Rangtest ist, dass sein Resultat von Datentransformationen abhängig sein kann! Dies ist völlig unüblich für einen nichtparametrischen Test. Der Wilcoxon Vorzeichen-Rangtest sollte daher nur bei einer symmetrischen Verteilung der Differenzen verwendet werden.

Manche Statistiker empfehlen die Verwendung des Vorzeichen-Test als nichtparametrische Alternative zum gepaarten t-Test. Der Vorzeichen-Test ist aber nicht sehr mächtig.

Andere Statistiker empfehlen, die Daten zu transformieren, wenn die Rohdaten größere Differenzen für größere Ausgangswerte zeigen, und dann den Wilcoxon Vorzeichen-Rangtest zu verwenden. Dies ist aber nur dann sinnvoll, wenn die Transformation zu einer symmetrischen Verteilung der Differenzen führt.

Wie sehen wir, ob größere Differenzen bei größeren Ausgangswerten vorliegen? Durch das Plotten von Differenz $(X-Y)$ versus Mittelwert $(X+Y)/2$.

Um den Wilcoxon Vorzeichen-Rangtest und den Vorzeichen-Test beim Beispiel 5.3.1. durchzuführen, klicken wir auf

Analysieren

Nichtparametrische Tests

Zwei verbundene Stichproben

Wir klicken dann die Variablen PREMENS und POSTMENS an. Diese scheinen daraufhin im Feld

Aktuelle Auswahl

als

Variable 1 und Variable 2

auf. Dann verschieben wir diese ins Feld

Ausgewählte Variablenpaare

und es erscheint die Differenz PREMENS--POSTMENS. Im Feld

Welche Tests durchführen?

wählen wir

Wilcoxon und Vorzeichen

aus. Dann klicken wir auf

Ok

und die beiden gewünschten Tests werden berechnet.

Für den Wilcoxon Vorzeichen-Rangtest ergibt sich ein p-Wert von 0.003, und für den Vorzeichen-Test ergibt sich ein p-Wert von 0.001.

Anmerkungen:

- Gepaarte Situationen sind in der Medizin nicht nur auf zwei Messungen innerhalb eines Patienten beschränkt. Z.B. auch bei einer gematchten Fall-Kontroll-Studie liegt eine gepaarte Situation vor. Die einzelnen Differenzen werden dann aus den beiden Beobachtungen eines jeden gematchten Paares (=Fall und seine Kontrolle) ermittelt. Achtung! So einfach gematchte Fall-Kontroll-Studien auch auszuwerten sind, so große Gefahren bergen sie in sich. Insbesondere die Auswahl der Kontrollen ist eine stete Quelle für verzerrte bzw. falsche Aussagen.
- Gepaarte Situationen sind der einfachste Fall für eine *Blockbildung*. Der Begriff *Block* kommt aus der Landwirtschaft. Jede Versuchseinheit, bei der mehr als einmal gemessen wurde, bildet demnach einen *Block*.

Beispiel 5.3.2.: Im Jahre 1968 wurde an der damaligen I. Chirurgischen Universitätsklinik (AKH, Wien) eine Diagnosestudie zum Thema Gefäßverschluss bei 121 Patienten durchgeführt. Die Diagnoseverfahren Klinisch und Doppler wurden verglichen. Der wahre Befund wurde damals mittels Venographie erstellt. Vorerst werden nur Patienten mit Gefäßverschluss betrachtet.

Klinisch	Doppler	Anzahl der Patienten mit Verschluss
+	+	22
+	-	3
-	+	16
-	-	3
		Summe 44

In 25 Fällen (22 plus 3) wurde klinisch ein vorliegender Verschluss korrekt diagnostiziert. Die korrekte Diagnose beim Doppler-Verfahren wurde in 38 Fällen gestellt (22 plus 16). Damit beträgt die Sensitivität für die klinische Diagnose $25/44=57\%$. Die Sensitivität für Doppler beträgt $38/44=86\%$.

Offenbar ist die Diagnose mit Doppler sensitiver als die klinische Diagnose. Die Frage ist nun, ob dieser Unterschied noch mit dem Zufall vereinbar ist? Wir benötigen also wieder einen statistischen Test.

Von der Struktur her sind die Beispiele 5.3.1. und 5.3.2. sehr ähnlich. Im ersten Fall wurde die Zielgröße pro Probandin jeweils zweimal gemessen (prae-, postmentstrual), nun wird die Zielgröße pro Patient ebenfalls jeweils zweimal gemessen (Klinisch, Doppler). Der wesentliche Unterschied liegt im Skalenniveau. Beim Beispiel 5.3.1. war die Zielgröße metrisch (Nahrungsaufnahme in kJ), nun ist sie dichotom (+, -).

Beim Beispiel 5.3.1. war der korrekte statistische Test letztendlich nichts anderes als die gepaarte Version des simplen ungepaarten t-Tests. Analog dazu ist der korrekte statistische Test für das Beispiel 5.3.2. eine gepaarte Version des simplen Chi-Quadrat-Tests, der sogenannte **McNemar-Test**.

Die Analogie geht weiter. Beim gepaarten t-Test wird nur die Differenz aus zwei gepaarten Messwerten verwendet. Beim McNemar-Test wird ebenfalls die Information aus zwei gepaarten Messwerten auf jeweils einen Wert reduziert.

Wir merken uns: Der McNemar-Test ist für gepaarte Situationen mit dichotomer Zielgröße gedacht. Er ist eine Variante des Vorzeichen-Tests.

