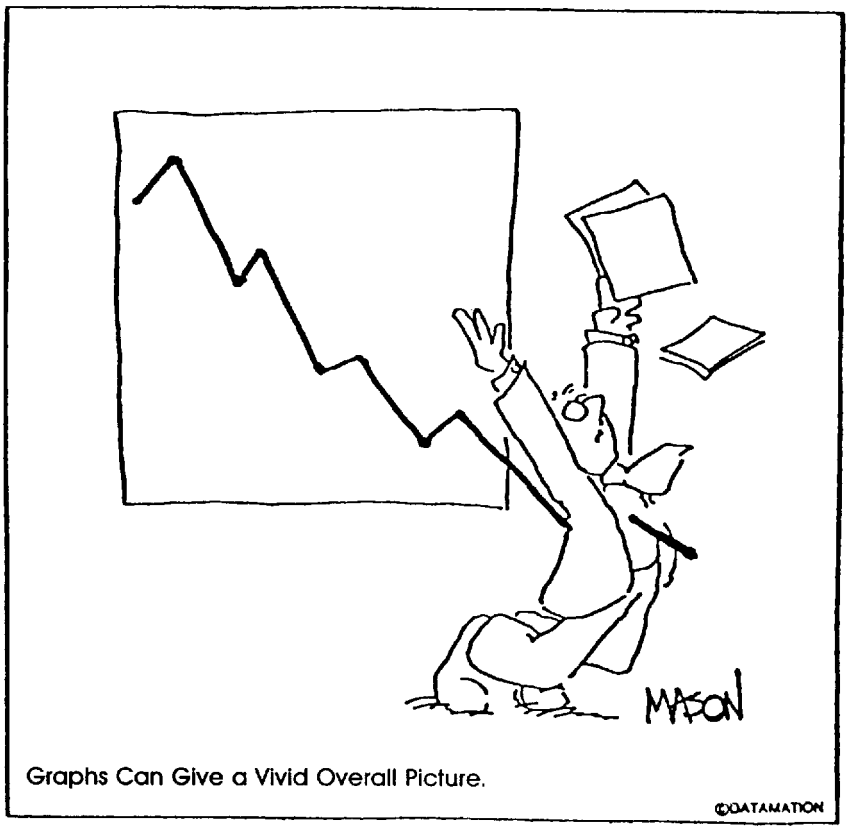
Grundbegriffe der Biostatistik

Theo Gasser & Burkhardt Seifert

Abteilung Biostatistik

Universität Zürich



3. Auflage 2006

Inhaltsverzeichnis

1 Einführung

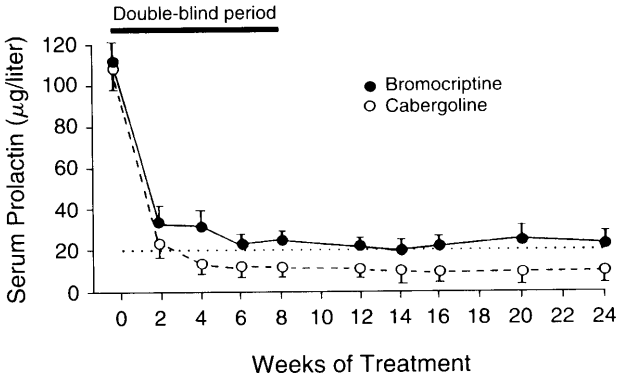
Der biomedizinische Bereich ist — neben den Sozialwissenschaften — derjenige universitäre Bereich, der am intensivsten moderne statistische Methodik benutzt. Auch wenn sich diese Methoden nur in wenigen Teilgebieten von jenen statistischen Methoden unterscheiden, die in anderen Wissenschaftszweigen Eingang gefunden haben, werden sie doch oft als Biostatistik oder auch als Medizinstatistik bezeichnet.
Das Spektrum der modernen Statistik reicht von einfachen quantitativen und graphischen Methoden bis hin zu komplexen Modellen, die nur mit Hilfe der höheren Mathematik behandelt werden können. Dabei stellt man in allen Anwendungen eine Verschiebung zugunsten des Einsatzes immer raffinierterer Methoden fest (z. B. logistische Regression oder Survivalanalyse). Dies hängt wesentlich damit zusammen, dass von der medizinischen Forschung auch immer anspruchsvollere Fragestellungen und Projekte angegangen werden. Der Erfolg der Statistik wäre aber auch nicht denkbar ohne die Entwicklung günstiger und leistungsfähiger Computer und die Entwicklung von statistischen Programmpaketen. Die letztere Entwicklung hinkt zwar immer etwas hinter dem technischen und methodischen Fortschritt hinterher, aber die Programmpakete sind heute teilweise sehr leistungsfähig und durchaus benutzerfreundlich, vorausgesetzt, man versteht etwas von statistischen Methoden.
In der biomedizinischen Forschung Tätige werden diese Methoden in verschiedenem Ausmass und verschiedener Vollständigkeit beherrschen (müssen). Ein Mediziner, der überwiegend in der Forschung tätig ist, sollte die Statistik besser beherrschen als ein überwiegend klinisch tätiger. Dabei spielen natürlich auch persönliche Neigungen eine Rolle. Jeder Mediziner sollte aber heute die Grundbegriffe der Statistik kennen und verstehen, um überhaupt medizinische Literatur lesen zu können (ein Blick in das New England Journal of Medicine oder Lancet genügt).
Dieses Skript ist eine Kurzfassung unseres bewährten Skripts „Biostatistik“. Die Betonung liegt auf den elementaren Kenntnissen der Biostatistik über Versuchsplanung, beschreibende Statistik, Testen von Hypothesen und Regression/Korrelation.
Nach unserer Bewichtung ist die deskriptive Statistik (Kapitel 2) zentral. Im Kapitel 3 werden wahrscheinlichkeitstheoretische Grundlagen für die folgenden Kapitel präsentiert; die Ergebnisse sollten jedoch weitgehend auch ohne Wahrscheinlichkeitstheorie verständlich sein. Wichtig für das weitere Vorgehen sind die Abschnitte 3.4 (wichtige Verteilungen) und 3.9 (Versuchsplanung). Die Prüfung von Hypothesen (Kapitel 4) ist ein schwieriges Gebiet, das aber wegen seiner Bedeutung zentral ist. Die Regressionsanalyse (Kapitel 5) ist ebenfalls zentral, aber im allgemeinen leichter verständlich. Kapitel 6 gibt einen Einblick in weiterführende statistische Verfahren, die in der medizinischen Literatur weit verbreitet sind.

1. Auflage 1997
2. überarbeitete Auflage 2003
3. überarbeitete Auflage 2006

1.1 Was bietet die Statistik ?

Statistische Methoden ermöglichen
- die Hervorhebung wesentlicher Zusammenhänge durch Datenreduktion und graphische Darstellungen.

Es folgt ein Beispiel aus dem New England Journal of Medicine (einer führenden klinischen Zeitschrift). Durch eine einfache Graphik wird die Wirkung von Behandlungen im Zeitverlauf auf einen Blick deutlich.



Betrachten wir einen Artikel aus dem gleichen Heft (siehe Kopie auf der nächsten Seite). Ohne statistische Kenntnisse sind viele Artikel in medizinischen Zeitschriften nur schwer verständlich. Emerson et al. (1986) stellten fest, dass jemand, der nur die deskriptive Statistik beherrscht, knapp 58% der Artikel des New England Journal of Medicine verstehen kann. Das Verständnis des t–Tests erhöht diesen Anteil auf 67%; das zusätzliche Verständnis von Kontingenztafeln erhöht ihn auf 73%. Seit 1986 ist der Anteil höherer statistischer Methoden deutlich gestiegen.

Ein Ziel dieser Biostatistik–Vorlesung ist es denn auch, das Studium wissenschaftlich–medizinischer Literatur zu erleichtern.

Statistische Methoden ermöglichen es,
- aus einer Stichprobe gültige Schlussfolgerungen zu ziehen,
- die Unsicherheit der Entscheidung zu quantifizieren.

Dazu stehen gute statistische Programmpakete zur Verfügung, die zum Teil auch benutzerfreundlich sind. Mit diesen und dem Stoff der Vorlesung sollte es möglich sein, einfachere Auswertungen selber durchzuführen. Die meisten Beispiele in diesem Skript wurden mit dem Paket SPSS analysiert.

896 THE NEW ENGLAND JOURNAL OF MEDICINE Oct. 6, 1994

INTENSIVE POSTREMISSION CHEMOTHERAPY IN ADULTS WITH ACUTE MYELOID LEUKEMIA

ROBERT J. MAYER, M.D., ROGER B. DAVIS, SC.D., CHARLES A. SCHIFFER, M.D., DEBORAH T. BERG, R.N., BAYARD L. POWELL, M.D., PHILIP SCHULMAN, M.D., GEORGE A. OMURA, M.D., JOSEPH O. MOORE, M.D., O. ROSS MCINTYRE, M.D., AND EMIL FREI III, M.D., FOR THE CANCER AND LEUKEMIA GROUP B*

Abstract *Background.* About 65 percent of previously untreated adults with primary acute myeloid leukemia (AML) enter complete remission when treated with cytarabine and an anthracycline. However, such responses are rarely durable when conventional postremission therapy is administered. Uncontrolled trials have suggested that intensive postremission therapy may prolong these complete remissions.

Methods. We treated 1088 adults with newly diagnosed AML with three days of daunorubicin and seven days of cytarabine and randomly assigned patients who had a complete remission to receive four courses of cytarabine at one of three doses: 100 mg per square meter of body-surface area per day for five days by continuous infusion, 400 mg per square meter per day for five days by continuous infusion, or 3 g per square meter in a 3-hour infusion every 12 hours (twice daily) on days 1, 3, and 5. All patients then received four courses of monthly maintenance treatment.

Results. Of the 693 patients who had a complete remission, 596 were randomly assigned to receive postremission cytarabine. After a median follow-up of 52 months, the disease-free survival rates in the three treatment groups were significantly different (P = 0.003). Relative to the 100-mg group, the hazard ratios were 0.67 for the 3-g group (95 percent confidence interval, 0.53 to 0.86) and 0.75 for the 400-mg group (95 percent confidence interval, 0.60 to 0.94). The probability of remaining in continuous complete remission after four years for patients 60 years of age or younger was 24 percent in the 100-mg group, 29 percent in the 400-mg group, and 44 percent in the 3-g group (P = 0.002). In contrast, for patients older than 60, the probability of remaining disease-free after four years was 16 percent or less in each of the three postremission cytarabine groups.

Conclusions. These data support the concept of a dose–response effect for cytarabine in patients with AML who are 60 years of age or younger. The results with the high-dose schedule in this age group are comparable to those reported in similar patients who have undergone allogeneic bone marrow transplantation during a first remission. (N Engl J Med 1994;331:896-903.)

1.2 Grundgesamtheit und Stichprobe

Zwei zentrale Begriffe der Statistik sind Population (oder: Grundgesamtheit) und Stichprobe. Individuen verhalten sich bei gleicher Behandlung unterschiedlich. Deshalb reicht es in der Medizin nicht, Einzelfälle zu dokumentieren. Es ist auf der anderen Seite unmöglich und auch nicht unbedingt wünschenswert, das Verhalten **aller** Patienten auf eine Behandlung zu evaluieren. Dies ist einerseits eine Kostenfrage, andererseits geht es darum, Behandlungen mit Nebenwirkungen so sparsam wie möglich einzusetzen. Als Ausweg können wir eine **Stichprobe** aus der **Grundgesamtheit** (Population, statistische Grundgesamtheit) ziehen.

> Die **Grundgesamtheit** ist die Gesamtheit aller Individuen, für welche Schlussfolgerungen gezogen werden sollen.

Beispiele:

1. Alle Hodgkin–Patienten der Welt, die die Krankheit überstanden haben (siehe Kapitel 2).
2. Alle männlichen Einwohner der Schweiz über 65 Jahre, die weder an einer neurologischen noch an einer psychiatrischen Krankheit leiden.

Eine **Stichprobe** aus einer statistischen Grundgesamtheit ist die Menge der Individuen, die tatsächlich beobachtet wurden.

Beispiele:

1. Stichprobe von $n = 20$ Hodgkin–Patienten aus einer Klinik.
2. Repräsentative Stichprobe von $n = 1000$ Männern aus städtischen und ländlichen Gebieten.

1.3 Empfohlene Literatur

Es gibt hunderte von Büchern zur Einführung in die Statistik. Ein grosser Teil davon ist brauchbar. Wir wollen nur einige der anwendungsorientierten Bücher erwähnen.

Altman, D. G. (1991). *Practical statistics for medical research.* Chapman and Hall.
Es bietet eine korrekte Statistik und sehr gute Beispiele. Aus diesem Buch haben wir einige der vorgestellten Beispiele. 600 S.

Armitage, P., Berry, G. & Matthews, J. N. S (1991). *Statistical methods in medical research* 4th ed., Blackwell.
Recht umfangreiches Lehrbuch. 817 S.

Bland, M. (1995). *An introduction to medical statistics.* Oxford Medical Publications.
Eine sehr gute Einführung mit vielen Beispielen und Aufgaben. 396 S.

Breiman, L. (1973). *Statistics. With a view towards applications.* Houghton Mifflin Comp.
Es bietet eine gute Statistik mit mathematischen Grundlagen und mit vielen Beispielen. 400 S.

Johnson, R. A. & Bhattacharyya, G. K. (1992). *Statistics. Principles and methods.* 2nd ed., Wiley.
Eine leichte Lektüre zum Feierabend. 700 S.

Matthews, D. E. & Farewell, V. T. (1988). *Using and understanding medical statistics.* 2nd ed., Karger.
Im Gegensatz zu vielen anderen Einführungen bietet dieses Buch logistische Regression und Survivalanalyse. 200 S.

Nachschlagewerk

Sachs, L. (2004). *Angewandte Statistik: Anwendung statistischer Methoden.* 10. Auflage, Springer.
Im deutschsprachigen Raum das klassische Kochbuch (Nachschlagewerk) für alle Lebenslagen seit 1968. Achtung: Kein Lehrbuch! 890 S.

Darstellung statistischer Analysen

Lang, T. A. & Secic, M. (1997). *How to report statistics in medicine.* Philadelphia: ACP.

2 Univariate deskriptive Statistik

Die Beschreibung von Daten ist die Grundlage jeder statistischen Analyse und ein wesentlicher Bestandteil jeder Publikation. Das Ziel besteht darin, die Daten einer Stichprobe kurz und prägnant zu charakterisieren. Dies erfolgt einerseits über statistische Kennwerte (z. B. Lage– und Streumasse) und andererseits durch graphische Verfahren. Dank PC und Laserdrucker und dank der Entwicklung guter, einfach handhabbarer Programme haben graphische Methoden an Bedeutung gewonnen.

Die deskriptive Statistik unterscheidet sich von der schliessenden Statistik dadurch, dass die Daten **ohne Signifikanz** präsentiert werden, und sie so ohne Wahrscheinlichkeitsannahmen auskommt. Trotzdem verweisen wir in diesem Kapitel gelegentlich auf Begriffe der Wahrscheinlichkeitsrechnung (Erwartungswert, Dichte, Normalverteilung). Diese Begriffe werden in Kapitel 3 erklärt.

Daten werden heute meistens mittels eines Tabellenkalkulations–Programmes oder einer Datenbank in den Computer eingegeben. Für erstere Option ist das Programm Excel eine beliebte und auch gute Wahl. Gut ist im Prinzip auch Filemaker, doch gibt es Probleme beim Export der Daten in Statistik–Programme. Ein Vorteil von Filemaker ist, dass Datenchecks leicht durchgeführt werden können. Für befristete und nicht zu komplexe Projekte ist die Benutzung einer Datenbank nicht angezeigt; die verbreitetsten Datenbanken erlauben aber den Export in die gängigen Statistik–Pakete. In der Regel werden die Daten als Tabelle eingegeben, wobei jede Zeile einen Patienten darstellt, die Kolonnen die jeweiligen Variablen. Nachstehend finden Sie die Daten einer Studie als Excel–Tabelle, die später immer wieder als Beispiel dient. Hier wurden immunologische Parameter (T_4– und T_8–Zellen) bei 20 Hodgkin und 20 non–Hodgkin–Patienten verglichen (siehe Abschnitt 2.4).

Patient	T4-cells	T8-cells	ln(T4-cells)	ln(T8-cells)	Disease	group
1	396	836	5.981	6.729	Hodgkin	1
2	568	978	6.342	6.886	Hodgkin	1
3	1212	1678	7.1	7.425	Hodgkin	1
...						
19	1283	336	7.157	5.817	Hodgkin	1
20	2415	936	7.789	6.842	Hodgkin	1
21	375	340	5.927	5.829	Non-Hodgkin	2
22	375	330	5.927	5.799	Non-Hodgkin	2
23	752	627	6.623	6.441	Non-Hodgkin	2
...						
39	377	108	5.932	4.682	Non-Hodgkin	2
40	503	163	6.221	5.094	Non-Hodgkin	2

2.1 Haupttypen von Daten

Am Anfang ist es wichtig, sich über die verschiedenen Typen von Daten klar zu werden. Verschiedene Datentypen bedingen nämlich verschiedene Präsentationen und Analysemethoden.

1. Qualitative oder **diskrete** Messdaten

   Diskrete Daten sind dadurch gekennzeichnet, dass sie nur bestimmte Werte annehmen können, z. B.: {rot, grün, blau}, {O, A, B, AB}, {männlich, weiblich} oder $\{0, 1, 2, \dots\}$, wobei diese Zahlen z. B. den Schweregrad einer Krankheit bedeuten können.