Im Folgenden verwenden wir die Datei b5_3_2.sav. Dabei dürfen wir auf „Fälle gewichten“ nicht vergessen!

5.3. Gepaarte Tests

Wenn wir die Diagnoseverfahren Klinisch und Doppler kreuzen, dann ergibt sich die folgende 2x2-Tabelle:

		Diagnose mittels Doppler-Verfahrens		Gesamt
		-	+	
Klinische Diagnose	-	3	16 (=f)	19
	+	3 (=g)	22	25
Gesamt		6	38	44

Bei 3 Patienten führen beide Verfahren zu einer negativen, und bei 22 Patienten führen beide Verfahren zu einer positiven Diagnose. Solche gleichlautenden Ergebnisse werden als **konkordant** bezeichnet.

Für die Frage, welches Diagnoseverfahren sensitiver ist, sind konkordante Ergebnisse unwichtig. Der Vergleich der Sensitivitäten $25/44$ und $38/44$ reduziert sich nämlich auf den Vergleich der Anzahl an jeweils korrekten Diagnosen, 25 und 38. Die 22 konkordanten Ergebnisse kommen dabei in beiden Anzahlen vor, was schlussendlich auf einen Vergleich der sogenannten diskordanten Ergebnisse, 3 und 16 hinausläuft (in der Tabelle mit "g" und "f" bezeichnet).

Unterschiedliche Diagnosen sind demnach **diskordante** Ergebnisse. Nur Patienten mit diskordanten Diagnoseergebnissen tragen relevante Information zur Beantwortung der Forschungsfrage bei. Und genau hier setzt der McNemar-Test an. Wir gehen vor, wie gehabt (Prinzip des statistischen Testens!)

Konkret:

- **Nullhypothese:** „Die Differenz zwischen den beiden Arten von diskordanten Ergebnissen (f minus g) ist gleich Null!“
(also: beide Diagnoseverfahren sind gleich sensitiv)
- **Zweiseitige Alternativhypothese:** „Diese Differenz ist ungleich Null!“
- **Intuitives Maß für Abstand zur Nullhypothese:**
Chi-Quadrat-Kriterium
- **Wann ist ein Ergebnis extremer als das Beobachtete?**
Je größer das Chi-Quadrat-Kriterium, desto extremer das Ergebnis!

Wenn die Nullhypothese wahr wäre:

19 diskordante Paare wurden beobachtet. Wenn wirklich Klinisch und Doppler gleich sensitiv wären, dann würden wir

- je 9.5 mal Klinisch „-“ und Doppler „+“ und
- je 9.5 mal Klinisch „+“ und Doppler „-“

erwarten. Dies müssen wir nur mehr noch ins Chi-Quadrat-Kriterium einsetzen. Erinnerung: (Beobachtet minus Erwartet) zum Quadrat, dann dividiert durch Erwartet, dann summieren. Im Gegensatz zum bereits bekannten "gewöhnlichen" Chi-Quadrat-Test werden beim McNemar-Test nur die beiden diskordanten Zellen verwendet:

$$\frac{(16-9.5)^2}{9.5} + \frac{(3-9.5)^2}{9.5} = 8.9$$

Man kann diese Formel auch vereinfachen und erhält: $\frac{(f-g)^2}{f+g}$

Wenn Sie für $f=16$ und $g=3$ einsetzen, dann liefert auch die vereinfachte Formel einen Wert von 8.9. Und natürlich gibt es auch eine exakte Variante des McNemar-Tests, die besonders bei kleinen Werten von f und g empfohlen wird.

Grundsätzlich bietet SPSS zwei Möglichkeiten, um den McNemar-Test zu berechnen.

1. Möglichkeit: Dazu klicken wir auf

Analysieren

Nichtparametrische Tests

Zwei verbundene Stichproben

Wir klicken dann die Variablen KLINISCH und DOPPLER an. Diese scheinen daraufhin im Feld

Aktuelle Auswahl

als

Variable 1 und Variable 2

auf. Diese verschieben wir ins Feld

Ausgewählte Variablenpaare

und es erscheint die Differenz KLINISCH--DOPPLER.

Im Feld

Welche Tests durchführen?

wählen wir

McNemar

aus. Dann klicken wir auf

Ok

und der gewünschte Test wird berechnet. SPSS gibt hier automatisch den exakten p-Wert von 0.004 an.

2. Möglichkeit: Dazu klicken wir auf

Analysieren

Deskriptive Statistiken

Kreuztabellen

Wir verschieben nun die Variable KLINISCH ins Feld

Zeilen

und die Variable DOPPLER ins Feld

Spalten

Wir klicken auf den Button

Statistik

und wählen den McNemar Test aus. Dann klicken wir auf

Weiter

und

Ok

und der gewünschte Test wird berechnet. Als exakter p-Wert ergibt sich wieder 0.004.

Am Ende fehlt uns noch die Interpretation des Testergebnisses. Der p-Wert von 0.004 ist kleiner als das in der Medizin übliche Signifikanzniveau von 0.05. Das Testergebnis ist daher statistisch signifikant, und wir können die Nullhypothese verwerfen. Demnach ist die Diagnose eines Gefäßverschlusses mit dem Doppler-Verfahren sensitiver als mit dem klinischen Verfahren.

5.4. Einseitige versus zweiseitige Tests

Zur Wiederholung:

Der p-Wert ist die Wahrscheinlichkeit, ein zumindest so extremes Resultat wie das Beobachtete zu erhalten. Dabei wird angenommen, dass die Nullhypothese gilt.

Es ist klar, wenn die Nullhypothese gilt, dann treten extreme Ergebnisse zufällig in jede Richtung gleich oft auf.

Konkret beim Rattenbeispiel: Wenn kein Unterschied in der Gewichtszunahme zwischen den beiden Proteindiäten bestünde, und wir den Versuch öfters wiederholen könnten, dann hätten einmal die Ratten mit dem schwach proteinhaltigen Futter und einmal die mit dem stark proteinhaltigen Futter stärker zugenommen. Rein zufällig eben.

Wir berücksichtigen diesen Umstand, in dem wir **zweiseitige Tests** durchführen und damit **zweiseitige p-Werte** errechnen. Das haben wir auch bis jetzt gemacht. In der Mehrzahl der Fälle wird dies auch die einzig richtige Vorgangsweise sein.

In seltenen Fällen kann es Forschungsfragen geben, bei denen substanzwissenschaftlich sinnvolle Unterschiede nur in eine Richtung auftreten können. Beobachtet man beim Experiment eine Differenz in die andere Richtung, dann würde man das immer dem Zufall zuschreiben - egal wie groß dieser Unterschied dann auch wäre. Ein typisches Beispiel dafür sind Dosis-Wirkungsbeziehungen in der Toxikologie - eine Erhöhung der Dosis kann zu keiner Verminderung der Toxizität führen.

In so einem Fall würde man die Alternativhypothese auf einen Effekt in nur eine Richtung beschränken.

Eine mögliche einseitige Alternativhypothese beim Rattenbeispiel könnte lauten: "Schwach proteinhaltiges Futter verursacht mehr Gewichtszunahme als stark proteinhaltiges Futter"

Wir würden also einen **einseitigen t-Test** durchführen und einen **einseitigen p-Wert** errechnen.

Einseitige Tests sind selten wirklich passend. Selbst wenn wir die starke Vermutung hätten, dass die neue Behandlung nicht schlechter als die bisherige Standardbehandlung sein könnte, könnten wir trotzdem nicht sicher sein. Wenn wir nämlich sicher wären, bräuchten wir kein Experiment mehr durchzuführen!!

5.4. Einseitige versus zweiseitige Tests

Wenn die Überzeugung besteht, dass ein einseitiger Test wirklich passend ist, dann muss die Entscheidung fallen, **bevor** die Daten analysiert werden. Besser noch, bevor die Daten erhoben werden. Bei prospektiven Studien vermerkt man daher die beabsichtigte Teststrategie (und die substanzwissenschaftlichen Gründe dafür) im Studienprotokoll, um unangenehme Diskussionen im nachhinein zu vermeiden.

Auf keinen Fall darf die Entscheidung für einen einseitigen Test vom Resultat des Experiments bzw. der Studie abhängen.

Wenn in der medizinischen Literatur einseitige Testergebnisse angegeben werden, dann liegen die p-Werte üblicherweise zwischen 0.025 und 0.05. Dies bedeutet: Ein zweiseitiger Test wäre nicht signifikant gewesen! Man kann wohl in vielen Fällen zu Recht annehmen, dass es keine im vorhinein festgelegten einseitigen Hypothesen gegeben hat.

Durch eine einseitige Alternativhypothese, die vom Resultat des Experiments abhängt, wird das angegebene Signifikanzniveau nicht eingehalten. Anstatt zum Beispiel 5 % wäre dann das tatsächliche Signifikanzniveau doppelt so hoch, nämlich 10 %.

- Verwenden Sie grundsätzlich zweiseitige Tests!
- Verwenden Sie einseitige Tests nur, wenn Sie dies vor Beginn der Studie geplant und die substanzwissenschaftlichen Gründe dafür schriftlich im Studienprotokoll festgehalten haben!
- Seien Sie gegenüber Studien, die einseitige Testergebnisse ohne vernünftige Begründung angeben, grundsätzlich misstrauisch!

5.5. Übungen

- 5.5.1. Bei einer Befragung von 16-jährigen Jugendlichen bezeichneten sich 27 von 40 Burschen und 12 von 32 Mädchen als Raucher. In der Datei b5_5_1.sav finden Sie die aggregierten Daten.

Ist das Rauchverhalten der Jugendlichen von ihrem Geschlecht abhängig?

- 5.5.2. In der Datei b5_5_2.sav finden Sie die Werte für Thyroxin im Serum (nmol/l) bei 16 Kindern mit Hypothyreose unterschieden nach Stärke der Symptome („keine bis geringe Symptome“ versus „ausgeprägte Symptome“).

Keine bis geringe Symptome:

34, 45, 49, 55, 58, 59, 60, 62, 86

Ausgeprägte Symptome:

5, 8, 18, 24, 60, 84, 96

Vergleichen Sie die Thyroxin-im-Serum-Werte zwischen den beiden Symptomgruppen.

- 5.5.3. In der Datei b5_5_3.sav finden Sie die Daten für zwei Patientengruppen. Die erste Gruppe (Hodgkin=1) besteht aus Hodgkin-Patienten in Remission. Die zweite Gruppe (Hodgkin=2) ist eine adäquate Vergleichsgruppe von Non-Hodgkin-Patienten ebenfalls in Remission.

Die Anzahl der T_4 und T_8 Zellen/mm³ Blut wurde ermittelt.

- (a) Gibt es Unterschiede im Auftreten von T_4 Zellen zwischen den beiden Gruppen?
- (b) Gibt es Unterschiede im Auftreten von T_8 Zellen zwischen den beiden Gruppen?