   Wir unterscheiden:

   - **nominal** oder kategoriell (Zuordnung zu Kategorien):
     Beziehung: gleich $\iff$ ungleich
     $\longrightarrow$ nur Anzahl und % sinvoll
     Beispiele: Geschlecht, Blutgruppe, Farbe
   - **ordinal** (geordnet kategoriell): Beziehung: grösser $\iff$ kleiner (Rangordnung)
     Beispiele: Schweregrad einer Krankheit, Items in Fragebogen

2. Quantitative oder **stetige** (numerische) Messdaten

   Stetige Daten können idealerweise alle Werte ohne Abstufungen annehmen. Oft ist es sinnvoll, auch ganzzahlige (also eigentlich diskrete) Werte als stetig zu behandeln (z. B. Grösse in cm oder Zähldaten).

   Es werden noch intervallskalierte und absolutskalierte Variablen unterschieden, doch hat dies für die Statistik selten eine Bedeutung. Bei absolutskalierten Grössen ist ein absoluter Nullpunkt vorgegeben, und dadurch kann man auch sinnvolle Quotienten bilden. Beispiel: Temperatur in Kelvin anstatt Grad Celsius.

In einem psychologischen Sinne ist Farbe kategoriell skaliert, durch den wissenschaftlichen Fortschritt kann man sie aber auch als stetig (Frequenzen von elektromagnetischen Wellen) auffassen.

2.2 Darstellung von diskreten Daten

Bei der Analyse von diskreten Daten spielt die **Wahrscheinlichkeit** eines einzelnen Ereignisses eine zentrale Rolle. Die relative Häufigkeit schätzt diese Wahrscheinlichkeit. Ein Schätzer (oder eine Schätzung, englisch estimator) ist eine statistische Approximation auf der Basis der Stichprobe an eine unbekannte wahre Grösse in der Grundgesamtheit.

$$\text{relative Häufigkeit} = \frac{\text{Anzahl Beobachtungen des Ereignisses}}{\text{totale Anzahl Beobachtungen}}$$

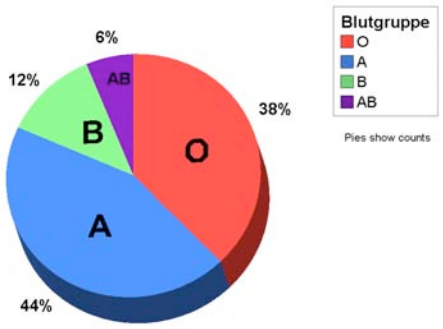
- multipliziert mit 100 erhält man Prozentsätze

Beispiel: Anteil von Blutgruppen in einer gesunden Population. So könnte die Präsentation der Daten aussehen:

Tabelle

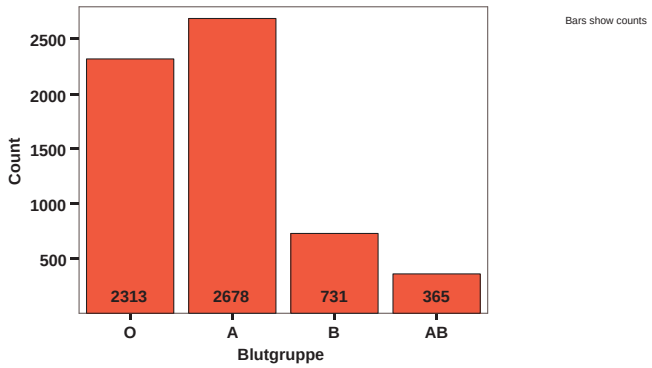
Blutgruppe	Anzahl	relative Häufigkeit
O	2313	38%
A	2679	44%
B	731	12%
AB	365	6%
total	5988	100%

Kuchendiagramm (pie chart)



Beachten Sie, dass in der 3–dimensionalen Darstellung die Proportionen durch den perspektivischen Effekt und den vorderen Rand visuell verfälscht werden.

Balkendiagramm (bar chart)



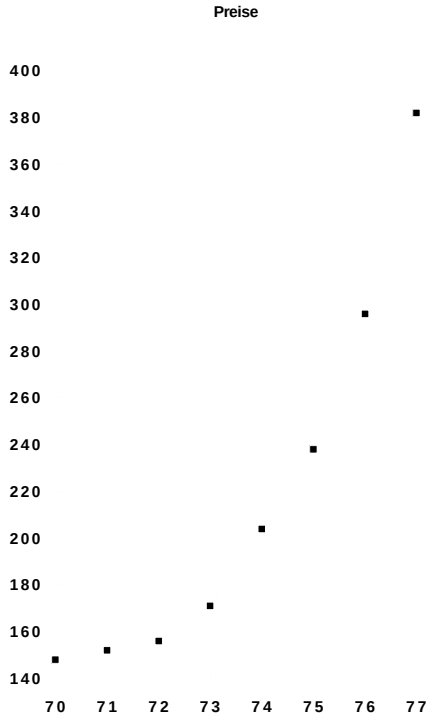
Beachten Sie, dass die Balken immer vom Nullpunkt ausgehen sollten.

2.3 Darstellung von Verläufen

- Ordinale und stetige Daten

Die Darstellung von Verläufen ist ein heikles Thema. Hier eine extrem missbräuchliche Darstellung:

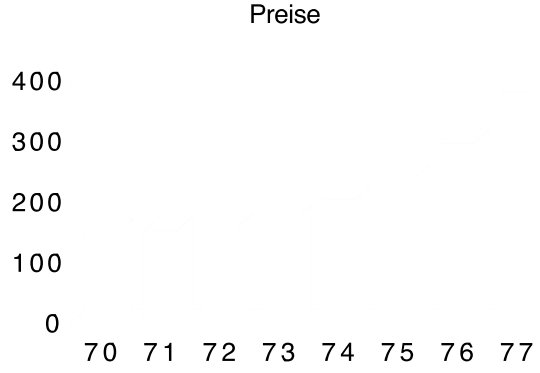
Liniendiagramm



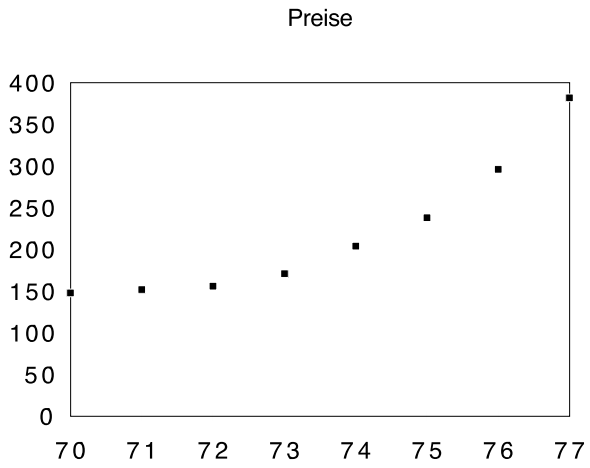
Die Gründe dafür sind:
- Das Verhältnis der x– zu der y–Skala ist extrem gewählt.
- Der Nullpunkt ist nicht in der Graphik enthalten!

Eine korrekte Darstellung sieht folgendermassen aus:

Balkendiagramm



Scattergramm



Es ist übrigens nicht einfach, in gewissen Graphikprogrammen den Nullpunkt in ein Scattergramm (Streudiagramm) einzubeziehen.

- Benutzen Sie Ihren gesunden Menschenverstand, um zu entscheiden, ob der Nullpunkt für die Darstellung bedeutsam ist.

2.4 Darstellung von stetigen Daten – das Histogramm

Ziel:
Verteilung der Daten graphisch zu charakterisieren, im Sinne einer „Datendichte“ durch das Histogramm.

Vorgehen:

- Bereich der Daten in gleiche, nicht überlappende Intervalle (Zellen, Klassen) zerlegen
- Anzahl Beobachtungen pro Intervall bestimmen

$$\text{relative Häufigkeit im Intervall} = \frac{\text{Anzahl Beobachtungen im Intervall}}{\text{totale Anzahl Beobachtungen}}$$

- relative (oder absolute) Häufigkeiten über Intervalle in Balkendiagramm darstellen

Beispiel: Immunologische Variablen als Anzahl von T_4– und T_8–Zellen bei verschiedenen Krebspatienten. Dies ist ein Beispiel aus dem Am. J. Med. Sci., das wir noch mehrmals benutzen werden:

Immunologic Status of Patients in Remission from Hodgkin's Disease and Disseminated Malignancies

BY CHARLES M. SHAPIRO, MD, ENRIQUE BECKMANN, MD, PhD, NEAL CHRISTIANSEN, MD, JACOB D. BITRAN, MD, MARK KOZLOFF, MD, ARTHUR A. BILLINGS, MD, MARGARET C. TELFER, MD

ABSTRACT: The immunologic phenotypes of peripheral blood lymphocytes of 20 patients cured of Hodgkin's disease (Group I) and of 20 patients cured of diverse, disseminated malignancies (Group II) were compared. Eleven patients in Group I had an increase in circulating B lymphocytes, and 6 patients in Group II had a decrease in this lymphocyte subset ($0.0005 < p < 0.0025$). The T_8 lymphocyte subset was increased in four patients in Group I and decreased in five patients in Group II ($0.005 < p < 0.0125$). The T_4/T_8 lymphocyte ratio was less than 1.0 in eight patients in Group I and in no patients in Group II ($p < 0.001$). These statistically significant differences between the two groups were unrelated to type or duration of therapy, and thus suggest basic biologic differences between the immune status of cured Hodgkin's disease patients and cured patients with other malignancies. KEY INDEXING TERMS: Immunologic Phenotypes; T_4/T_8 Lymphocyte Ratio; Hodgkin's Disease; Cancer. [Am J Med Sci 1986; 293(6):366–370.]

A large body of literature documents the presence of multiple immunologic abnormalities in patients with Hodgkin's disease (HD).[1–6] These abnormalities occur early in the clinical course and are more pronounced in advanced stages of the disease.[7–9] Less defined are the status of the immune system in patients in remission from HD and other malignancies and the effect of different prior therapies on immune function.[10,11]

In 1983, Lauria et al described several abnormalities in T lymphocyte subsets in disease-free HD patients, 5 years after therapy.[12] They noted a decrease in the proportion of T_4 lymphocytes. The T_8 lymphocyte subset was increased, both proportionally and in absolute numbers, resulting in an abnormal T_4/T_8 ratio. This study was undertaken to confirm these findings and to determine whether they are unique to patients with HD.

Materials and Methods

Forty patients with disseminated malignancies were entered into the study. They were disease-free for periods of 3–20 years. There were 20 HD patients who had been staged according to the Ann Arbor classification (Table 1).[13] All but six patients had splenectomies during staging laparotomies. The modes of treatment consisted of standard MOPP chemotherapy (nitrogen mustard, vincristine, procarbazine, prednisone) (six patients), extended mantle or total nodal radiation (ten patients), or a combination of radiation and chemotherapy (four patients). The remaining 20 patients had diverse, disseminated malignancies (Table 2). There were four patients with non-Hodgkin's lymphoma, three patients with inoperable lung carcinoma, three patients with Stage II–III breast carcinoma, three patients with Stage III ovarian carcinoma, three patients with Stage III testicular carcinoma, two patients with Stage C colon carcinoma, and one patient each with multiple myeloma and cholangiocarcinoma. They were treated with a variety of multidrug combinations, except for one patient who received only radiation therapy.

Mononuclear cells were purified from 30 ml of heparinized or EDTA anticoagulated peripheral blood, by centrifugation over Ficoll–Hypaque cushions (Pharmacia Fine Chemicals, Piscataway, NJ).[14]

From the Department of Medicine, Division of Hematology/Oncology, and Department of Pathology, Division of Microbiology and Diagnostic Immunology, Michael Reese Hospital and Medical Center, and the University of Chicago Pritzker School of Medicine, Chicago, Illinois.

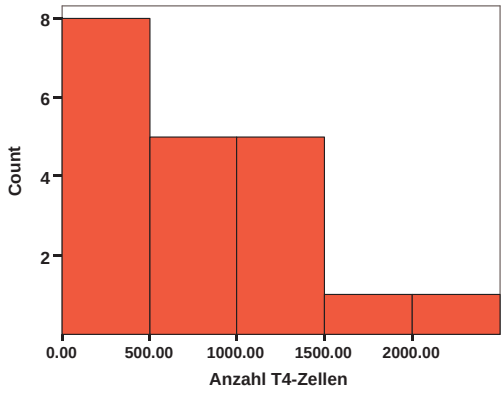
Supported in part by the Arthur A. Billings Research Fund for Lymphomas and Allied Diseases, Michael Reese Hospital and Medical Center, Chicago, Illinois.

Reprint requests: Charles M. Shapiro, MD, Division of Hematology/Oncology, L-231 Blood Center, Michael Reese Hospital and Medical Center, Lake Shore Drive at 31st Street, Chicago, ILL 60616.

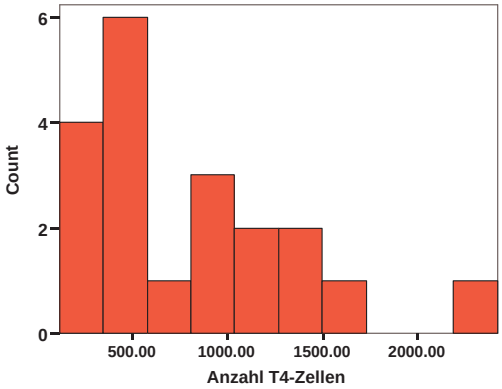
Für das Merkmal „Anzahl T_4–Zellen bei Hodgkin–Patienten“ sieht die Berechnung eines Histogramms etwa so aus, wenn man Intervalle zu 500 bildet:

Intervall	Anzahl T_4–Zellen	Anzahl Beob. im Intervall	relative Häufigkeit
1–500	171		
	257		
	288		
	295		
	396		
	397		
	431		
	435	8	40%
501–1000	554		
	568		
	795		
	902		
	958	5	25%
1001–1500	1004		
	1104		
	1212		
	1283		
	1378	5	25%
1501–2000	1621	1	5%
2001–2500	2415	1	5%
total		20	100%

Aus der Tabelle ergibt sich das folgende Histogramm („Count“ bedeutet Anzahl Patienten pro Intervall):

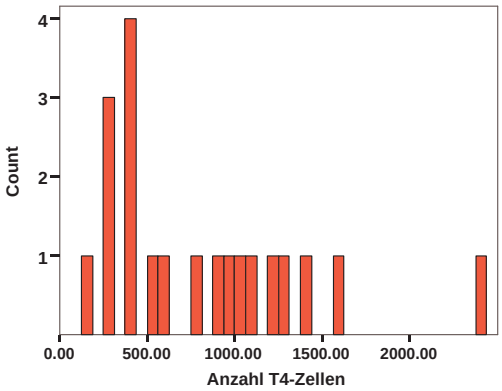


Das Standard–Histogramm von SPSS sieht so aus:



- Vergleichen Sie diese beiden Histogramme und Sie bemerken, dass man die Graphiken unterschiedlich interpretieren kann.
  - Steigt die relative Häufigkeit am Anfang an?
  - Gibt es Lücken?
- Eine wesentliche Erkenntnis ist, dass die Anzahl T_4–Zellen (rechts–)schief verteilt ist, d. h. dass rechts vom Häufigkeitszentrum mehr Werte als links davon auftreten. Eine logarithmische Transformation führt zu einer eher symmetrischen Verteilung, ähnlich einer Normalverteilung (Gaussche Glockenkurve).
- Die Aussage eines Histogramms hängt wesentlich von der **Klassenbreite** und dem **Klassenzentrum** ab.

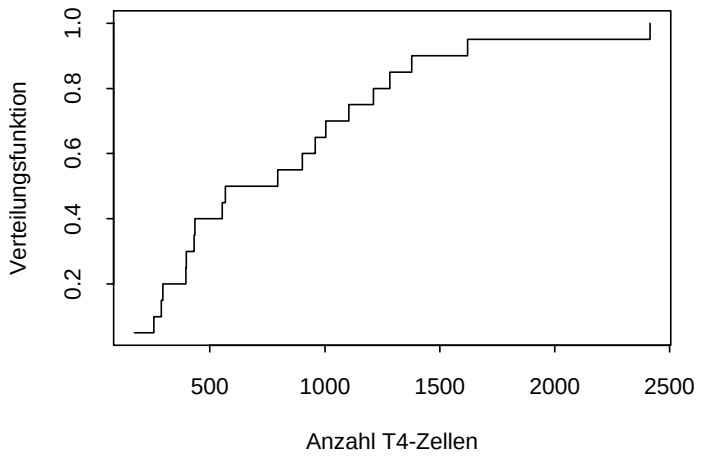
Wenn man eine deutlich zu kleine Intervallänge wählt, erhält man z. B. ein sehr variables Histogramm:



Das Histogramm ist ein Schätzer der Wahrscheinlichkeitsdichte.

- Histogramme sind einfach und verbreitet.
- Es gibt bessere Dichteschätzer, die aber nur in einigen Statistikpaketen verfügbar sind.

Eine kumulative Darstellung der Daten ist durch die empirische Verteilungsfunktion gegeben. Dabei steigt die Funktion treppenartig um $1/n$, wenn ein Datenpunkt dazukommt (n = Stichprobengrösse). Die Verteilungsfunktion beginnt dadurch bei $y = 0$ und steigt dann an bis $y = 1$.



Im allgemeinen kann man aus der kumulativen Darstellung weniger ersehen als aus dem Histogramm, was die grössere Verbreitung des Histogramms erklärt.

2.5 Lage– und Streumasse

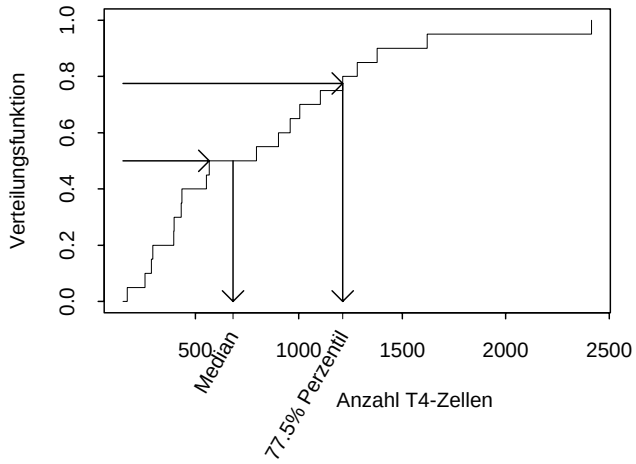
2.5.1 Perzentile oder Quantile

Perzentile sind Hilfsmittel zur Beschreibung der Verteilung von ordinalen und stetigen Daten ohne irgendwelche Verteilungsannahmen (wie z. B. Normalverteilung).