- 5.5.4. Rechtsschief verteilte Daten werden gerne logarithmisch transformiert. Der Grund dafür ist simpel: Wenn die logarithmierten Werte in beiden Gruppen annähernd normalverteilt sind, dann wird mit diesen logarithmierten Werten ein t-Test durchgeführt.

Was bedeutet dies für die Interpretation des Ergebnisses auf der nicht-logarithmierten Originalskala?

Anleitung:

Mit welcher mathematischen Operation kommen Sie von einer logarithmischen Skala zurück auf die Originalskala?

Was bewirkt die Anwendung besagter mathematischer Operation auf eine Differenz?

Was bewirkt die Anwendung besagter mathematischer Operation auf die Nullhypothese beim t-Test?

- 5.5.5. Ändert sich aufgrund Ihrer Überlegungen beim Beispiel 5.5.4. Ihre Analyse des Beispiels 5.5.3.?

Anleitung:

Versuchen Sie beim Beispiel 5.5.3. eine logarithmische Transformation der Zellen-Anzahlwerte.

Scheint danach die Verwendung von t-Tests sinnvoll zu sein?

Wenn ja, führen Sie die t-Tests durch und interpretieren Sie deren Ergebnisse.

Wenn nein, begründen Sie Ihre Antwort ausführlich.

- 5.5.6. Überspielen Sie die Daten des Rattendiäts-Beispiels 4.1.1. von SPSS nach EXCEL. Führen Sie dort einen t-Test mit gleicher und ungleicher Varianz sowie einen Test auf gleiche Varianzen durch.

- 5.5.7. Verwenden Sie die Daten aus Beispiel 5.3.1. und bilden Sie die Variable DIFF aus der Differenz von PREMENS minus POSTMENS. Klicken Sie auf

Analysieren

Mittelwerte vergleichen

T-Test bei einer Stichprobe

und verschieben Sie die Variable DIFF ins Feld Testvariable(n).

Klicken Sie dann auf Ok.

(a) Was fällt Ihnen auf im Vergleich zum gepaarten t-Test? Haben Sie eine Erklärung dafür?

(b) Können Sie herausfinden, was das Feld Testwert bedeutet? Warum steht da eine Null? Was passiert, wenn Sie diesen Wert abändern?

5.5. Übungen

- 5.5.8. Die folgende Studie wurde bei insgesamt 173 Patienten mit Hautkrebs durchgeführt. Die Hautreaktion der Patienten auf das Kontaktallergen Dinitrochlorobenzol (DNCB) als auch auf das hautreizende, entzündungsfördernde Krotonöl wurde erhoben:

+ve ... Hautreaktion vorhanden
 -ve keine Hautreaktion vorhanden

Der Zweck dieser Untersuchung war die Klärung der Frage, ob bei den Patienten mit Hautkrebs der Kontakt zu DNCB unterschiedlich oft Hautreaktionen als der Kontakt zu Krotonöl hervorruft.

Hier sind die Ergebnisse der Studie:

Hautreaktion auf		Anzahl
DNCB	Krotonöl	
+ve	+ve	81
+ve	-ve	23
-ve	+ve	48
-ve	-ve	21

Führen Sie die entsprechenden Auswertungen durch.

- 5.5.9. Hier sind die vollständigen Daten für die Diagnosestudie zum Thema Gefäßverschluss (Beispiel 5.3.2). Die Diagnoseverfahren *Klinisch* und *Doppler* wurden verglichen. Der *wahre* Befund wurde damals mittels Venographie erstellt.

Klinisch	Doppler	Anzahl der Patienten	
		mit Verschluss	ohne Verschluss
+	+	22	4
+	-	3	27
-	+	16	5
-	-	3	41
		Summe 44	Summe 77

Vergleichen Sie die Spezifitäten der beiden Diagnoseverfahren.

- 5.5.10. Um festzustellen, ob ein Hodenhochstand (maldescensus testis) bei Geburt zu einem erhöhten Hodenkrebsrisiko führt, wurde eine gematchte Fall-Kontrollstudie durchgeführt. Als Fälle wurden 259 Hodenkrebspatienten identifiziert. Zu jedem Fall wurde innerhalb des gleichen Spitals ein Kontrollpatient gesucht (dazugematcht), der gleich alt (± 2 Jahre) war, zur selben ethnischen Gruppe gehörte, und an einer anderen Krankheit als Hodenkrebs litt. In der Datei b5_5_10 finden Sie die entsprechenden Daten. Beantworten Sie damit die Forschungsfrage.

Noch eine Anmerkung: Hier müssen wir mit der Begriffsbezeichnung „Fall“ vorsichtig umgehen, da zwei verschiedene Begriffe dahinter stecken:

- (i) Ein Fall in der Fall-Kontrollstudie ist ein Hodenkrebspatient.
- (ii) Ein Fall in der Datenmatrix ist ein **Paar**, welches aus einem Hodenkrebspatienten und dem dazugematchten Kontrollpatienten besteht.

Anhänge

D. Exakte Tests

Wiederholung - Beispiel 5.2.1.: Therapievergleich von Dr. X: Standardtherapie wird mit neuer Therapie verglichen, Zielgröße ist binär (geheilt versus nicht geheilt). Die Forschungsfrage lautet: Gibt es Unterschiede in den Heilungsraten zwischen den beiden Therapien?

	Geheilt		
	Ja	Nein	
Standard Th.	4	12	16
Neue Therapie	9	9	18
	13	21	34

Dazu wurde ein Chi-Quadrat Test gerechnet:

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	2.242 ^b	1	0.134		
Kontinuitätskorrektur ^a	1.308	1	0.253		
Likelihood-Quotient	2.286	1	0.131		
Exakter Test nach Fisher				0.172	0.126
Zusammenhang linear mit-linear	2.176	1	0.140		
Anzahl der gültigen Fälle	34				

- Wird nur für eine 2x2-Tabelle berechnet
- 0 Zellen (0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 6.12.

D. Exakte Tests

Wir wissen bereits, dass das Pearson'sche Chi-Quadrat Kriterium eine Teststatistik ist, die bei der 2x2-Kreuztabelle unter Gültigkeit der Nullhypothese asymptotisch eine Chi-Quadratverteilung mit einem Freiheitsgrad ("df") besitzt.

Für den beobachteten Teststatistik-Wert "Wert" von 2.242 errechnet sich ein p-Wert von 0.134.

Zwei Dinge sind aber auffallend, zum einen betrifft es die Begriffe "**Asymptotische Signifikanz**" und "**Exakte Signifikanz**", zum anderen die Anmerkung "**0 Zellen (.0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 6.12.**"

Das wollen wir uns jetzt näher ansehen. Dazu rufen wir uns als erstes noch einmal die erwarteten Patientenzahlen unter Gültigkeit der Nullhypothese in Erinnerung:

Erwartet, wenn Nullhypothese gilt	Geheilt		
	Ja	Nein	
Standard Th.	6.1	9.9	16
Neue Therapie	6.9	11.1	18
	13	21	34

D. Exakte Tests

Wenn wir beim Beispiel 5.2.1. nur die sogenannten Randverteilungen wissen würden, dann wären grundsätzlich die folgenden 14 verschiedenen 2×2-Kreuztabellen als beobachtbares Ergebnis möglich:

0	16	16
13	5	18
13	21	34

7	9	16
6	12	18
13	21	34

1	15	16
12	6	18
13	21	34

8	8	16
5	13	18
13	21	34

2	14	16
11	7	18
13	21	34

9	7	16
4	14	18
13	21	34

3	13	16
10	8	18
13	21	34

10	6	16
3	15	18
13	21	34

4	12	16
9	9	18
13	21	34

11	5	16
2	16	18
13	21	34

5	11	16
8	10	18
13	21	34

12	4	16
1	17	18
13	21	34

6	10	16
7	11	18
13	21	34

13	3	16
0	18	18
13	21	34

Beachte dabei: Es genügt bei fixen Randverteilungen nur die Angabe eines Feldes dieser 2×2-Kreuztabelle. Wir nehmen die linke obere Ecke (das entspricht der Anzahl der Geheilten unter der Standardtherapie).

D. Exakte Tests

Angenommen, die Nullhypothese (=beide Therapien wirken gleich gut) gilt, dann können wir uns die Wahrscheinlichkeiten ausrechnen, zufällig eine der möglichen Kreuztabellen zu beobachten. Für Interessierte: Diese Wahrscheinlichkeiten werden unter Verwendung der hypergeometrischen Verteilung errechnet.

X	Pearson'sches Chi-Quadrat Kriterium	Wahrscheinlichkeit (unter Annahme der Gültigkeit der Nullhypothese)
0	18.7	0.0000092
1	13.1	0.0003201
2	8.5	0.0041152
3	4.9	0.0264062
4	2.2	0.0953555
5	0.62	0.2059680
6	0.0069	0.2746240
7	0.39	0.2288533
8	1.8	0.1188277
9	4.2	0.0377231
10	7.5	0.0070416
11	11.9	0.0007202
12	17.3	0.0000353
13	23.7	0.0000006

X steht dabei für die Eintragung in der linken oberen Ecke der Kreuztabelle. Wir sehen, dass X=5, 6 und 7 am wahrscheinlichsten sind. Dies ist kein Wunder, denn die dazugehörigen Kreuztabellen entsprechen am ehesten der erwarteten Kreuztabelle unter der Nullhypothese.

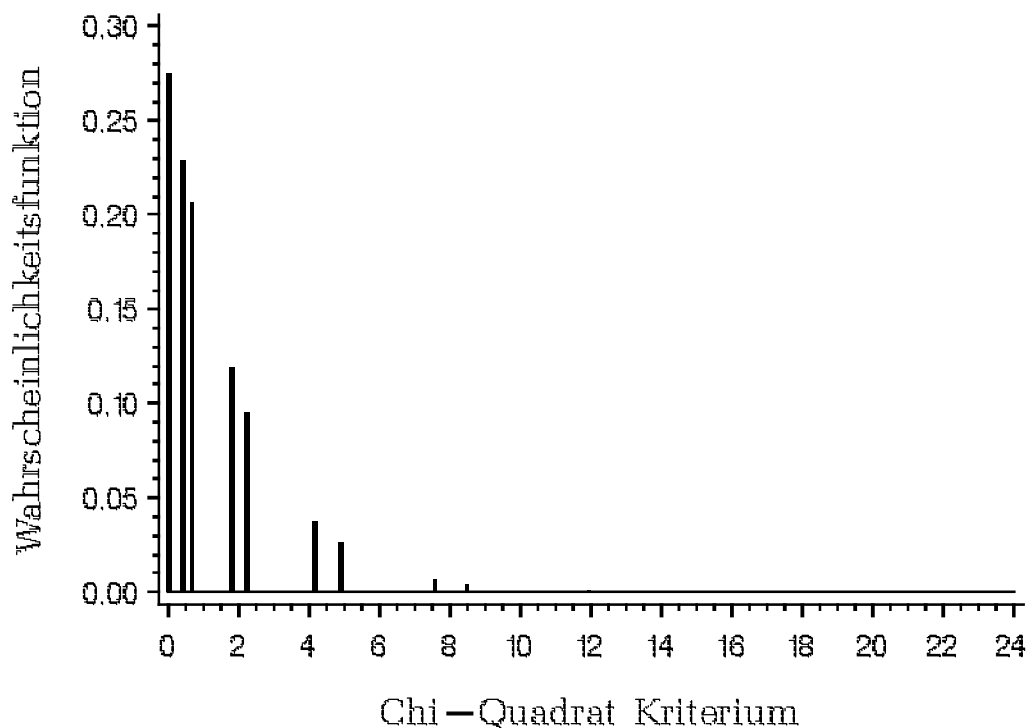
In der mittleren Spalte wurde auch das Pearson'sche Chi-Quadrat Kriterium für jede mögliche Kreuztabelle angegeben. Für X=6 ergibt sich der kleinste Wert, für X=7 der zweitkleinste, usw.