Bedeutung:
1. zur Charakterisierung einer Stichprobe
2. zur Konstruktion von Normwerten (biochemisch, anthropometrisch, psychometrisch)

Erklärung $\alpha\%$ – **Perzentil:** $\alpha\%$ der Daten sind kleiner als das $\alpha\%$ – Perzentil.

In der folgenden Graphik finden Sie eine Erklärung für Perzentile. Die Treppenlinie stellt die empirische (Stichproben–) Verteilungsfunktion der Anzahl T_4–Zellen für $n = 20$ Hodgkin–Patienten dar. Wenn man von $y = \alpha$ eine Horizontale zieht und beim Schnittpunkt mit der Verteilungsfunktion den zugehörigen x–Wert abliest, erhält man das $\alpha\%$ – Perzentil. Für das 77.5% Perzentil trifft man eine Stufe, und damit ist dieses Perzentil eindeutig definiert. Beim 50% Perzentil oder Median trifft man auf ein horizontales Stück und trifft damit 2 Datenpunkte. Als Median wählt man dann den Mittelwert der 2 x–Werte.



Genaue **Definition:** Mindestens $\alpha\%$ der Daten sind gleich oder kleiner und mindestens $(100 - \alpha)\%$ sind gleich oder grösser als das $\alpha\%$ – Perzentil.

Quantile, in der Statistik gebräuchlich, sind das gleiche wie Perzentile bis auf den Faktor 100 (also z. B. α – Quantil).

Wichtige Perzentile:

- **Median** = 50% Perzentil
  Der Median ist ein Lagemass, der das Zentrum der Daten charakterisiert.
- **Quartile** = 25% und 75% Perzentile
  Der Abstand zwischen den Quartilen (Interquartilabstand oder englisch „interquartile range“) ist ein Streumass. Dieser Bereich enthält die zentralen 50% der Daten.

Achtung! Die Werte sind nicht eindeutig!
Verschiedene Programme können also etwas verschiedene Werte liefern. Dies hat damit zu tun, dass man auf verschiedene Arten Anteile von Daten verrechnen kann, wenn die Prozentrechnung nicht eine eindeutige Beobachtung ergibt.

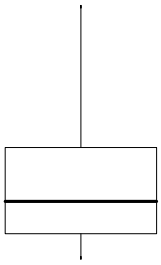
Der Median von 9 Beobachtungen berechnet sich z. B. so, dass man diese ordnet und die 5. Beobachtung als Median nimmt. Bei 10 Werten liegt der Median zwischen dem 5. und 6. geordneten Wert, und man nimmt dann üblicherweise das Mittel beider Werte.

So sieht das Vorgehen am Beispiel der Anzahl T_4 – Zellen aus:

Nr.	Anzahl T_4–Zellen	Perzentil
1	171	
2	257	⟵ 10% Perzentil = 272.5
3	288	
4	295	
5	396	⟵ unteres Quartil = 396.5
6	397	
7	431	
8	435	
9	554	
10	568	⟵ Median = 681.5
11	795	
12	902	
13	958	
14	1004	
15	1104	⟵ oberes Quartil = 1158
16	1212	
17	1283	
18	1378	⟵ 90% Perzentil = 1499.5
19	1621	
20	2415	

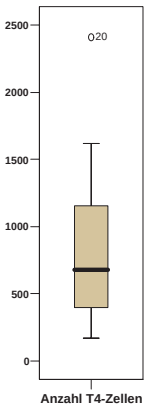
Boxplots

Boxplots sind ein gutes Hilfsmittel, um optisch die Verteilung der Daten zu erfassen und Ausreisser zu erkennen. Es handelt sich um eine Methode, die auf Perzentilen basiert. Das ist die einfachste Form:



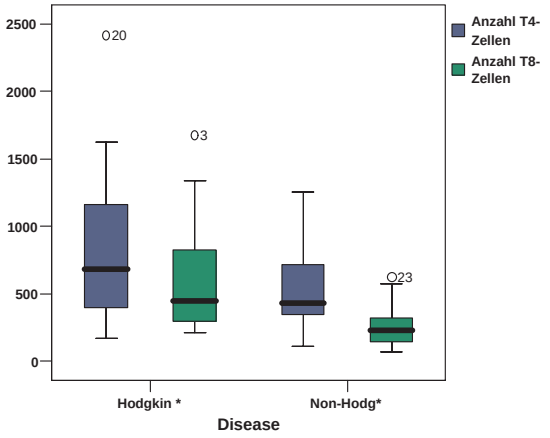
Die „Box“ gibt den Bereich vom 25. zum 75. Perzentil an, der horizontale Strich in der Box den Median. Die Stäbe (whiskers), die aus der Box hinausführen, sind nicht einheitlich definiert. In vielen Statistikprogrammen gehen sie vom Minimum zum Maximum oder sie charakterisieren die Grenzen für Beobachtungen, die keine Ausreisser sind. In SPSS z. B. ist ein Ausreisser eine Beobachtung, die weiter als 1.5 Interquartilabstände (Box – Längen) von der Box entfernt ist.

So sieht der Boxplot in SPSS aus:



Die Asymmetrie des Boxplots deutet auf eine rechtsschiefe Verteilung hin, die Quartile und die Whiskers haben gegen oben einen grösseren Abstand als gegen unten. Sehr klar tritt der Extremwert bei 2415 T_4–Zellen/mm^3 hervor.

Besonders hilfreich sind Boxplots, wenn mehrere Merkmale bzw. Gruppen verglichen werden:



Was sieht man auf einen Blick?

- Die immunologischen Variablen sind schief verteilt, denn der Median teilt den Interquartilsabstand nicht gleichmässig und Extremwerte sind ebenfalls asymmetrisch.
- Offensichtlich liegen die Anzahlen von T_4– und T_8–Zellen bei non–Hodgkin–Patienten tiefer und streuen weniger als bei Hodgkin–Patienten. Dies entspricht der wissenschaftlichen These der Hodgkin–Studie.
- Die Anzahlen von T_8–Zellen liegen tiefer als die von T_4–Zellen und streuen weniger, und zwar bei beiden Gruppen.

2.5.2 Charakterisierung des Zentrums der Daten

- Was ist ein typischer, mittlerer Wert ?

Möglichkeiten:

1. graphisch: Histogramme und Boxplots vermitteln einen Eindruck, wie die Daten verteilt sind. Um eine objektive, quantitative Zusammenfassung der Daten zu erhalten, benötigen wir aber klar definierte Masszahlen für das Zentrum der Daten.

2. quantitativ durch den **Mittelwert**: Der Mittelwert (mean, average) beschreibt das Verhalten der Daten „im Mittel“ ($\sum$ = Summenzeichen).

$$\overline{x} = (x_1 + x_2 + \ldots + x_n) \,/\, n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Bei **Normalverteilung** ist der Mittelwert optimaler Schätzer des **Erwartungswertes**.

3. quantitativ durch den **Median**: Der Median (median) beschreibt die „Mitte“ der Daten. Er ist der 50% – Punkt, d. h. die Hälfte der Stichprobe liegt über dem Median, und die andere Hälfte liegt darunter (siehe Perzentile).

Ein anschauliches Beispiel für die Unterschiede zwischen Mittelwert und Median gibt die Einkommensverteilung. Wenn z. B. in einer Gemeinde ein extrem gutverdienender Einwohner wohnt, so ist der Mittelwert relativ hoch, auch wenn die übrigen Einwohner weniger verdienen. Der Median charakterisiert dann besser das typische Einkommen der Bürger, während das Mittel für die Steuerkraft ein aussagekräftiger Index ist.
Bei Verteilungen, die rechtschief sind, d. h. wo grössere Abweichungen nach oben als nach unten auftreten, ist der Mittelwert grösser als der Median (Beispiel: Einkommen und viele biochemische Variablen). Bei symmetrischer Verteilung (im besonderen Normalverteilung) sind Median und Mittelwert identisch (Beispiel: Körpergrösse).

Vorteile der verschiedenen Methoden: Die Wahl zwischen Mittelwert und Median ist

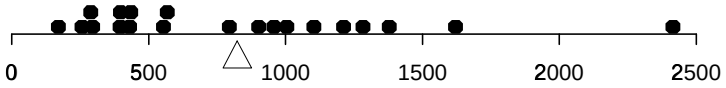
- abhängig davon, ob ein typischer oder mittlerer Wert gesucht wird,
- abhängig von der Verteilung (normal, schief, gibt es Ausreisser?),
- abhängig davon, ob statistische Präzision oder Robustheit im Vordergrund stehen.

Der folgende Vergleich soll die Unterschiede zwischen Mittelwert und Median illustrieren.

Der **Mittelwert** ist derjenige Wert, der die Daten auf einer „Waage“ ausbalanciert.

Wir gehen von einer Waage ohne Gewicht und gleich schweren Beobachtungen aus.

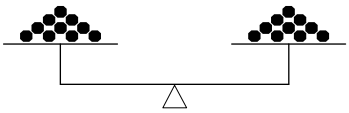
Mittelwert: $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$



- Entfernte Beobachtungen haben eine starke „Hebelkraft“.

Der Mittelwert ist **empfindlich** gegen Ausreisser.

Median: Beim Median spielt der Abstand der Beobachtungen vom Zentrum keine Rolle.



- Der Median ist **robust** gegen Ausreisser.

2.5.3 Streuung oder Variabilität einer Stichprobe

- Wie stark variieren die Daten um die mittlere Lage?

Jetzt sind also nicht die x_i, sondern die $(x_i - \overline{x})$ relevant.

Möglichkeiten:

1. graphisch: Histogramme und Boxplots vermitteln einen visuellen Eindruck der Variabilität der Daten.

2. quantitativ durch die **Varianz** s^2: Die Varianz ist ein Mass für die Variabilität der Abweichungen $(x_1 - \overline{x}), \ldots, (x_n - \overline{x})$ vom Mittelwert.
   Einfach Aufsummieren funktioniert aber nicht, die Summe ist 0. Also werden die Abweichungen quadriert:

$$\begin{aligned} s^2 &= \{(x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2\} / (n-1) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \end{aligned}$$

Der Nenner $(n-1)$ (anstatt n) überrascht zunächst; er ist mathematisch erklärbar.

Achtung: Die Varianz s^2 ist in quadrierten Einheiten (z. B. cm^2), was störend sein kann. Deshalb ist es gebräuchlich, die Wurzel aus der Varianz zu benützen:

Standardabweichung: $s = \sqrt{\text{Varianz}}$ (z. B. in cm)
(standard deviation, SD)

- Die Standardabweichung und die Varianz sind empfindlich gegen Ausreisser.

Bei der **Normalverteilung** liegen 68% der Daten im Bereich Mittelwert ± 1 SD und 95% im Bereich Mittelwert ± 2 SD.

3. quantitativ durch die **Spannweite**
= Maximum − Minimum
- Die Spannweite gibt den Bereich (range) aller Daten an. Sie ist stark durch Extremwerte beeinflusst und hängt zudem von n ab.

4. quantitativ durch den **Interquartilabstand** (interquartile range)
= 75% Perzentil − 25% Perzentil
= oberes Quartil − unteres Quartil
= 3. Quartil − 1. Quartil
$\Longrightarrow$ umfasst zentrale 50% der Daten

Der Interquartilabstand ist wie die Standardabweichung ein Mass für die Grösse des Bereichs der zentralen Daten.

Bei der **Normalverteilung** ist der halbe Interquartilabstand 0.67 SD.

Es gibt keine Masszahl, die unter allen Umständen optimal ist. Weitaus am verbreitetsten ist die Standardabweichung.

Mean ± SD

- Daten werden in medizinischen Zeitschriften oft als Mittelwert plus–minus Standardabweichung summarisiert.

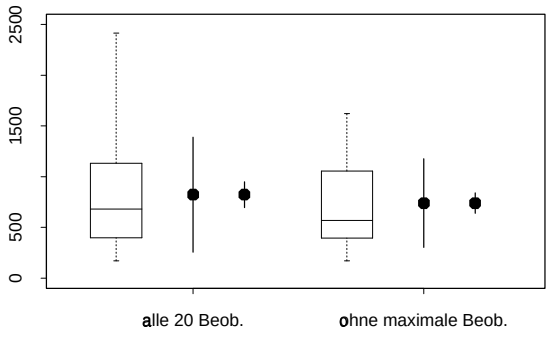
Mean ± SEM

- Der **Standardfehler** des Mittelwertes (standard error of the mean, SE($\overline{x}$), SEM) ist die Standardabweichung des Mittelwertes, beschreibt also die Streuung der Masszahl „Mittelwert“:

$$\text{SEM} = \text{SD}/\sqrt{n}$$

Der Standardfehler beschreibt **nicht** die Daten, sondern die Genauigkeit einer Schätzung.

Er hat in diesem Kapitel eigentlich nichts zu suchen. Der SEM wird trotzdem häufig anstatt der Standardabweichung SD verwendet, da eine kleinere Masszahl für die Variabilität besser wirkt (Vorsicht Falle!). Der folgende Plot zeigt die verschiedenen Masszahlen des Zentrums und der Variabilität am Beispiel der Anzahl von T_4–Zellen. Links: Boxplot (whiskers: Minimum–Maximum), Mitte: mean ± SD, rechts: mean ± SEM. In der rechten Hälfte wurde die maximale Beobachtung gestrichen, um einen Eindruck von der Abhängigkeit der Masszahlen von Einzelwerten zu geben.



Beispiel: Statistische Masszahlen für Anzahl T_4– und T_8–Zellen von $n = 20$ Hodgkin– und $n = 20$ non–Hodgkin–Patienten:

Descriptive Statistics

Disease		N	Minimum	Maximum	Mean	Std. Deviation
Hodgkin	Anzahl T4-Zellen	20	171	2415	823.20	566.385
	Anzahl T8-Zellen	20	212	1678	613.90	397.854
	Valid N (listwise)	20				
Non-Hodgkin	Anzahl T4-Zellen	20	116	1252	522.05	292.989
	Anzahl T8-Zellen	20	72	627	260.00	154.654
	Valid N (listwise)	20				

Statistics

Disease			Anzahl T4-Zellen	Anzahl T8-Zellen
Hodgkin	N	Valid	20	20
		Missing	0	0
	Percentiles	25	396.25	292.25
		50	681.50	447.50
		75	1185.00	829.75
Non-Hodgkin	N	Valid	20	20
		Missing	0	0
	Percentiles	25	330.00	143.25
		50	433.00	231.50
		75	727.00	327.50

3 Wahrscheinlichkeitsrechnung und Versuchsplanung

3.1 Ereignisse und ihre Wahrscheinlichkeiten

Wozu benötigen wir Wahrscheinlichkeitsrechnung? Oft sind wir nicht sicher, ob ein gewisses Ereignis eintreffen wird oder nicht. Beispiele sind eine 6 bei einmal würfeln oder schönes Wetter am nächsten Sonntag oder ob ein Patient tatsächlich Krebs hat, wenn ein verdächtiges Röntgenbild vorliegt. Die Wahrscheinlichkeitsrechnung ist eine mathematische Methode, um diese Unsicherheit quantitativ einzugrenzen. In der Statistik benötigen wir die Wahrscheinlichkeitsrechnung vor allem beim Hypothesentesten und für Konfidenzintervalle (siehe Kapitel 4). Dabei möchten wir z. B. nachweisen, dass ein empirischer Mittelwertsunterschied nur mit kleiner Wahrscheinlichkeit per Zufall auftreten kann (Signifikanztest). Weiter möchten wir um einen empirischen Wert herum Intervalle angeben, in denen der unbekannte wahre Wert mit 95% Wahrscheinlichkeit liegt (95%-Konfidenzintervalle). Insofern erlaubt uns die Wahrscheinlichkeitsrechnung Aussagen, die auf einer Stichprobe beruhen, auf die Grundgesamtheit zu verallgemeinern. Die Wahrscheinlichkeitsrechnung hat zudem Bedeutung für die stochastische Modellierung, z. B. die Ausbreitung von Epidemien oder in der Populationsgenetik.
Die Wahrscheinlichkeitsrechnung basiert auf Ereignissen, die mit einfachen Beispielen illustriert werden:

- **Ereignisraum** Ω = Menge der möglichen Ergebnisse (Elementarereignisse) eines Zufallsexperiments

  Beispiele:

  - Diagnose $\longrightarrow \Omega = \{$„krank“, „gesund“$\}$
  - Körpergrösse $\longrightarrow \Omega$ = Menge der positiven reellen Zahlen

- **Ereignis** E = Teilmenge von Ω

  Beispiele:

  - $E = \{$„gesund“$\}$ = Der Patient ist gesund.
  - $E = \{$Körpergrösse > 180 cm$\}$ = Der Patient ist grösser als 1.80 m.
  - $E = \{170 \text{ cm} \leq$ Körpergrösse ≤ 180 cm$\}$ = Der Patient ist zwischen 1.70 m und 1.80 m gross.
  - $E = \Omega$ = sicheres Ereignis. Beispiele: Der Patient ist krank oder gesund. Der Patient ist grösser als 0 cm.
  - $E = \emptyset$ (leere Menge) = unmögliches Ereignis. Beispiel: Der Patient ist gesund und krank.

  Für Ereignisse gelten die Rechenregeln der Mengenlehre.