Im Beispiel 5.2.1. haben wir eine Kreuztabelle mit X=4 beobachtet. Wir können nun die oben dargestellte Wahrscheinlichkeitsverteilung benutzen, um einen **exakten p-Wert** zu ermitteln. Dazu müssen wir die Wahrscheinlichkeiten aller Kreuztabellen addieren, deren Chi-Quadrat Kriterium gleich oder größer als 2.2 ist (grau markiert). Wir erhalten einen exakten p-Wert von 0.1717. Diese Vorgangsweise ist auch unter dem Namen **Fisher's exakter Test** bekannt.

Der SPSS-Output zum Beispiel 5.2.1. bestätigt unsere Rechnung.

Anmerkung: Der SPSS-Output gibt für das Beispiel 5.2.1. auch eine "exakte 1-seitige Signifikanz" von 0.126 für den exakten Test von Fisher an. Aufgrund unserer Berechnungen können wir nachvollziehen, wie SPSS auf diesen Wert kommt: Konkret werden die Wahrscheinlichkeiten von $X=0$ bis 4 aus der Tabelle von Seite 75 addiert.

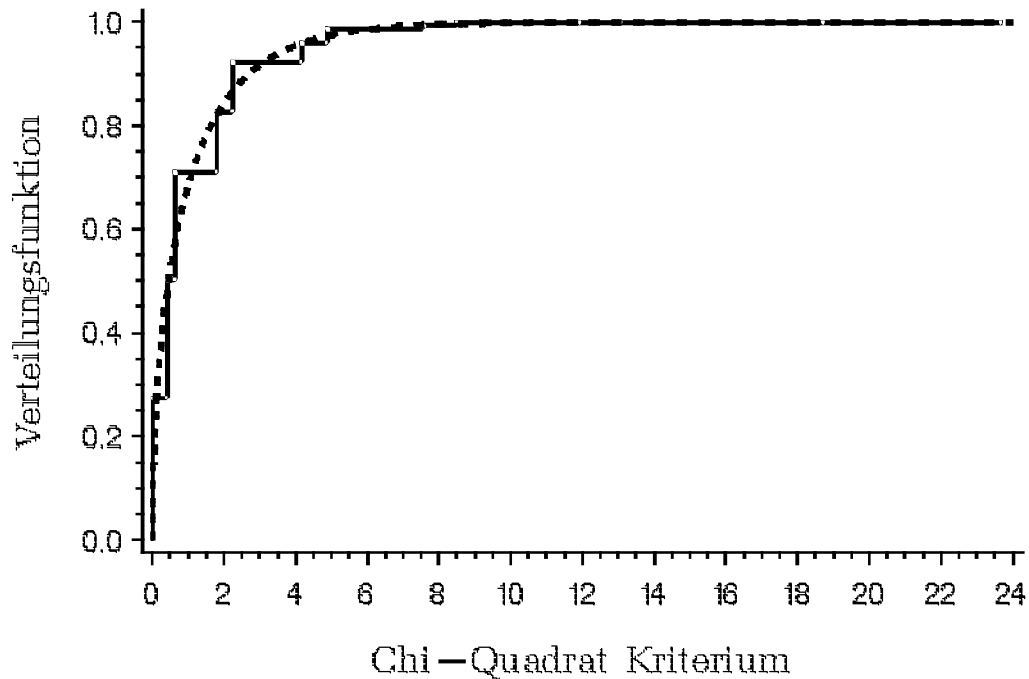
Beachte, die einseitige Alternativhypothese wird hier anhand der beobachteten Daten automatisch generiert und getestet. Das ist eine unsaubere wissenschaftliche Vorgangsweise. D.h., 1-seitige p-Werte sollten nicht in dieser Form automatisch angegeben werden.



Graphische Darstellung der Wahrscheinlichkeitsfunktion des Chi-Quadrat Kriteriums beim Beispiel 5.2.1. unter der Annahme, dass die Nullhypothese gilt.

Es fragt sich nun, warum wir beim Beispiel 5.2.1. einen asymptotischen p-Wert beim Chi-Quadrat Test von 0.134 angegeben haben, wenn der exakte p-Wert 0.1717 beträgt?

Vorerst: Der asymptotische p-Wert (der auf der Chi-Quadrat Verteilung beruht) ist eine Approximation des exakten p-Werts. Wie fast alle Approximationen in der Statistik wird sie mit zunehmendem Stichprobenumfang immer besser, das heißt genauer.



Graphische Darstellung der asymptotischen (gestrichelte Linie) und exakten (durchgezogene Linie) Verteilungsfunktionen des Chi-Quadrat Kriteriums beim Beispiel 5.2.1. unter der Annahme, dass die Nullhypothese gilt.

Diese Approximation der exakten durch die asymptotische Verteilung scheint ganz gut zu sein. Trotzdem, wenn wir den exakten p-Wert zur Verfügung haben, warum verwenden wir dann den asymptotischen?

Antwort: Obwohl die Begriffsbezeichnung "exakt" den Eindruck vermittelt, hier handelt es sich um die *bessere*, weil genauere Variante des Tests, muss man zwei Dinge wissen. Erstens, exakte Tests erfordern im allgemeinen Spezialprogramme und leistungsfähige Computer. Zweitens, exakte Tests gelten aufgrund der diskreten Teststatistik als konservativ. Konservativität eines statistischen Tests bedeutet, dass der Test zu lange an der Nullhypothese festhält. Oder anders formuliert, die zugestandene Fehlerwahrscheinlichkeit (das Signifikanzniveau) wird nicht vollständig ausgenützt.

Als Faustregel für die Entscheidung zwischen asymptotischer und exakter Variante beim Chi-Quadrat Test wird gerne die folgende verwendet:

- Wenn alle unter der Nullhypothese erwarteten Zellbesetzungen größer sind als 5, dann kann ruhig der asymptotische p-Wert verwendet werden. Ansonsten verwende den exakten p-Wert.

Beim Beispiel 5.2.1. ist die kleinste erwartete Zellbesetzung gleich 6.1.

Manchmal findet man noch eine andere Faustregel als Zwischenschritt:

- Wenn die Stichprobenzahl kleiner ist als 60, dann verwende die **nach Yates korrigierte** Version des asymptotischen p-Werts. Diese Version heißt auch stetigkeits- oder kontinuieritätskorrigiert. Im SPSS Output für den Chi-Quadrat Test findet man diesen p-Wert in der Zeile "Kontinuitätskorrektur".

Für das Beispiel 5.2.1. ergibt sich ein kontinuieritätskorrigierte p-Wert von 0.253.

Gibt es auch noch andere exakte Tests?

Ja. Theoretisch könnte man jeden Test in einer "exakten" Variante durchführen. Praktisch wird es aber nur bei nichtparametrischen Tests gemacht. So z.B. auch beim uns bereits bekannten Wilcoxon-Mann-Whitney U-Test. Im Prinzip laufen alle exakten Tests in etwa so ähnlich ab, wie es oben am Beispiel 5.2.1. vorgeführt wurde. Der notwendige Rechenaufwand kann aber beträchtlich werden.

Heutzutage bieten immer mehr Programmpakete exakte Tests an (so auch SPSS, SAS, usw.). Für viele Probleme muss man aber immer noch auf Spezialprogramme ausweichen.

Bei sehr kleinen Stichproben

⇒ exakte anstatt asymptotisch ermittelte p-Werte angeben!!

E. Äquivalenzstudien

Die meisten klinischen Studien werden gemacht, um Unterschiede zwischen zwei Behandlungen festzustellen. Manchmal möchte man aber bloß zeigen, dass eine bestimmte neue Therapie (NT) gleich gut wie die bisher verwendete Standardtherapie (ST) wirkt. Solche Studien heißen Äquivalenzstudien. Der Grund dafür könnte sein, dass NT geringere Nebenwirkungen zeigt, weniger kostet, oder leichter zu verabreichen ist.

Im Englischen heißen *Äquivalenzstudien* entweder *equivalence trials* oder *similarity trials*. Der zweite Ausdruck weist darauf hin, dass exakte Gleichheit von zwei Behandlungsarten niemals gezeigt werden kann, selbst wenn diese Gleichheit tatsächlich bestünde. Was aber gezeigt werden kann, ist die "ausreichende Ähnlichkeit" der Behandlungen.

Zuerst müssen wir überlegen, was "ausreichende Ähnlichkeit" in so einem Fall bedeutet. Wenn NT besser als ST wirkt, dann stört uns das nicht. Wenn NT=ST gilt, ist alles in Ordnung. Und selbst wenn NT ein bisschen schlechter als ST wirkt, finden wir NT immer noch akzeptabel. Damit haben wir es mit einer einseitigen Hypothese zu tun.

Was bedeutet aber "ein bisschen schlechter"? Dazu bedarf es klinischer Überlegungen, welche Differenz als noch akzeptabel angesehen werden kann. Diese aus klinischen Gründen gerade noch akzeptable Differenz kürzen wir mit θ_0 ab.

Beispiel: Vergleich von Heilungsraten, $\theta_0 = 0.03$

Nullhypothese: ST ist um zumindest 3 Prozentpunkte besser als NT,
 $ST-NT \geq 0.03$, (keine Äquivalenz – hier wird wieder die Forschungshypothese negiert!)