Wahrscheinlichkeiten

Wahrscheinlichkeit = relative Häufigkeit in der Grundgesamtheit

Die Wahrscheinlichkeit wird mit P bezeichnet (für "probability"). Wenn die Wahrscheinlichkeit P für ein bestimmtes Ereignis E $\mathrm{P}(E) = 0.4$ ist, so tritt in der Grundgesamtheit das Ereignis E im Mittel in 40 von 100 Fällen auf. $\mathrm{P}(E)\times 100$ ist der Prozentsatz, $\mathrm{P}(E)$ die wahre relative Häufigkeit. Beim simplen Experiment eines Münzwurfes sind die möglichen Ereignisse die Augenzahlen 1, 2, 3, 4, 5, 6, und bei einem gleichmässigen Würfel gilt $\mathrm{P}(E = I) = 1/6$, wo $I = 1, 2, \ldots, 6$. Wenn ein Patient die Praxis betritt, so hat er mit einer gewissen Wahrscheinlichkeit eine der vier Blutgruppen, es gibt vier mögliche Ereignisse. Als Konvention gilt, dass $\mathrm{P}(E) = 0$ bedeutet, dass das Ereignis E unmöglich eintreten kann, bei $\mathrm{P}(E) = 1$ tritt das Ereignis sicher auf (z. B. tritt sicher eine Augenzahl von 1–6 bei einmal würfeln auf). Weiter gilt für das Ereignis „Nicht–E", (abgekürzt E^c) $\mathrm{P}(E^c) = 1 - \mathrm{P}(E)$, da ja entweder E oder E^c eintreten muss.

3.2 Bedingte Wahrscheinlichkeit und Unabhängigkeit

Von zentraler Bedeutung sind die Begriffe der bedingten Wahrscheinlichkeit und der stochastischen Unabhängigkeit.

Definition **bedingte Wahrscheinlichkeit** = Wahrscheinlichkeit, dass das Ereignis E_1 eintritt, gegeben, dass das Ereignis E_2 eingetreten ist.

Beispiele:

- Wahrscheinlichkeit, dass ein Patient einen Tumor hat, wenn ein auffälliges Röntgenbild vorliegt.
- Wahrscheinlichkeit, dass ein Kind grösser als 1.50 m ist, wenn bekannt ist, dass es 10.5 Jahre alt ist.

Definition **Unabhängigkeit von Ereignissen:** E_1 und E_2 sind dann (stochastisch) unabhängig, wenn die bedingte Wahrscheinlichkeit von E_1 gegeben E_2 gleich der Wahrscheinlichkeit von E_1 ist (d. h. E_2 hat keinen Einfluss auf E_1).

Aus der Unabhängigkeit der Ereignisse E_1 und E_2 folgt das Produktgesetz für deren Wahrscheinlichkeiten, das Sie vielleicht aus der Mittelschule kennen:
$P[E_1 \text{ trifft ein und } E_2 \text{ trifft ein}] = P[E_1 \cap E_2] = P[E_1] \times P[E_2]$.
Körpergrösse und Körpergewicht sind z. B. keine unabhängigen Variablen, während der Gesundheitszustand nacheinander aufgenommener Patienten meist unabhängig ist (siehe auch Abschnitt 3.9). Die Unabhängigkeit von Ereignissen hat wichtige Implikationen für die Gesetze der grossen Zahlen (siehe Abschnitt 3.5), und sie hilft, viele Formeln zu vereinfachen.

Beispiele:
Die Häufigkeit einer Krankheit bei einem Patientenkollektiv sei $p = 0.2$ (20%).

- Wahrscheinlichkeit, dass 3 aufeinanderfolgende Patienten an dieser Krankheit leiden:

$$\text{P[Patient 1 und Patient 2 und Patient 3 krank]} = 0.2 \times 0.2 \times 0.2 = 0.008$$

  Das heisst, diese Wahrscheinlichkeit ist nur noch 0.8%, also sehr klein.

- Wahrscheinlichkeit, dass von 3 aufeinanderfolgenden Patienten mindestens einer krank ist:

$$\text{P[mind. 1} \times \text{krank]} = 1 - \text{P[dreimal gesund]} = 1 - (0.8 \times 0.8 \times 0.8) = 0.49,$$

  da $\text{P}(E^c) = 1 - \text{P}(E)$. Das heisst, diese Wahrscheinlichkeit ist fast 50%.

3.3 Zufallsvariable und Verteilungen

Für die Statistik sind Zufallsvariablen wichtiger als Ereignisse:
Den Ausgang einer Messung oder einer Beobachtung können wir als Zufallsvariable X bezeichnen (die Bezeichnung X ist eine Konvention). Beim Würfeln kann X per Zufall die ganzzahligen Werte 1 bis 6 annehmen, beim Messen einer Körpergrösse sind es positive reelle Werte. Die Körpergrösse ist insofern eine Zufallsvariable, als wir nicht wissen, wie gross der nächste Patient sein wird.
Der erhaltene Wert ($X = x$) nach einer Messung heisst **Realisierung** der Zufallsvariablen X.

Definition **Stichprobe:** n Realisierungen einer Zufallsvariablen X, die uns interessiert: $x_1, \ldots, x_n$.

Beispiel: Wenn die interessierende Zufallsvariable die Körpergrösse 10–jähriger Knaben ist, kann eine Stichprobe aus allen 10–jährigen Knaben eines Schulhauses bestehen.

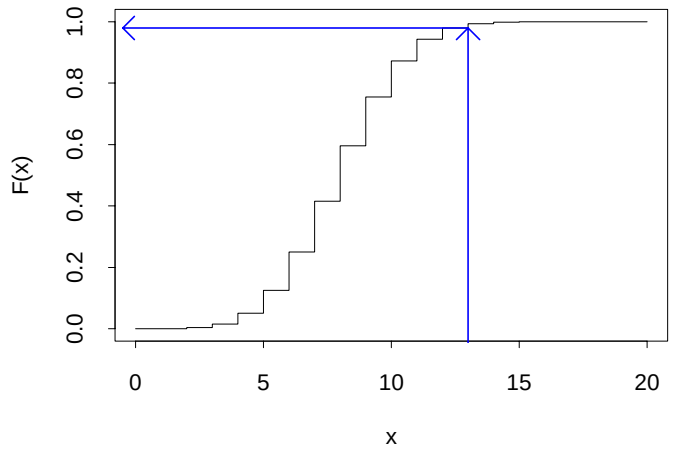
Die Wahrscheinlichkeitsrechnung beruht auf Ereignissen, und man muss deshalb für Zufallsvariablen passende Ereignisse konstruieren. Ereignisse der Form $X \leq z$ sind wichtig, da alle anderen interessierenden Ereignisse daraus konstruierbar sind. Dabei ist z eine von uns vorgegebene Zahl.

Definition **Verteilungsfunktion** F einer Zufallsvariablen X :

$$F(z) = \text{P}[X \leq z]$$

Die empirische (kumulative) Verteilungsfunktion, die bei den graphischen Methoden eingeführt wurde (siehe Abschnitt 2.4), ist das Datenäquivalent zum theoretischen Begriff der Verteilungsfunktion F(ein „Schätzer“ für die Verteilungsfunktion F). Verteilungsfunktionen sind monoton steigend von 0 auf 1. Beispiele, die wir nachher kennenlernen werden sind die Normalverteilung, die χ^2–Verteilung und die Binomialverteilung.

Für diskrete Zufallsvariable wird F eine Treppenfunktion; im Beispiel nimmt die Zufallsvariable X ganzzahlige Werte ab 0 an (F entspricht einer Binomial–Verteilung), definiert in Abschnitt 3.4:



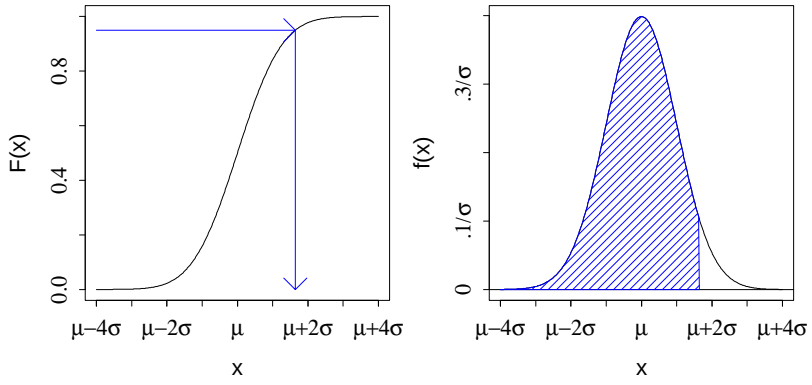
Werte kleiner als 0 kommen mit Wahrscheinlichkeit 0 vor. Werte grösser als 12 kommen mit Wahrscheinlichkeit 2.1% vor, denn $P[k < 13] = 97.9\%$ (siehe Grafik). Die Wahrscheinlichkeit für das Auftreten einer 2 ist 0.3% – es ist die Höhe der Stufe der Treppenfunktion bei $x = 2$.

Definition **Wahrscheinlichkeitsdichte** f:
a) diskrete Variable $X : f(z_i) = P[X = z_i]$
b) stetige Variable $X : f(z) = F'(Z)$ (Ableitung von F)

Das Histogramm ist ein Schätzer für die Wahrscheinlichkeitsdichte. Die Dichtefunktion ist visuell informativer als die Verteilungsfunktion, was z. B. die Form und Breite der Verteilung angeht, enthält aber grundsätzlich dieselbe Information wie die Verteilungsfunktion.

Dichtefunktionen sind nichtnegativ, und die Wahrscheinlichkeit, dass die Zufallsvariable X in einem Intervall $[a, b]$ liegt, ist die **Fläche unter der Kurve** über dem Intervall $[a, b]$.

Sie sehen nachstehend eine Graphik der Verteilungsfunktion der Normalverteilung (mit theoretischem Mittelwert μ und theoretischer Standardabweichung σ).
Die Körpergrösse ist approximativ normalverteilt (da multifunktionell genetisch) und Männer haben einen Mittelwert $\mu = 178$ cm und eine Standardabweichung $\sigma = 7.0$ cm. Informativer als die Verteilungsfunktion ist die Wahrscheinlichkeitsdichte $f = F'$ (rechts in der Graphik, „Glockenkurve").



Bei normalverteilten Grössen ist die Wahrscheinlichkeit, Messungen im Intervall $(\mu \pm \sigma)$ zu erhalten, 68% (siehe Abschnitt 2.5.1). Die Wahrscheinlichkeit für Messungen ausserhalb des Intervalls $(\mu \pm 3\sigma)$ ist sehr klein, nämlich ca. 0.3%. Die Wahrscheinlichkeit, dass eine Zufallsvariable x in einem Intervall a bis b liegt, ist gleich der Fläche der Wahrscheinlichkeitsdichte über diesem Intervall (oder gleich der Differenz $F(b)-F(a)$). Verteilungsfunktion F und Wahrscheinlichkeitsdichte f sind demnach geeignet, um mit Wahrscheinlichkeiten zu rechnen.

Im Abschnitt 2.5.1 haben Sie empirische Perzentile kennen gelernt. Bei den statistischen Tests werden wir theoretische Perzentile benötigen, und zwar die extremen. Typischerweise sind dies das 2.5% und das 5% Perzentil bzw. das 97.5% oder 95% Perzentil, sodass darunter bzw. darüber nur mit kleiner Wahrscheinlichkeit Resultate zu erwarten sind. In der Graphik der Normalverteilungs–Funktion können Sie sehen, wie man das 95% Perzentil findet: man startet bei $F(x) = 0.95$, geht horizontal zum Kreuzpunkte mit F und dann vertikal hinunter. Jetzt hat man das gesuchte 95. Perzentil, in Kapitel 4 jeweils mit $z_{0.95}$ abgekürzt. In der rechten Graphik können Sie sehen, wie Wahrscheinlichkeiten mit Hilfe der Dichte berechnet werden: Die Fläche unter der Dichte bis $z_{0.95}$ ist gleich 0.95.

Sie haben statistische Kennzahlen wie Mittelwert $\overline{x}$ und Standardabweichung s kennengelernt, um Daten einer Stichprobe zu charakterisieren (exakter: um deren empirische Verteilung zu charakterisieren). Es gibt auch die entsprechenden theoretischen Grössen, welche die wahre Verteilung in der Grundgesamtheit charakterisieren. Zu $\overline{x}$ gehört μ und zu s^2 gehört σ^2. In der Statistik sagt man, dass $\bar{x}$ ein Schätzer für μ ist, und s ein Schätzer für σ.

Erwartungswert $$\mu = \mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$$

Varianz $$\sigma^2 = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Standardabweichung $$\sigma = \sqrt{\mathrm{E}\left[(X - \mathrm{E}[X])^2\right]}$$

Im Falle diskreter Zufallsvariabler werden aus Integralen Summen (gegebenenfalls mit unendlich vielen Summanden): $\mu = \sum x_i \mathrm{P}[X = x_i]$

Der Begriff der Unabhängigkeit lässt sich von Ereignissen auf Zufallsvariable übertragen:

(stochastische) **Unabhängigkeit** von X und $Y \iff f_{XY}(x, y) = f_X(x)\, f_Y(y)$

Dabei ist f_{XY} die gemeinsame Dichte von (X, Y), die formal gleich wie die univariate Dichte definiert ist.

3.4 Wichtige Verteilungen

Verteilungen werden einerseits als Wahrscheinlichkeitsmodell für Daten gebraucht. Andererseits sind sie in der Statistik sehr wichtig beim Testen, wo wir die Verteilung statistischer Kenngrössen benötigen (z. B. wenn man entscheiden will, ob eine gemessene Differenz von zwei Mittelwerten wahrscheinlich oder unwahrscheinlich ist; siehe t–Test).
Es gibt eine sehr grosse Zahl von Verteilungen, die bedeutsam geworden sind. Hier seien 6 wichtige Verteilungen vorgestellt.

1. **Normalverteilung** $\mathcal{N}(\mu, \sigma^2)$.

$\mathcal{N}$ ist das Symbol der Normalverteilung (auch Gauss–Verteilung genannt), die folgende Dichte hat:

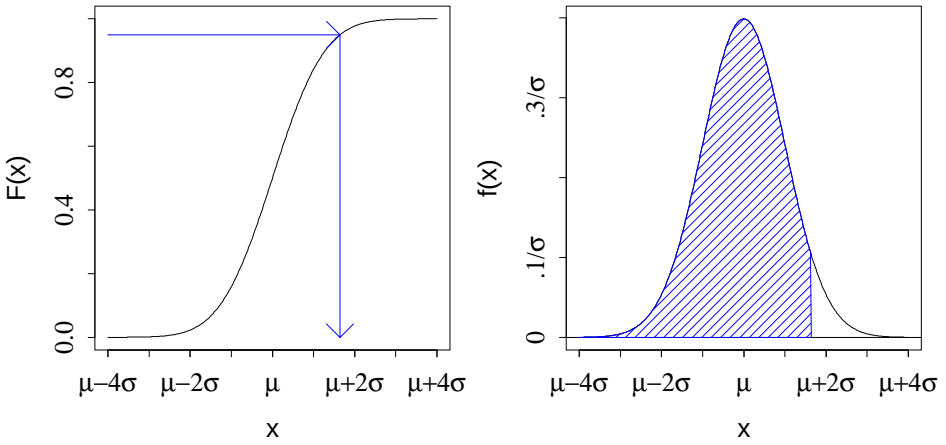
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Dabei ist „exp“ die Exponentialfunktion. Für $\mu = 0$ und $\sigma^2 = 1$ heisst die Verteilung Standardnormalverteilung. Die $\alpha \times 100\%$–Perzentile der Standardnormalverteilung werden konventionell mit z_α bezeichnet. Die Normalverteilung ist omnipräsent wegen des zentralen Grenzwertsatzes (siehe Abschnitt 3.5).

Eigenschaften:

- symmetrische Verteilung (siehe auch Plot der Dichte), Median = theoretischer Mittelwert
- charakterisiert durch 2 Parameter μ (= Populationsmittelwert) und σ (= Standardabweichung), die intuitiv besonders einfach sind. Wenn man diese kennt, kennt man das Wahrscheinlichkeitsgesetz.
- Die Wahrscheinlichkeiten für grosse Abweichungen vom Erwartungswert sind klein. Die Wahrscheinlichkeit von Abweichungen grösser als 2σ vom Erwartungswert μ beträgt etwa 5% („dünne Schwänze“ der Verteilung).

Hier noch einmal die Verteilungsfunktion und Dichte der Normalverteilung mit Illustration der Berechnung des 95. Perzentils.



2. χ^2–Verteilung

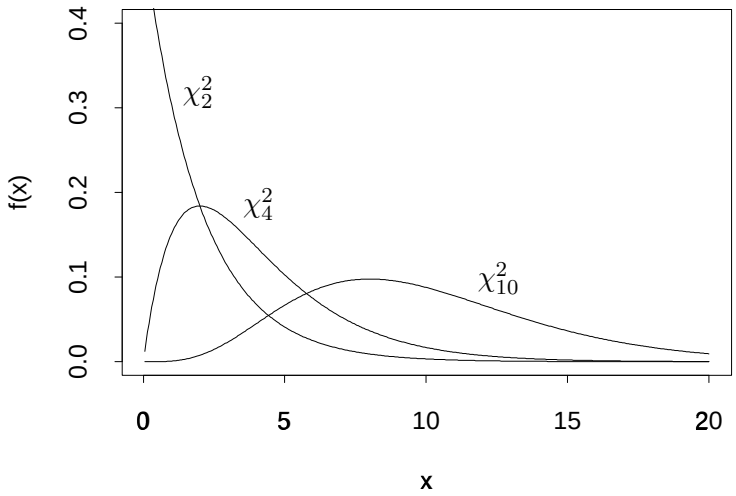
Die χ^2–Verteilung wird häufig gebraucht, wenn man in Kreuztabellen auf Unterschiede für kategorielle Daten testen will (siehe Kapitel 4). Dort werden Unterschiede zwischen beobachteten und hypothetischen Häufigkeiten quadriert und aufsummiert. Seien $Z_1, \ldots, Z_\nu$ unabhängige standardnormalverteilte Zufallsvariablen $\mathcal{N}(0,1)$ (d. h. Erwartungswert $\mu = 0$ und Varianz $\sigma^2 = 1$). Dann ist die χ^2–Verteilung definiert als Summe der quadrierten Z_i:

$$\chi^2_\nu = \sum_{i=1}^{\nu} Z_i^2 \quad \chi^2\text{–verteilt mit } \nu \text{ Freiheitsgraden } (\nu = \text{Anzahl Summanden})$$

Die Freiheitsgrade werden in Programmen oft mit „df“ (*degrees of freedom*) bezeichnet. Eigenschaften:

- asymmetrisch (rechts–schief)
- Erwartungswert und Varianz werden durch einen gemeinsamen Parameter ν beschrieben $\mu = \nu$, $\sigma^2 = 2\nu$.

Hier sehen Sie die Dichten der χ^2–Verteilung mit 2, 4 und 10 Freiheitsgraden (χ^2_2, χ^2_4, χ^2_{10}).

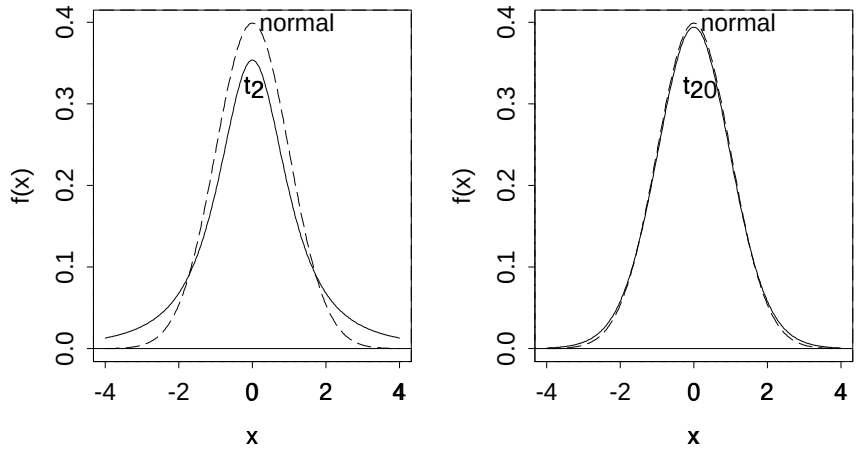


3. *t*–Verteilung

Die t–Verteilung wird unter anderem für den Test von Mittelwertsunterschieden und entsprechende Konfidenzintervalle gebraucht (Abschnitte 4.2 und 4.6).
Seien $X_1, \ldots, X_n$ unabhängige standardnormal–verteilte Zufallsvariablen $N(0,1)$.

Dann ist: $t = \dfrac{\bar{x}}{s/\sqrt{n}}$ t–verteilt mit $(n-1)$ Freiheitsgraden.

Die folgenden Abbildungen zeigen die Dichten der t–Verteilung mit 2 und 20 Freiheitsgraden im Vergleich zur Normalverteilung. Sie sehen, dass die t–Verteilung mit wenig Freiheitsgraden (d. h. n klein) beträchtlich von der Normalverteilung abweicht, für viele Freiheitsgrade wenig (für $n \longrightarrow \infty$ wird die Approximation dann exakt).



Instruktiv ist auch eine Tabelle für das 97.5%–Perzentil der t–Verteilung im Vergleich zur Normalverteilung. Für $\alpha = 0.05$ ergäbe die Normalverteilungsapproximation $z_{.975} = 1.96$, während wir für die t–Verteilung mit $n - 1$ Freiheitsgraden folgende Werte erhalten:

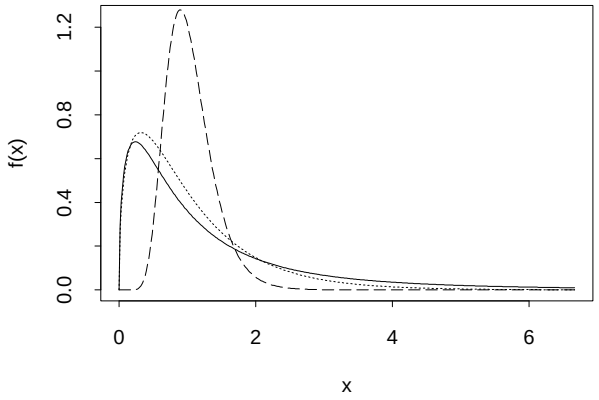
n	5	10	15	20	30	60	120	∞
$t_{.975}$	2.78	2.26	2.14	2.09	2.05	2.00	1.98	1.96

Offensichtlich ist die Approximation durch die Normalverteilung spätestens ab $n = 60$ gut.

4. ***F*–Verteilung**

Die F–Verteilung wird unter anderem in der Varianzanalyse (Abschnitt 6.1) gebraucht. Seien χ^2_m und χ^2_n zwei unabhängige χ^2–verteilte Zufallsgrössen mit m bzw. n Freiheitsgraden. Dann hat $F_{m,n} = \dfrac{\chi^2_m/m}{\chi^2_n/n}$ eine ***F*–Verteilung** mit (m, n) Freiheitsgraden.

Die folgende Abbildung zeigt die Dichten von $F_{m,n}$ für $(m, n) = (3, 5)$ (durchgezogene Linie), $(3, 50)$ (gepunktete Linie), und $(30, 50)$ (gestrichelte Linie).



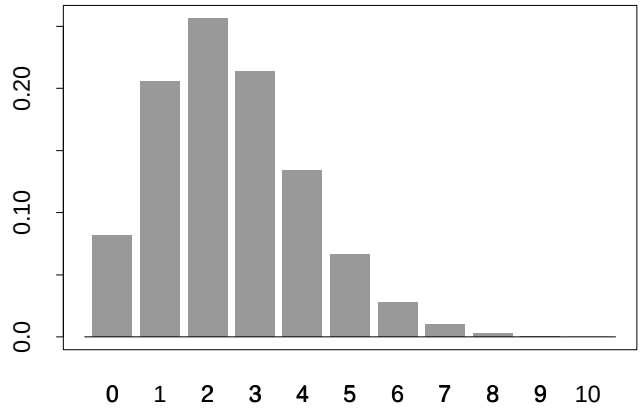
5. **Poisson–Verteilung**

Es handelt sich um eine diskrete Zufallsvariable, die ganzzahlige Werte $k = 0, 1, \ldots, \infty$ annimmt. Eine Poisson–Variable eignet sich demnach, um Daten zu beschreiben, die auf Zählungen beruhen, z. B. die Anzahl Klinikaufenthalte.

$$\boxed{\mathrm{P}[X = k] = \frac{\lambda^k}{k!} \exp(-\lambda)}$$

- Es gilt $\mu = \lambda$, $\sigma^2 = \lambda$, d. h. Erwartungswert und Varianz werden durch einen gemeinsamen Parameter beschrieben.
- Die Poisson–Verteilung ist u. a. relevant für die Modellierung seltener Ereignisse (radioaktiver Zerfall, Statistiken für Todesfälle durch seltene Krankheiten).

Hier sehen Sie die Einzelwahrscheinlichkeiten einer Poissonverteilung mit $\lambda = 2.5$:



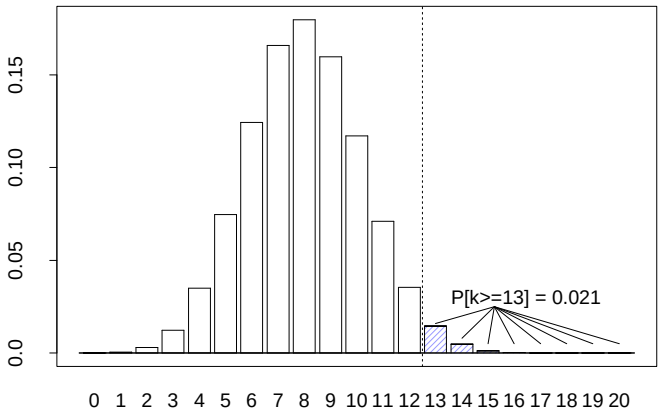
Beispiel: Modellierung der Anzahl Verbrechen an Voll– und Neumondtagen durch Poissonverteilung (Obs = beobachtet, Exp = erwartet nach Poisson–Modell mit Mittelwert 1.4 bei Vollmond und 0.5 bei Neumond)

Anzahl Verbrechen	Vollmondtage		Neumondtage	
pro Tag	Obs	Exp	Obs	Exp
0	40	45.2	114	112.8
1	64	63.1	56	56.4
2	56	44.3	11	14.1
3	19	20.7	4	2.4
4	1	7.1	1	0.3
5	2	2.0	0	0.0
6	0	0.5	0	0.0
7	0	0.1	0	0.0
8	0	0.0	0	0.0
9	1	0.0	0	0.0
Total Anzahl Tage	183	183.0	186	186.0

Wie man sieht, modelliert die Poissonverteilung die Daten gut. Da pro Person ein krimineller Akt ein seltenes Ereignis ist, erscheint dies nicht ganz überraschend.

6. **Binomialverteilung**

Bei der Einführung des Testens von Hypothesen (siehe Kapitel 4) wird an $n = 20$ Patienten geprüft, ob ein neues Medikament eine höhere Heilungswahrscheinlichkeit als $p = 0.4$ (d. h. 40%) hat. Die Anzahl k der geheilten Patienten ($k = 0$ bis 20 möglich) folgt einer Binomialverteilung. Falls man eine Wahrscheinlichkeit von $p = 0.4$ annimmt, ergibt sich nachfolgende Wahrscheinlichkeitsverteilung für die Anzahl Heilungen:



Dies bedeutet, dass 13 oder mehr Heilungen nur mit einer Wahrscheinlichkeit von 2.1% zu erwarten sind.

Allgemein hat die diskrete Dichtefunktion folgende Form:

$$\mathrm{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k} \quad 0 \le k \le n$$

- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, wobei $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$ (Binomialkoeffizient)

- Erwartungswert einer binomialverteilten Variablen $= n\,p$, Varianz $= n\,p\,(1-p)$

3.5 Gesetze der grossen Zahlen

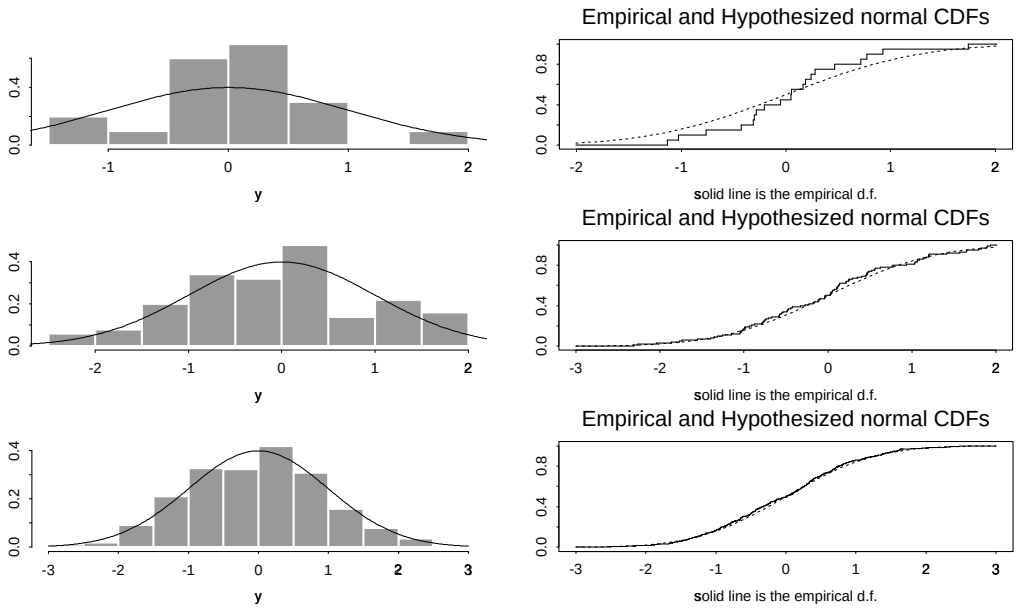
Eine Stichprobe von n unabhängigen Zufallsvariablen liefert Masszahlen, zum Beispiel den Mittelwert $\overline{x}$, die diese Stichprobe beschreiben. Eigentlich interessiert uns der zugrundeliegende Populationswert μ.

Fragestellung: Streben die empirischen Werte („Schätzer“, $\overline{x}$) gegen die wahren (theoretischen) Werte (μ), wenn die Stichprobe immer grösser wird ($n \longrightarrow \infty$)?

Die **Gesetze der grossen Zahlen** zeigen mathematisch, dass diese erwünschte Eigenschaft gilt. Der **zentrale Grenzwertsatz** besagt, dass die empirischen Kennzahlen approximativ normalverteilt sind, falls n gross ist.

Der zentrale Grenzwertsatz erklärt aber auch, weshalb die Normalverteilung so oft in der Natur gilt und für die Statistik wichtig ist:
Viele Phänomene der Natur entstehen durch Überlagerung vieler kleiner Effekte, so dass das Resultat etwa normalverteilt ist. Multiplikative Effekte führen hingegen auf eine schiefe Verteilung, z. B. die χ^2-Verteilung oder die lognormale Verteilung. Multiplikative Regeln haben wir z. B. bei Messungen, die durch die Zellteilung beeinflusst sind, weshalb viele biomedizinische Messgrössen nicht normalverteilt sind sondern schief verteilt, (z. B. Anzahl T_4- und T_8-Zellen in Abschnitt 2.4).

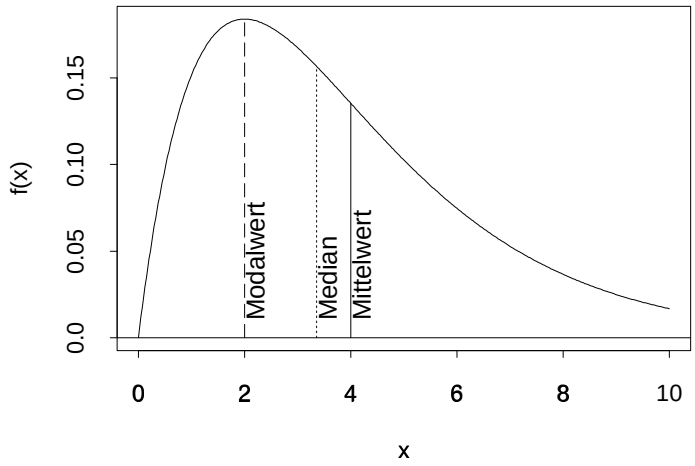
Nachstehend wird illustriert, wie das Histogramm und die empirische Verteilungsfunktion gegen die wahre Grösse (im Beispiel die Normalverteilung bzw. deren Dichte) konvergieren. Das Beispiel basiert auf simulierten Pseudo–Zufallsvariablen, $n = 20,\ 100,\ 500$.



Man sieht, dass das geschätzte Histogramm für kleine und mittlere Stichproben von der wahren Wahrscheinlichkeitsdichte abweichen kann. Entsprechend vorsichtig muss man bei der Interpretation sein.

3.6 Transformationen von Daten und Verteilungen

Bei vielen biochemischen Daten, aber auch bei anderen Variablen wie Körpergewicht oder Vermögen, haben wir bei weitem keine Normalverteilung. Vielmehr erhalten wir oft eine Dichte etwa der folgenden Form (Modalwert = Wert mit maximaler Wahrscheinlichkeitsdichte):



Die Verteilung ist **rechtsschief** (wie z. B. die χ^2–Verteilung) und nicht wie die Normalverteilung symmetrisch. Mittelwert, Median und Modalwert repräsentieren unterschiedliche Aspekte eines typischen Wertes der Verteilung. Ein Nachteil einer schiefen Verteilung ist diese Uneindeutigkeit des Zentrums der Daten; mindestens so gravierend ist, dass viele statistische Verfahren — die auf der Normalverteilung basieren — nicht mehr ohne weiteres anwendbar sind.

Konventionellerweise werden solche Variablen X in Medizin und Ingenieurwissenschaften oft logarithmiert, um eine approximativ symmetrische Verteilung zu erhalten: $Y = \log(X)$ Damit erreicht man, dass Standardverfahren der Statistik anwendbar sind, und zumeist wird auch die Interpretation der statistischen Analyse einfacher. Neben der logarithmischen Transformation gibt es auch andere Transformationen.

Falls die Verteilung rechtsschief ist, soll die Transformation grosse Werte klein machen, was z.B. durch folgende Transformationen erreicht wird:

$$\sqrt{X}, \quad \sqrt[3]{X}, \quad \log(X), \quad -\frac{1}{\sqrt{X}}, \quad -\frac{1}{X}$$

Die Transformation $\sqrt{X}$ ist die schwächste, $-1/X$ die stärkste. Welche ist in einer gegebenen Situation die richtige oder gar optimale? Man konsultiere dazu einen Statistiker. In einer unserer Evaluationen erwies sich für das Körpergewicht eine Transformation $-1/\sqrt{\text{Gewicht}}$ als angemessen, für den „body mass index“ (Gewicht/Grösse^2) die Transformation $-1/(\text{body mass index})$. Konventionell wurde in der Medizin die log–Transformation benutzt, wenn überhaupt eine zum Einsatz kam.
Damit erreicht man oft, dass konventionelle statistische Verfahren brauchbar sind, weil die Normalverteilung approximativ gilt. Die Alternative ist die Anwendung von Rangverfahren, auf Rängen der Daten beruhend, die aber nicht immer zur Verfügung stehen.

3.7 Schätzverfahren für statistische Kennwerte

Studien kosten Geld und Zeit, und Patienten sind nicht beliebig verfügbar. Man möchte deshalb die Daten statistisch effizient nutzen, **„gute Schätzer“** für interessierende wahre Kennwerte erhalten.
Angenommen, man möchte einen unbekannten Parameter θ schätzen, z. B. $\theta = \mu$, den theoretischen Mittelwert. Für einen Schätzer $\hat{\theta}$ eines Populationsparameters θ fordern wir mathematisch als Minimum, dass für wachsende Stichprobengrösse n sich der Schätzer $\hat{\theta}$ dem wahren Wert θ immer mehr annähert: Wenn man einen interessierenden wahren Parameter θ durch den empirischen Wert („Schätzer“) $\hat{\theta}$ bestimmt, so stellt man folgende Minimalforderungen:

- $\hat{\theta}$ soll mit wachsender Stichprobengrösse n den wahren Wert θ immer besser approximieren.
- $\hat{\theta}$ sollte für grosse n etwa normalverteilt sein.