Alternativhypothese: ST ist entweder schlechter, gleich gut oder um höchstens 3 Prozentpunkte besser als NT
 $ST-NT < 0.03$ (Äquivalenz)

Lösung: einseitiger Test auf einem Signifikanzniveau von 5 % oder Angabe eines einseitigen 95 % oder zweiseitigen 90 % Konfidenzintervalls

Beachte,

Fehler 1. Art: Irrtümlich Äquivalenz zu behaupten, obwohl Nullhypothese richtig

Fehler 2. Art: Irrtümlich keine Äquivalenz zu behaupten, obwohl Alternative richtig

Wenn wir zeigen wollen, dass der Effekt von zwei Behandlungen sich nicht allzu viel in beide Richtungen unterscheidet, haben wir es mit einer zweiseitigen Fragestellung zu tun. Zuerst müssen wir zwei gerade noch akzeptable Differenzen (je eine für Abweichungen nach oben wie nach unten) festlegen. Dazu definieren wir zwei einseitige Nullhypothesen:

Zum Beispiel: Vergleich von Heilungsraten, $\Theta_{01} = 0.03$, $\Theta_{02} = 0.07$

Nullhypothese 1: ST ist besser als NT um zumindest 3 Prozentpunkten
 $ST - NT \geq 0.03$ (keine Äquivalenz)

Nullhypothese 2: NT ist besser als ST um zumindest 7 Prozentpunkten
 $NT - ST \geq 0.07$ (keine Äquivalenz)

Alternativhypothese: ST ist um höchstens 3 Prozentpunkte besser als NT, gleich gut oder um höchstens 7 Prozentpunkte schlechter als NT
 $-0.07 < ST - NT < 0.03$ (Äquivalenz)

Erst wenn beide einseitigen Nullhypothesen auf dem 5 % Signifikanzniveau verworfen worden sind, können wir Äquivalenz annehmen.

Zweiseitige Hypothesen treten vor allem bei den sogenannten *bioequivalence trials* oder *bioavailability trials* auf. Dahinter stehen zumeist pharmakokinetische oder pharmakodynamische Fragestellungen.

Anmerkung: Man kann natürlich auch für Äquivalenzstudien Fallzahlberechnungen durchführen. Die "besseren" Stichprobenplanungsprogramme haben eigene Menüpunkte dafür.

F. Beschreibung von statistischen Methoden in medizinischen Publikationen

Wenn in einer medizinischen Publikation statistische Methoden verwendet werden, dann müssen diese auch adäquat beschrieben werden. Der "statistical methods"-Teil wird üblicherweise am Ende des "material and methods"-Kapitels angehängt (vor dem "results"-Kapitel).

Beachten Sie dabei bitte folgende Grundsätze:

- Einerseits: Die Beschreibung der statistischen Methoden sollte kurz und prägnant sein, denn in einer medizinischen Arbeit stehen medizinische Aspekte im Vordergrund.
- Andererseits: Die Beschreibung sollte detailliert genug sein. In anderen Worten: Wenn jemand die gleiche Studie wie wir durchgeführt hätte, dann sollte er anhand unseres statistischen Methodenteils in die Lage versetzt werden, die gleichen Auswertungen wie wir durchzuführen, um dann seine Ergebnisse mit den von uns publizierten vergleichen zu können.
- Empirische Resultate gehören nicht in den statistischen Methodenteil.

Wie soll ein statistischer Methodenteil aufgebaut werden?

- Beschreibung der verwendeten deskriptiven Maßzahlen, Daten-transformationen, etc.
- Beschreibung der verwendeten statistischen Tests, statistischen Modelle, Methoden zur Multiplizitätskorrektur, etc.
- Beschreibung der verwendeten Software
- Beschreibung des verwendeten Signifikanzniveaus und ob ein- oder zweiseitige Tests verwendet wurden

Beispiel (statistischer Methodenteil für das Rattendiätbeispiel 4.1.1):

Weight gain in both groups was described by mean and standard deviation. Differences between both groups were assessed with unpaired t-test. SPSS statistical software system (SPSS Inc., Chicago, IL) was used for calculations. The reported p-value is a result of a two-sided test. A p-value smaller or equal to 5 % is considered statistically significant.

G. Literaturverzeichnis

Die fett gedruckten Literaturzitate sind Einführungsbücher, die mit dem Seminarinhalt harmonisieren und ihn ergänzen.

- **D.G. Altman (1992): Practical statistics for medical research. Chapman and Hall.**
 - **M. Bland (1995): An Introduction to Medical Statistics. Second Edition. Oxford University Press.**
 - G. Gigerenzer (2002): Das Einmaleins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken. Berlin Verlag.
 - **I. Guggenmoos-Holzmann und K.-D. Wernecke (1995): Medizinische Statistik. Blackwell Wiss.-Verlag.**
 - A.R. Feinstein, D.M. Sosin and C.K. Wells (1985): The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. The New England Journal of Medicine, 312(25), 1604-1608.
 - **R.-D. Hilgers, P. Bauer und V. Scheiber (2003): Einführung in die Medizinische Statistik. Springer-Verlag.**
 - **M.H. Katz (1999): Multivariable Analysis. A Practical Guide for Clinicians. Cambridge University Press.**
 - D.E. Matthews and V.T. Farewell (1988): Using and Understanding Medical Statistics. 2nd, revised edition. Karger.
 - **H. Motulsky (1995): Intuitive Biostatistics. Oxford University Press.**
 - G. v. Randow (1992): Das Ziegenproblem. Denken in Wahrscheinlichkeiten. Rowohlt Verlag.
 - **M. Schumacher und G. Schulgen (2002): Methodik klinischer Studien. Methodische Grundlagen der Planung, Durchführung und Auswertung. Springer-Verlag.**
-