Beides ist i. A. erfüllt, und man möchte zusätzlich – quantitativ – dass die Abweichung des Schätzwertes $\hat{\theta}$ vom wahren Wert θ möglichst klein ist.

Wenn der Schätzwert $\hat{\theta}$ systematisch zu hoch oder zu niedrig liegt, so nennt man diese Abweichung Bias (systematischer Fehler). Idealerweise ist der Bias null, und solche Schätzer sind „erwartungstreu“, d. h. in diesem Sinne optimal. Viele — aber nicht alle — Verfahren der Statistik erfüllen diese Voraussetzung. Zusätzlich möchte man, dass ein Schätzwert $\hat{\theta}$ nicht zu sehr variiert, von Stichprobe zu Stichprobe möglichst stabil bleibt. Optimal ist es, wenn die Varianz von $\hat{\theta}$ möglichst klein ist („Minimum–Varianzschätzer“).

Bei vielen praktisch wichtigen Problemen ist es nicht möglich, einen Schätzwert $\hat{\theta}$ zu berechnen, der zugleich keinen Bias (d. h. keine systematische Abweichung) und minimale Varianz erzielt. Man hat deshalb nach einem weiteren allgemeinen Berechnungsprinzip gesucht und es in der **Maximum–Likelihood–Schätzung** gefunden.

Für diejenigen unter Ihnen, die es gern etwas genauer wüssten, geben wir noch eine mehr mathematisch orientierte Erklärung (die Anderen können zu Abschnitt 3.8 weitergehen).

Ein Teil des Schätzfehlers kommt vom systematischen Fehler her, und dieser wird durch die Forderung der Erwartungstreue ausgeschlossen:

Kriterium 1: **Erwartungstreue** von $\hat{\theta}$ („im Mittel richtig“)

$$\boxed{\mathrm{E}\left[\hat{\theta} - \theta\right] = 0 \text{ oder } \mathrm{E}\left[\hat{\theta}\right] = \theta}$$

$\mathrm{E}\left[\hat{\theta} - \theta\right]$ heisst **Bias** von $\hat{\theta}$ („systematischer Fehler“).

Beispiele:

- Für die Ausfallstatistik eines medizinischen Gerätes wird jeden Tag festgehalten

  $x_i = 0$: Gerät nicht ausgefallen

  $x_i = 1$: Gerät ausgefallen

  Ein natürlicher Schätzwert $\hat{p}$ für die Ausfallwahrscheinlichkeit p ist

  $$\hat{p} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  Es gilt:

  $$\mathrm{E}[\hat{p}] = \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[X_i] = p$$

  Also ist $\hat{p} = \bar{x}$ erwartungstreu.

- Wenn man zwei Geräte zugleich betreibt, so interessiert man sich für die Wahrscheinlichkeit, dass beide zugleich kaputtgehen. Naiv würde man nehmen $\hat{p}^2 = \bar{x}^2$. Aber wegen

  $$\mathrm{E}[\bar{x}^2] = \mathrm{Var}(\bar{x}) + \mathrm{E}[\bar{x}]^2 = \frac{p(1-p)}{n} + p^2$$

  resultiert ein Bias $= \dfrac{p(1-p)}{n} \neq 0$, der für grosse n aber unbedeutend wird.

Kriterium 2: **Minimum–Varianz–Schätzung**

Konstruiere $\hat{\theta}$ so, dass

$$\boxed{\mathrm{Var}(\hat{\theta}) = \text{minimal}}$$

Denn es ist naheliegend, dass ein Schätzwert nicht zu variabel sein sollte.

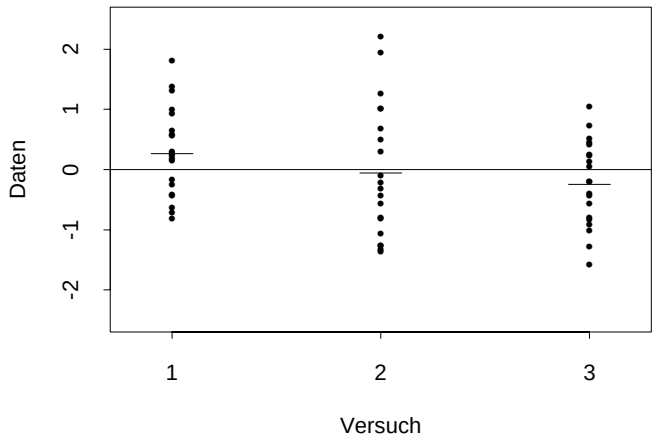
Erwartungstreue Schätzer kleinster Varianz sind meistens gut.

Leider sind solche Schätzer nicht immer konstruierbar und auch nicht immer optimal, deshalb wurde in der mathematischen Statistik nach einem weiteren allgemeinen Prinzip gesucht. Ein solches — und auch das am weitesten verbreitete — stellt die **Maximum–Likelihood–Schätzung** dar. Diese ist so definiert, dass die Übereinstimmung des Schätzers mit den Beobachtungen optimal wird. Dabei muss für die Daten eine Wahrscheinlichkeitsverteilung angenommen werden, z. B. die Normalverteilung.

3.8 Variabilität zwischen Datensätzen

Bei Wiederholung einer Studie erhalten wir andere statistische Kennzahlen. Dies ist erklärbar durch die unterschiedlichen Stichproben, was notwendigerweise einen Zufallseffekt mit sich bringt. Man möchte diese Zufallsschwankungen in den statistischen Kennzahlen quantitativ fassen (Achtung: bisher hatten wir mittels Varianz bzw. Standardabweichung Variabilität in **einem** Datensatz erfasst!). Dazu dient wieder die Varianz bzw. die Standardabweichung ($= \sqrt{\text{Varianz}}$); die Standardabweichung von Kennzahlen wird aber Standardfehler genannt (standard error, SE).

Sie sehen nachstehend 3 auf dem Computer simulierte normalverteilte Stichproben mit $\mu = 0,\ \sigma^2 = 1$ und $n = 20$.



Wie Sie an nachfolgender Tabelle sehen, schwanken die Mittelwerte $\bar{x}$ um den Wert $\mu = 0$ und die Standardabweichungen s um den Wert $\sigma = 1$, mit zum Teil merklichen Abweichungen.

	1.	2.	3.
$\bar{x}$	0.27	-0.06	-0.25
s	0.74	1.10	0.71

Mit den Methoden der Wahrscheinlichkeitsrechnung (siehe dieses Kapitel) kann man die Varianz des Mittelwertes herleiten:

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{n},$$

wobei σ^2 die Populations–Varianz einer Einzelmessung ist. Durch Ersetzen von σ durch den Schätzwert s erhält man den Standardfehler (standard error of the mean, SEM).

$$s_{\bar{x}} = s \,/\, \sqrt{n}$$

Man muss zwischen s (= Standardabweichung = Variabilität in der Stichprobe) und $s_{\bar{x}}$ (=

Standardfehler = Standardabweichung des Mittelwerts = Variabilität des Mittelwerts) klar unterscheiden. Beide haben ihren – unterschiedlichen – Platz. Für die log–T_4–Zellanzahl bei Hodgkin–Patienten erhielten wir in Kapitel 2 $\bar{x} = 6.49$ und $s = 0.71$. Bei $n = 20$ ergibt sich ein $s_{\bar{x}} = 0.16$.

Beispiel: Genauigkeit von Verhältnissen

$n = 80$ Personen wurden über Asthma befragt

$k = 7$ Asthmatiker wurden gefunden

Also ist die relative Häufigkeit von Asthma: $\hat{p} = k/n = 0.088$, in der Epidemiologie Prävalenz genannt.

Wie genau ist $\hat{p}$ bestimmt (n ist klein)? Dies basiert auf der Binomialverteilung, da eine ja/nein–Befragung eine binäre Variable ergibt:

$$\begin{aligned} \mathrm{Var}(\hat{p}) &= \frac{p(1-p)}{n} \\ \Longrightarrow s_{\hat{p}} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \end{aligned}$$

$$\begin{aligned} \text{Prävalenz–Beispiel: } s_{\hat{p}} &= 0.032 \\ \text{falls } \hat{p} = 0.5\text{: } s_{\hat{p}} &= 0.056 \end{aligned}$$

Die Formel für $s_{\hat{p}}$ zeigt, dass mittlere Häufigkeiten mehr streuen als Extreme.
Wie wir nachher sehen werden, liegt die wahre Prävalenz etwa im Intervall $(\hat{p} \pm 2s_{\hat{p}})$, d. h. (0.088 ± 0.064), sie ist also schlecht bestimmt.

Bei beiden Beispielen verbessert sich die Genauigkeit des Kennwerts mit $1/\sqrt{n}$ und nicht etwa mit $1/n$ („Wurzel–n–Gesetz“). Dies gilt für fast alle statistischen Fragestellungen und hat grosse praktische Konsequenzen: Um eine Verdoppelung der Genauigkeit der späteren Datennanlayse zu erhalten, müssen viermal soviele Patienten in die Studie einbezogen werden. Umgekehrt steigt auch der Aufwand für grosse Studien stärker als poportional zu n, so dass eine seriöse Versuchsplanung angezeigt ist, gerade was die Wahl der Stichprobengrösse angeht. (Siehe auch Kapitel 4.)
Die Streuung eines Schätzers — den Standardfehler — erhalten wir bei den meisten Programmpaketen mitgeliefert und damit auch ein Bild über die Genauigkeit der Statistik.

3.9 Versuchsplanung

3.9.1 Repräsentative Stichprobe

Bei jeder Studie wollen wir — implizit oder explizit — über eine Grundgesamtheit (Population) von Versuchseinheiten (Menschen, Tiere, Proben, Spitäler) Aussagen machen. Üblicherweise wird eine Stichprobe, also nicht eine Population, in eine Studie einbezogen, d. h. es werden n Versuchseinheiten aus der Grundgesamtheit gezogen. Die Stichprobe bedingt den Zufallseffekt. Sehr wichtig ist die Wahl der Stichprobengrösse, das „n“. Es gibt dazu viel Literatur; einige Gesichtspunkte werden in Kapitel 5 diskutiert. Selten unternimmt man auch eine Gesamterhebung. Beispiele sind die Volkszählungen oder Krankenregister (z. B. Krebsregister des Kantons Zürich). Gesamterhebungen bedingen einen hohen Aufwand, und es können nur relativ grobe Parameter erfasst werden. Deshalb wird neben der Volkszählung ein Mikrozensus durchgeführt, um detailliertere sozialwissenschaftliche Ergebnisse zu erhalten.
Das Beispiel eines Abstracts aus dem „British Medical Journal“ von 1996 zeigt, dass Studien detailliert bezüglich ihres Versuchsplans („design“) umschrieben werden müssen, um akzeptiert zu werden.

Adequacy of cervical cytology sampling with the Cervex brush and the Aylesbury spatula: a population based randomised controlled trial

Paola Dey, Stuart Collins, Minaxi Desai, Ciaran Woodman

Abstract

***Objective*—To compare the adequacy of cervical cytology sampling with two sampling instruments commonly used in primary care—namely, the Aylesbury spatula and the Cervex brush.**

***Design*—Pair matched, population based randomised controlled trial.**

***Setting*—86 general practices and family planning clinics in Greater Manchester.**

***Subjects*—15 882 cervical smears taken from women aged 20-64 years as part of the national cervical screening programme.**

***Interventions*—Participating centres were allo-**

Introduction

Inadequate cervical smears are not only a cause of needless anxiety and inconvenience to women but are also an additional cost to the NHS. Attempts to reduce the rate of inadequate smears have focused on improving the competence of smear takers and the design of sampling instruments, but these instruments have rarely been evaluated in population based settings. We compared the adequacy of cytological sampling in a primary care setting with two commonly used instruments, the Cervex brush and the Aylesbury spatula.

Wie soll man die Stichprobe auswählen („ziehen“) ?

- Sie soll repräsentativ für die Population sein (Verallgemeinerungsfähigkeit).
- Wenn mehrere Gruppen untersucht werden, sollten sie in den wesentlichen Merkmalen vergleichbar sein (Beispiel: Alter, Schweregrad der Krankheit).
- Unabhängigkeit der Versuchseinheiten muss gewährleistet sein; dies schliesst Familienangehörige aus, da diese Genetik und Umwelt teilen.
- Es gibt wichtige Ausnahmen von der Unabhängigkeit:
  prä–post–Vergleiche, Versuche mit Messwiederholungen, Longitudinalstudien.

Die Repräsentativität für die Population wird durch ein Bündel von Massnahmen sichergestellt. Am wichtigsten ist die **Randomisierung**, was bedeutet, dass Patienten nach dem Zufallsprinzip einer oder mehreren Gruppen zugewiesen werden. Aber auch auf eine möglichst volle Teilnahme bis zum Schluss ist sehr zu achten.
Wegen der mangelnden Randomisierung lagen die Wahlprognosen in den USA (1948) zwischen Truman und Dewey völlig daneben. Aufgrund (bequemer) Telefonumfragen wurde ein klarer Sieg von Dewey prognostiziert, gewonnen hat aber Truman. 1948 waren eben Privattelefone nicht gleichmässig über alle Bevölkerungsschichten verteilt, so dass sich eine Verfälschung zugunsten der Mittel– und Oberschicht einschlich.

Merke:

- Freiwillige sind nicht repräsentativ.
- Verweigerer sind nicht repräsentativ.
- Patienten von Uni–Kliniken sind nicht repräsentativ für stationäre Patienten an sich (z. B. Kreisspitäler).

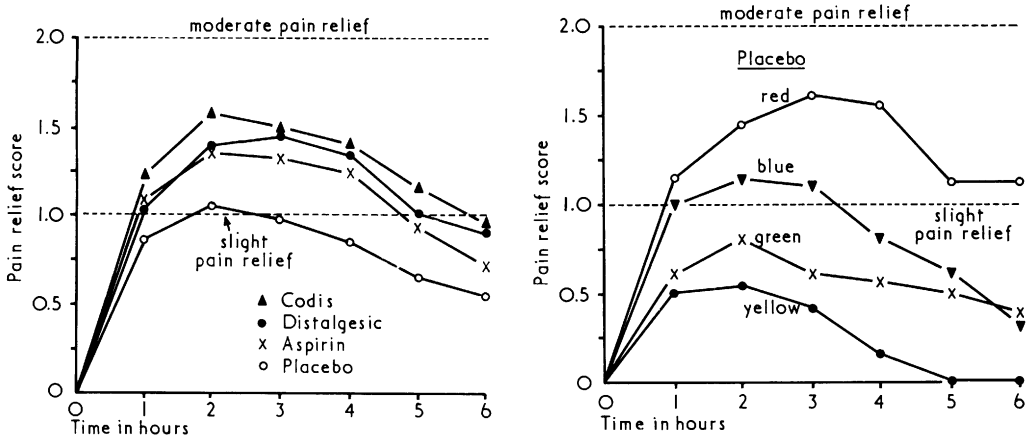
3.9.2 Arten von Studien

- Beobachtungsstudien werden experimentellen Studien gegenübergestellt. Bei letzteren wird eine Zielgrösse beeinflusst, es wird interveniert. Beispiele aus der Medizin sind:
  - Therapiestudien
  - Neurophysiologische Studien
  - Tierexperimente

  Beispiele von Beobachtungsstudien sind:
  - Bestimmung der Prävalenz einer Krankheit
  - Studie der kardiovaskulären Risikofaktoren
  - Vergleich dementer und gesunder alter Menschen
- Des weiteren werden prospektive und retrospektive Studientypen unterschieden. Retrospektive Studien aus Krankenblatt–Archiven werden wegen ihrer beschränkten Aussagekraft nicht mehr so häufig durchgeführt. In der Epidemiologie sind prospektive Studien durch Kohortenstudien, retrospektive durch Fall–Kontroll–Studien („case–control studies“) repräsentiert.
- Eine wichtige Unterscheidung besteht noch zwischen Querschnitts– und Längsschnitt–Studien. Letztere werden immer wichtiger, da z. B. aus dem Verlauf einer Krankheit wichtige Erkenntnisse gewonnen werden können. Es steigt dabei nicht nur der empirische, sondern auch der statistische Aufwand.

Klinische Therapiestudien

Ziel ist der Nachweis der Wirksamkeit bzw. Unwirksamkeit einer Therapie, oder die Überlegenheit einer neuen Therapie über eine Standardtherapie. Im ersteren Fall wird im allgemeinen eine Placebogruppe geführt, da allein die Verabreichung irgendeiner Pille einen (psychologischen) Effekt haben kann. Zielgrösse ist der Therapieerfolg, der qualitativ („geheilt“) oder quantitativ („Veränderung des Blutdrucks“) erhoben werden kann.
Die Resultate einer Studie im British Medical Journal 1974 über die Wirkung von Schmerzmitteln bei Arthritis (siehe Graphiken unten) zeigen, dass Placebo einen nicht sehr viel schwächeren Effekt hat als z. B. Aspirin. Wenn man die Farbe von Placebo variiert, erhält man verblüffende Unterschiede, die über die therapeutischen Differenzen hinausgehen.



Die verschiedenen Gruppen („Therapiearme“) sollten in allen wichtigen Punkten (Alter, Geschlecht, Schweregrad der Krankheit) vergleichbar sein. Die Randomisierung sollte dies approximativ garantieren; raffinierte Versuchspläne randomisieren „geschichtet“, z. B. nach Altersklasse. Trotzdem sollte man mögliche „Störgrössen“ (eben z. B. den Alterseffekt) miterheben. Man kann diese nämlich auch später, als sogenannte Kovariaten, in der statistischen Analyse quantitativ berücksichtigen.

Zusammengefasst einige Prinzipien der Versuchsplanung:

Randomisierung:
- Gleiche Chance für alle (einer Population), in eine Stichprobe zu kommen (Repräsentativität)
- gleiche Chance für alle (einer Stichprobe), in eine Gruppe zu kommen (Vergleichbarkeit)

Standardisiertes Vorgehen:
- klare Einschluss–/Ausschlusskriterien
- klare diagnostische und experimentelle Bedingungen
  Beispiel: In der Neurophysiologie ist genau festgelegt, wie und wo Elektroden auf die Kopfhaut zu kleben sind.

Doppelverblindung:
- Verfälschung durch Subjektivität vermeiden.
  Beispiel: Bei einer Therapiestudie sollte der Patient (einfach–blind) und wenn möglich auch der Arzt (doppel–blind) nicht wissen, welches Medikament gegeben wird.

Kontrolle:
- neue Methode mit Placebo oder Standardtherapie vergleichen

Unabhängigkeit der Versuchseinheiten:
- Beine eines Versuchstieres sind nicht unabhängig.

Adäquate Stichprobengrösse:
- Zu kleine Studien können keine klaren Ergebnisse liefern; zu grosse Stichproben sind unter Umständen unethisch, wenn Versuchspersonen z. B. weiterhin Placebo erhalten, obwohl die Wirksamkeit des neuen Medikaments längst feststeht. Wie man die Stichprobengrösse gut wählt, wird ab Seite 53 illustriert.

Einfache Versuche:
- nur **zwei** Gruppen oder **zwei** Zeitpunkte vergleichen

„informed consent“:
- Ohne das Einverständnis einer Ethikkommission sind Versuchsergebnisse nicht mehr publizierbar.

3.9.3 Wo kommt der Zufall her ?

Gründe für unterschiedliche Ergebnisse können sein:
1. systematisch („Bias“) oder
2. zufällig („Variabilität“)

Beispiele für systematische Diskrepanzen

- Messinstrument verstellt
- in Gruppe A mehr alte Patienten als in Gruppe B
- zwei behandelnde Ärzte mit unterschiedlichen („nicht operationalisierten“) Kriterien

Vermeidung systematischer Abweichungen durch:

- angemessene(n) Versuchsplan / Versuchsdurchführung
- statistische post–hoc Prüfungen (eventuell Kovariaten berücksichtigen)
- Trendanalysen
- Analyse der Untersuchereffekte

Zufällige Schwankungen

Zufällige Schwankungen enthalten neben dem Messfehler eine Vielzahl biologischer Schwankungen. Wenn man alle möglichen Einflussfaktoren unter Kontrolle hätte — wie angenähert in physikalischen Experimenten, wo man z. B. Druck, Temperatur und Magnetfeld konstant halten kann — gäbe es nur noch den Messfehler. Die möglichen Schwankungen zwischen Patienten sind genetischer Art (dies macht oft einen grossen Teil aus), können aber auch der Umwelt zugeschrieben werden (Essen, Schlaf) oder der Lebensgeschichte (früherer Stress, Übergewicht).
Beispiel: Das Geburtsgewicht eines Kindes ist teilweise genetisch bestimmt, hängt aber auch von der Grösse der Mutter ab (kleine Mütter haben statistisch häufiger kleine Babies). Es hängt z. B. weiter vom Essen der Mutter ab, ob sie raucht, oder eine chronische Erkrankung hat.
Man versucht die zufälligen Schwankungen möglichst klein zu halten, um damit schärfere Aussagen zu erhalten. Einige Möglichkeiten sind:

- Standardisierung der Messmethode
- Kontrolle der potentiellen Einflussfaktoren
- Homogenisierung der Grundgesamtheit durch Ein– und Ausschlusskriterien (z. B. Alter, Schweregrad der Krankheit)

Die stochastische Modellierung im biomedizinischen Bereich ist nicht primär die Modellierung des Zufalls, sondern die Modellierung von Komplexität. Die vielen Einflussfaktoren, die man nicht alle erfassen kann, auch gar nicht alle kennt, werden — mit Erfolg — als stochastisch und nicht als deterministisch aufgefasst.

4 Prüfung von Hypothesen

4.1 Was ist ein statistischer Test?

Bei der Analyse einer Stichprobe erhalten wir mittels statistischer Kennwerte nie Sicherheit über einen Sachverhalt, denn Kennwerte schwanken von Stichprobe zu Stichprobe. Statistische Tests schränken diese Unsicherheit quantitativ ein: Es sind Entscheidungsregeln, ob eine wissenschaftliche Hypothese mit grosser Wahrscheinlichkeit zutrifft. Man möchte *objektiv und quantitativ* beurteilen, ob eine Differenz oder ein Kennwert zufällig so herausgekommen ist, oder ob durch eine experimentelle Bedingung (z. B. eine Therapie) ein systematischer Effekt vorliegt. *Subjektive* Beurteilungen sind anfechtbar, weil die selbe Datenlage durch einen Forscher eher optimistisch, durch einen anderen eher pessimistisch eingeschätzt wird.

Statistische Tests werden heute in der medizinischen Literatur standardmässig angewandt. Im Kapitel „Subjects and Methods“ einer wissenschaftlichen Arbeit gibt es üblicherweise einen Abschnitt, der darlegt, welche statistischen Methoden zum Einsatz kommen. Nachstehend ein Beispiel aus dem „British Medical Journal“ von 1996. Untersucht wird die Frage, ob der mittlere Cholesterinwert oder eine Veränderung des Cholesterinwerts das Suizidrisiko beeinflussen. Bei solch nicht naheliegenden Fragestellungen ist das Bedürfnis nach einer statistischen Prüfung besonders evident. Die Spalte „P value“ gibt an, mit welcher Wahrscheinlichkeit die gefundenen Differenzen durch Zufall zustande kommen konnten.

STATISTICAL ANALYSIS

We used Student's *t* test and Cox's proportional hazards model for analysis. All analyses were stratified by the number of measurements of serum cholesterol for each subject (three to five).

Table 1—*Relative risks (95% confidence interval) of suicide among 6393 men by average serum cholesterol concentration and change in cholesterol concentration*

	No of subjects	No of suicides	Adjusted relative risk (95% confidence interval)*	P value
Average serum cholesterol concentration (mmol/l)†				
<4.78	827	10	3.16 (1.38 to 7.22)	0.007
4.78-6.21	3600	13	1.00	
>6.21	1966	9	1.28 (0.55 to 3.01)	0.56
Change in serum cholesterol concentration (mmol/l a year)‡				
Decline >0.13	1143	11	2.17 (0.97 to 4.84)	0.056
Change ≤0.13	2795	13	1.00	
Increase >0.13	2455	8	0.72 (0.30 to 1.72)	0.46

*Relative risks for average cholesterol concentration were adjusted, using Cox's proportional hazards model, for age, smoking habits (never, former, or current), and mean corpuscular volume at first examination. Relative risks for change in cholesterol concentration were adjusted as above and for average serum cholesterol concentration.

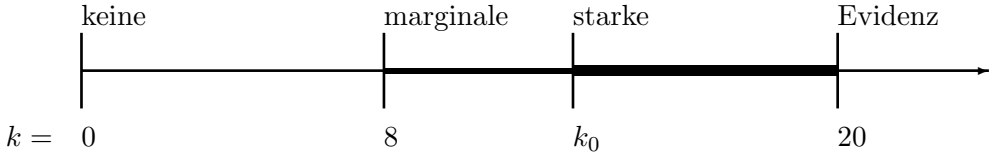
†Mean of serum cholesterol concentrations from all examinations.

‡Estimated using within person linear regression method (0.13 mmol/l equivalent to 5 mg/dl).

Einführendes Beispiel: Ein Standardmedikament wirkt in 40% aller Fälle. Ist ein neues Medikament besser?

In einer Studie wurden $n = 20$ Patienten mit dem neuen Medikament behandelt. Falls alt und neu gleich gut wären, würden im Mittel $k = 8$ Patienten geheilt. Falls weniger als 8 Patienten geheilt werden, spricht statistisch nichts für das neue Medikament. Falls es gerade etwas mehr als 8 Patienten sind, bleiben wir unsicher, ob dies nicht zufällig so ist. Ab einem gewissen k_0, das deutlich grösser als 8 ist, glaubt man subjektiv an die Überlegenheit des neuen Medikaments. Kann man dies wahrscheinlichkeitsmässig erhärten?

Die Evidenz, dass für die Erfolgswahrscheinlichkeit $p_{neu} > 0.4$ gilt, ist für verschiedene mögliche Resultate von $k =$ „Anzahl geheilter Patienten“ etwa wie folgt:



Frage: Wie wahrscheinlich ist ein $k \geq k_0$, falls doch $p_{neu} = 0.4$ gilt? Anders: Mit welcher Wahrscheinlichkeit schliesst man auf Verbesserung, obwohl keine da ist?
Falls $p_{neu} = 0.4$ gilt, ist k binomialverteilt mit $p = 0.4$. Daraus ergeben sich folgende Wahrscheinlichkeiten (aus Tabellen der Binomialverteilung):

$$\begin{aligned} \mathrm{P}[k \geq 11] &= 0.128 \\ \mathrm{P}[k \geq 12] &= 0.057 \\ \mathrm{P}[k \geq 13] &= 0.021 \\ \mathrm{P}[k \geq 14] &= 0.006 \end{aligned}$$

Was schliessen wir daraus? Falls das neue Medikament nicht besser ist (d. h. $p_{neu} = 0.4$), dann ist es sehr unwahrscheinlich, dass mehr als 12 Patienten geheilt werden. Falls wir also 13, 14 oder gar mehr Heilungen feststellen, dann ist das neue Medikament mit grosser Wahrscheinlichkeit besser (d. h. $p_{neu} > 0.4$). Angenommen man führt die Studie oft durch. Dann würden nur in ca. 2% der Studien mehr als 12 Patienten geheilt werden, obwohl das neue Medikament nicht besser ist.

Allgemeine Formulierung:

Ein Studienplan startet mit einer wissenschaftlichen Hypothese. Diese kann aus klinischen oder wissenschaftlichen Erfahrungen entstehen oder der Literatur entnommen werden.

H_1 : Wissenschaftliche Hypothese oder Alternativhypothese

Beispiel: $\mathrm{H}_1 : p_{neu} > 0.4$

Eine Alternativhypothese der Form $p_{neu} > 0.4$ heisst einseitig, weil man sich für Abweichungen in nur einer Richtung interessiert (hier: Verbesserung). Falls man die wissenschaftliche Hypothese $p_{neu} \neq 0.4$ (Verbesserung oder Verschlechterung) prüfen möchte, so nennt man dies eine zweiseitige Alternative.

Sowohl die wissenschaftliche Hypothese als auch die anschliessend zu formulierende statistische Hypothese (oder Nullhypothese) beziehen sich auf die unbekannten wahren Werte, nicht auf statistische Kennwerte (diese sind ja bekannte Zahlen!).

H_0 : Statistische Hypothese oder Nullhypothese
Beispiel: $H_0 : p_{neu} = 0.4$

Ein statistischer Test prüft wahrscheinlichkeitsmässig die Nullhypothese, obwohl man bei der Planung einer Studie von der wissenschaftlichen Hypothese ausgeht. Die Schlussweise ist indirekt: Wenn die Nullhypothese aufgrund der Daten mit hoher Wahrscheinlichkeit abgelehnt werden kann, entscheidet man sich aufgrund dieser Evidenz für die wissenschaftliche Hypothese.
Man hat demnach zwei mögliche Entscheidungen zur Auswahl:

- verwerfe H_0 und bejahe H_1 oder
- verwerfe H_0 nicht und betrachte H_1 als nicht nachgewiesen.

Man hat dabei gemäss nachstehender Tabelle zwei Möglichkeiten, richtig zu entscheiden, und zwei Fehlermöglichkeiten.

	Wahrheit	
Entscheidung	H_0 stimmt	H_0 stimmt nicht
H_0 nicht verworfen	richtig	Fehler 2. Art: „β“
H_0 verworfen	Fehler 1. Art: „α“	richtig

Es gibt eine Analogie zum Gerichtsverfahren eines nicht geständigen Angeklagten, wo auch eine Entscheidung bei Unsicherheit zu fällen ist.

	Gerichtsverfahren	Hypothesentesten
braucht starke Evidenz	Schuldspruch	neue Hypothese bejahen
Nullhypothese H_0	nicht schuldig	alte Theorie stimmt
Alternativhypothese H_1	schuldig	neue Theorie stimmt
Haltung	plädiere nicht schuldig ohne starke Evidenz für Schuld	behalte Nullhypothese, ausser sie sei mit den Daten sehr unverträglich

Wir führen jetzt wichtige statistische Begriffe ein:

Definition **Irrtumswahrscheinlichkeit** oder **Signifikanzniveau** eines Tests $= \alpha$
= (maximale) Wahrscheinlichkeit eines Fehlers 1. Art
= Wahrscheinlichkeit, neue Therapie oder Theorie als besser zu betrachten, obwohl dies nicht der Fall ist.

Konventionell wird $\alpha = 5\%$ oder selten 1% vorgegeben. Wenn die Daten beim Testen eine kleinere Wahrscheinlichkeit als α ergeben, lehnt man die Nullhypothese ab und nimmt die wissenschaftliche Hypothese an.
Meist geben Programmpakete nicht an, ob ein Resultat signifikant oder nicht signifikant ist. Stattdessen geben Sie einen p–Wert.

Definition ***p*–Wert** eines Tests

1. korrekt: $p =$ kleinstes α, das noch signifikant würde.
2. lax, aber verständlicher: $p =$ Wahrscheinlichkeit, gegeben H_0 stimme, für die Testgrösse einen so grossen oder grösseren als den aus den Daten berechneten Wert zu erhalten.

Falls $p \leq \alpha$, wird die Null–Hypothese abgelehnt (und die wissenschaftliche Hypothese angenommen); falls $p > \alpha$, wird die Null–Hypothese nicht abgelehnt. In der Praxis wird aber ein ganz kleiner p–Wert (z. B. < 0.0001) als „starke“ Ablehnung von H_0 gewertet und als „hochsignifikant“ eingestuft. Auch wenn dies nicht ganz korrekt ist, ist eine solche Interpretation tolerierbar.

Definition **Trennschärfe** oder **Macht** („power“) eines Tests $= 1 - \beta$
$= 1-$Wahrscheinlichkeit eines Fehlers 2. Art
= Wahrscheinlichkeit, neue Therapie oder Theorie als besser nachzuweisen, wenn sie tatsächlich besser ist.

Die Macht eines Tests hängt von der Stichprobengrösse n und der **Effektgrösse** ab.

Wir illustrieren die Begriffe mit dem Medikamentenbeispiel, indem wir ein Signifikanzniveau von $\alpha = 0.05$ vorgeben:

- Falls das Ergebnis $k = 13$ Heilungen ist, erhalten wir $p = \mathrm{P}[k \geq 13] = 0.021$ („p–Wert“). Da demnach $p \leq \alpha = 0.05$ gilt, können wir H_0 verwerfen und mit 5% Irrtumswahrscheinlichkeit schliessen, dass das neue Medikament besser ist.

- Falls aber nur $k = 12$ Patienten geheilt werden, ergibt sich $p = \mathrm{P}[k \geq 12] = 0.057$, so dass $p > \alpha = 0.05$. Man kann also H_0 nicht verwerfen, der statistische Nachweis der Überlegenheit des neuen Medikaments ist nicht gelungen.

Eventuell war die Stichprobe in letzterem Fall nicht gross genug gewählt, um eine relevante Verbesserung mit genügender Wahrscheinlichkeit festzustellen (zu kleine Trennschärfe). Deswegen darf man in einem solchen Fall auch nichtsagen, dass statistisch bewiesen wurde, dass das neue Medikament nicht besser ist. Richtig ist zu sagen, dass man eine Verbesserung nicht nachweisen konnte.

Die Trennschärfe des Tests hängt von der wahren Erfolgswahrscheinlichkeit des neuen Medikaments ab. Wenn $p_{neu} = 0.8$ ist (grosser Effekt), dann ist die Trennschärfe $1 - \beta = \mathrm{P}[k \geq 13] = 0.97$, und nichtsignifikante Ergebnisse sind selten ($\beta = 3\%$). Wenn das neue Medikament nur unwesentlich besser ist als das alte ($p_{neu} = 0.5$), ist $1 - \beta = \mathrm{P}[k \geq 13] = 0.13$. Die Stichprobe ist also viel zu klein, um derartige Effekte nachweisen zu können ($\beta = 87\%$).

Beispiel für die Konstruktion eines Tests

Man möchte prüfen, ob herzkranke Babies später zu laufen beginnen als gesunde Babies. Dafür wird eine empirische Studie mit $n = 20$ herzkranken Kindern durchgeführt. Deren Werte sollen mit Normwerten aus der Literatur verglichen werden, die man als nicht zufällig annimmt. Für den Beginn des Laufens liefert die Normpopulation ein Mittel von $\mu_0 = 12$ Monaten mit einer Populationsstreuung von $\sigma_0 = 1.8$ Monaten. Die Studie mit den herzkranken Babies ergibt ein Mittel von $\bar{x} = 12.8$ Monaten (das zugehörige, unbekannte Populationsmittel sei μ). Der Einfachheit halber nehmen wir an, dass die Standardabweichung der Norm $\sigma_0 = 1.8$ auch gültig ist.

- Wissenschaftliche Hypothese: Kinder mit angeborenem Herzleiden laufen später.
  $H_1 : \mu > \mu_0$
- Statistische (Null–) Hypothese:
  $H_0 : \mu = \mu_0$

Die statistische Prüfung will nun quantitativ nachweisen, dass die Grösse der Abweichung $(\bar{x} - \mu_0)$ nicht durch den Zufall erklärt werden kann:

- Berechne Differenz $(\bar{x} - \mu_0)$.
- Die Grösse von $(\bar{x} - \mu_0)$ wird auf den Standardfehler $\sigma_0/\sqrt{n}$ der Differenz bezogen. Dies führt zur Teststatistik (Testgrösse) z:

$$z = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{0.8}{1.8/\sqrt{20}} = 1.99$$

Die Wahrscheinlichkeit, einen so grossen oder noch grösseren Wert per Zufall zu erhalten — obwohl die Null–Hypothese gilt — bezeichnen wir mit p („p–Wert“).
Um die Wahrscheinlichkeit p berechnen zu können, müssen wir eine Annahme über die Verteilung der Daten machen. Der Einfachheit halber nehmen wir Normalverteilung an. In der Praxis muss das geprüft werden.

Die folgende Tabelle zeigt die Resultate nicht nur für $n = 20$, sondern auch für $n = 10$, 40, 80 bei identischem $\bar{x}$.

n	10	20	40	80
z	1.41	1.99	2.81	3.98
p	0.079	0.023	0.0025	0.0003

Falls $p \leq \alpha$ gilt, ist die Differenz statistisch signifikant zum Signifikanzniveau α. Demnach gilt: Bei $\alpha = 0.05$ ist das Resultat signifikant für $n \geq 20$, bei $\alpha = 0.01$ ist das Resultat signifikant für $n \geq 40$. Man sieht, dass bei grösserem Stichprobenumfang n die gleiche Differenz eher signifikant wird. Mathematisch gesprochen nimmt die Trennschärfe mit $\sqrt{n}$ zu, ähnlich wie sich die Variabilität von Schätzern verbessert (siehe Abschnitt 3.7). Bei gleichem n nimmt die Trennschärfe mit $(\mu - \mu_0)$ und mit $1/\sigma_0$ zu.

Da es in der Praxis oft um den Nachweis von Mittelwertsunterschieden geht, haben viele Testgrössen eine Form, die ähnlich zu der von z ist.

Allgemeines Prozedere für einen statistischen Test

- Formuliere wissenschaftliche Hypothese H_1 und Nullhypothese H_0 (bezogen auf Populationswerte).
- Setze Irrtumswahrscheinlichkeit α fest.
- Es werden Daten $x_1, \ldots, x_n$ gesammelt.
- Definiere Test–(Prüf–)Statistik $T(x_1, \ldots, x_n)$.
  Erwünscht ist:
  - T soll empfindlich auf H_1 reagieren.
  - Die Verteilung von T (für H_0) soll mathematisch berechenbar sein.

  Typische Form für T:
  - bei einseitiger Alternative:

$$T = \frac{\text{beobachteter Wert} - \text{hypothetischer Wert}}{\text{Standardfehler des beobachteten Wertes}}$$

  - bei zweiseitiger Alternative:

$$T = \left| \frac{\text{beobachteter Wert} - \text{hypothetischer Wert}}{\text{Standardfehler des beobachteten Wertes}} \right|$$

- Berechne Teststatistik T für gegebene Daten $x_1, \ldots, x_n \longrightarrow T_0$.
- Die Verteilung $F_T(x)$ zu T sei unter der Annahme der Hypothese H_0 bekannt.
- Ermittle p–Wert zu beobachtetem T_0:
  = Wahrscheinlichkeit, per Zufall (bei der Nullhypothese) einen so extremen Wert wie T_0 zu erhalten:

$$p = 1 - F_T(T_0)$$

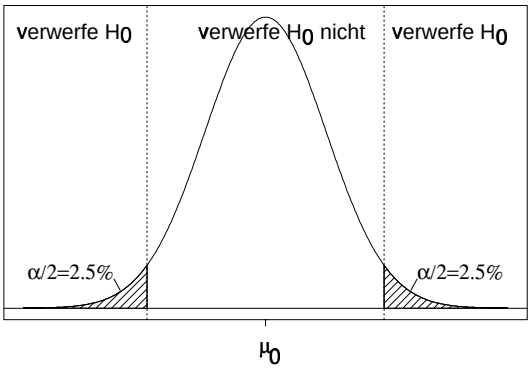
- Entscheide: Falls $p \leq \alpha \Longrightarrow$ verwerfe H_0.
  Falls $p > \alpha \Longrightarrow$ verwerfe H_0 nicht.

Für die verschiedenen wissenschaftlichen Fragestellungen gibt es eine Vielzahl von Tests. In diesem Kapitel behandeln wir einige von grosser praktischer Bedeutung; weitere Beispiele werden in den folgenden Kapiteln bezogen auf die dort behandelten Probleme vorgestellt.

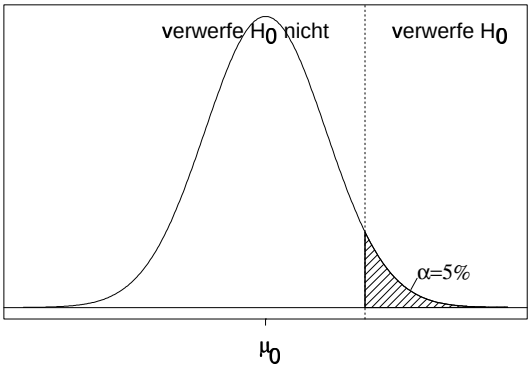
Nachfolgend werden graphisch die Bereiche dargestellt, in denen die Nullhypothese verworfen bzw. nicht verworfen wird, sowie die Wahrscheinlichkeiten α ($= 0.05$) und $1 - \beta$. Dies erfolgt für den Nachweis eines Mittelwertsunterschiedes von einem vorgegebenen Wert, d. h. $H_0 : \mu = \mu_0$. Für die Daten wird eine Normalverteilung bekannter Varianz angenommen.

Falls die Nullhypothese stimmt, d. h. $\mu = \mu_0$, stellen sich die Entscheidungsbereiche wie folgt dar:

a) Alternative zweiseitig ($\mu \neq \mu_0$):



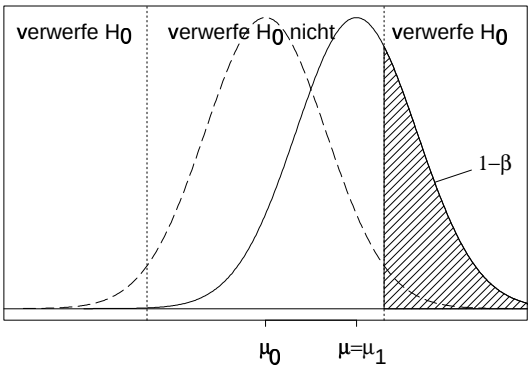
b) Alternative einseitig ($\mu > \mu_0$ oder je nach Fragestellung auch $\mu < \mu_0$):



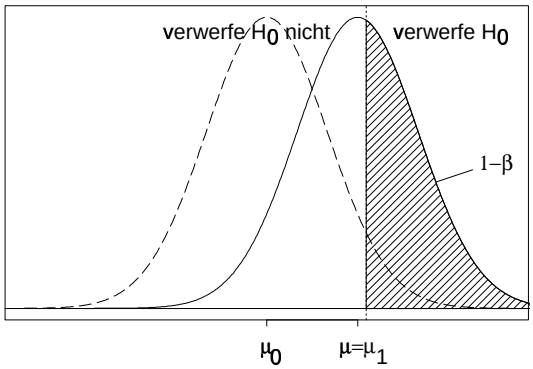
Die schraffierten Flächen geben die Wahrscheinlichkeiten an, dass die Nullhypothese fälschlicherweise verworfen wird.

Falls die Nullhypothese nicht stimmt und $\mu = \mu_1 > \mu_0$ gilt, führt dies zu folgenden Bereichen:

a) Alternative zweiseitig ($\mu \neq \mu_0$):



b) Alternative einseitig (in der Graphik $\mu > \mu_0$):



Man sieht, dass die Trennschärfe $(1 - \beta)$ bei gleicher Differenz $(\mu - \mu_0)$ grösser wird, wenn man einseitig testet (die Verwerfungsgrenze „rutscht hinunter“). Da man manchmal so bei einseitigem Testen Signifikanz erhält, wenn dies zweiseitig nicht möglich ist, und da Signifikanz die Publizierbarkeit verbessert, sind viele Zeitschriften einseitigen Tests gegenüber skeptisch (Gefahr, post–hoc eine Richtung zu postulieren).

Macht (Trennschärfe) eines Tests

- Optimale Tests sind so definiert, dass sie bei vorgegebenem α maximale Trennschärfe haben (Beispiel: t–Tests, falls Normalverteilung vorliegt).
- Die Macht sinkt, wenn α kleiner wird („Unschärferelation“: wenn man den einen Fehler kleiner macht, wird der andere grösser).
- Die Macht steigt, wenn die Variabilität kleiner wird. Dies bedeutet, dass homogenere Gruppen oder bessere Messmethodik von Vorteil sind.
- Die Macht ist bei einseitigen Tests besser.
- Die Stichprobengrösse n kann mit dem Versuchsplan so gewählt werden, dass zum Beispiel $\beta = 0.10$ oder 0.05 erreicht wird (d. h. vorgegebene Macht 90% oder 95%). Damit ist ein klarer Entscheid bezüglich der Nullhypothese und der Alternative möglich.

Wir werden die Bestimmung der Stichprobengrösse jetzt näher erläutern („Power Analyse“). Wem das zu speziell ist, kann direkt bei Abschnitt 4.2 weiterlesen.

Bestimmung der Stichprobengrösse

Die Bestimmung der notwendigen Stichprobengrösse soll am Beispiel des Nachweises eines Mittelwertsunterschiedes von einem vorgegebenen Wert μ_0 bei bekannter Varianz σ_0^2 und unabhängigen normalverteilten Daten $x_1, \ldots, x_n$ erläutert werden. Die Teststatistik ist

$$z = \sqrt{n}\,\frac{\bar{x} - \mu_0}{\sigma_0}\,.$$

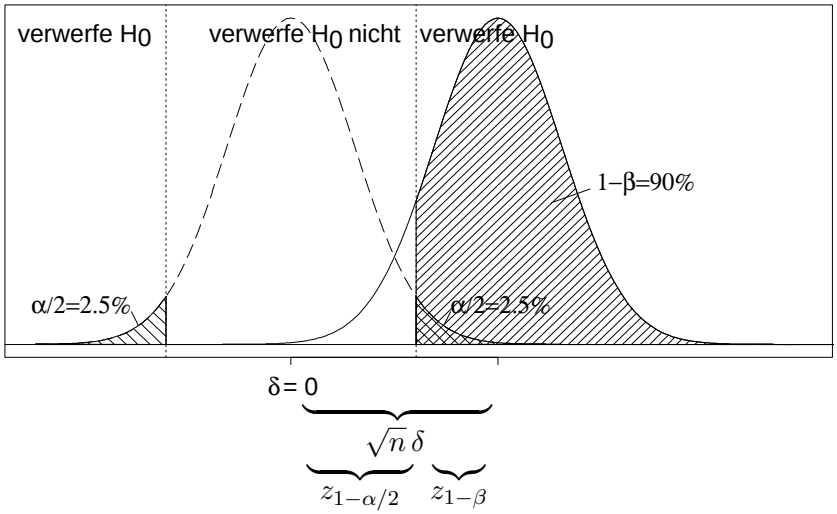
Unter der Nullhypothese $\mathrm{H}_0 : \mu = \mu_0$ ist z standardnormalverteilt. Zur Erinnerung: Mit z_α bezeichnen wir das $\alpha\times$ 100 %–Perzentil der Standardnormalverteilung. H_0 wird verworfen, wenn $|z| > z_{1-\alpha/2}$, da

$$\mathrm{P}_o[|z| > z_{1-\alpha/2}] = \alpha$$

(siehe auch die Graphik zum Fehler erster Art bei zweiseitiger Alternative, oben). Wenn $\mu = \mu_1 > \mu_0$ gilt, dann ist

$$z = \sqrt{n}\,\frac{\bar{x} - \mu_1}{\sigma_0} + \sqrt{n}\,\frac{\mu_1 - \mu_0}{\sigma_0} \sim \mathcal{N}(\sqrt{n}\delta, 1)\,.$$

Die Grösse $\delta = \frac{\mu_1 - \mu_0}{\sigma_0}$ ist hier die **Effektgrösse**.



Die Trennschärfe $1-\beta$ erhält man aus der obigen Graphik. Für eine vorgegebene Effektgrösse δ und eine vorgegebene Trennschärfe $1-\beta$ berechnet sich die notwendige Stichprobengrösse damit aus

$$\sqrt{n}\delta = z_{1-\alpha/2} + z_{1-\beta}$$

und folglich

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\delta^2}\,.$$

Die notwendige Stichprobengrösse n ist proportional zur Varianz σ_0^2 und indirekt proportional zum quadrierten Mittelwertsabstand $(\mu_1 - \mu_0)^2$. Das ist analog zur Genauigkeit von

Schätzungen: Um einen halb so grossen Effekt nachweisen zu können, brauchen wir 4 mal so viele Beobachtungen. Die Stichprobengrösse n wächst ausserdem bei einer Reduktion des Signifikanzniveaus α und bei einer Erhöhung der Trennschärfe $1-\beta$. Diese Zusammenhänge sind aber nichtlinear und deshalb nicht sofort zu überblicken.

4.2 Tests auf Mittelwertsunterschiede

Wenn man eine experimentelle Bedingung ändert oder eine neue Therapie erprobt, wird man sich zuerst für Veränderungen im Mittelwert interessieren. Beim Vergleich von 2 Gruppen gibt es im wesentlichen 3 Situationen:

- Vergleich eines Mittelwertes mit einem bekannten festen Wert (Einstichprobenproblem),
- Vergleich der Mittelwerte zweier unabhängiger Stichproben (Zweistichprobenproblem),
- Vergleich der Mittelwerte zweier verbundener Stichproben (gepaartes Testproblem).

Wenn wir eine Normalverteilung der Beobachtungen voraussetzen, werden die Hypothesen mit t–Tests geprüft, sonst mit entsprechenden Rangverfahren. Im folgenden wollen wir die verschiedenen Testverfahren vorstellen.
Dabei zeigen wir auch, wie die entsprechenden Testgrössen konstruiert werden. Dadurch soll der methodisch interessierte Leser ein Gefühl dafür bekommen, wie das soeben beschriebene allgemeine Procedere bei diesen Problemen funktioniert. Für die praktische Anwendung ist die Kenntnis der Konstruktion der Testgrössen nicht notwendig, da die Verfahren in allen gängigen Statistikprogrammen verfügbar sind.

4.2.1 Einstichproben–t–Test

Beim einführenden Beispiel des Laufens von Babies benutzten wir einen Einstichprobentest, da wir mit einer bekannten Norm — als wahr vorausgesetzt — und nicht mit einer Gruppe von gesunden Kindern verglichen haben. Letzteres führt auf einen Zweistichproben–Test (siehe unten). Da wir normalverteilte Daten und eine bekannte Standardabweichung angenommen haben, resultiert eine normalverteilte Teststatistik z.

Beispiel: Beginn des Laufens von Babies $\Longrightarrow$ Einstichproben–z–Test
$x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma_0^2)$, wo σ_0^2 bekannt ist
$\mathrm{H}_0 : \mu = \mu_0$

$$\Longrightarrow z = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \quad \text{unter der Nullhypothese normalverteilt } \mathcal{N}(0,1)$$

Falls aber — was im allgemeinen realistisch ist — die Varianz σ^2 nicht bekannt ist, muss sie durch s^2 aus den Daten geschätzt werden. Da man somit durch eine Zufallsvariable teilt, ist die entsprechende Statistik t nicht mehr normalverteilt.

Dies führt auf den t–Test:

Teststatistik	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	Einstichproben–t–Test

Die Teststatistik t ist t–verteilt mit $(n-1)$ Freiheitsgraden (siehe Abschnitt 3.4).
Unter der Annahme, dass $s = 1.8$ gilt, erhalten wir beim vorangehenden Beispiel:

n	10	20	40	80
t	1.41	1.99	2.81	3.98
p	0.0961	0.0306	0.0039	0.0008

Die einseitigen p–Werte sind im Vergleich zur z–Teststatistik (Seite 49 mit σ_0 anstatt s) etwas grösser geworden.

4.2.2 Zweistichproben–t–Test für unabhängige Stichproben

Man möchte die Mittelwerte zweier Gruppen auf der Grundlage von zwei Stichproben statistisch vergleichen. In der Praxis ist diese Aufgabenstellung viel häufiger als das Einstichproben–Problem.

Beispiel: Vergleich der log–T_4–Zellanzahl für Hodgkin– und non–Hodgkin–Patienten

Gruppe 1 (Hodgkin): $n = 20$, $\bar{x} = 6.49$, $s_x = 0.71$

Gruppe 2 (non–Hodgkin): $m = 20$, $\bar{y} = 6.09$, $s_y = 0.63$

Die wissenschaftliche Hypothese ist: Die T_4–Zellanzahl bei Hodgkin–Patienten ist auch nach der Remission erhöht gegenüber non–Hodgkin–Patienten.

$$\mathrm{H}_1 : \mu_x > \mu_y \text{ (einseitige Alternative)} \qquad \mathrm{H}_0 : \mu_x = \mu_y \Longrightarrow \mu_x - \mu_y = 0$$

Die Konstruktion einer Zweistichproben–Teststatistik erfolgt nach dem bewährten Prinzip:

$$\text{Beobachtet} - \text{Erwartet bei } \mathrm{H}_0 = (\bar{x} - \bar{y}) - 0 = 0.4$$

Liegt 0.4 genügend weit von 0 weg, so dass die Nullhypothese verworfen werden kann? Dividiere durch den Standardfehler von $(\bar{x} - \bar{y})$!

Machen wir die Annahme, dass $\sigma_x^2 = \sigma_y^2 = \sigma^2$ gilt, d. h. gleiche Varianz in beiden Stichproben. Dann erhalten wir für die Standardabweichung von $(\bar{x} - \bar{y})$:

$$\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}$$

Die gemeinsame Standardabweichung σ ist nicht bekannt und muss durch s geschätzt werden.

Die Formel für die aus beiden Stichproben kombinierte (gepoolte) Standardabweichung s lautet:

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

Damit ergibt sich als Testgrösse:

Teststatistik	$t = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$	Zweistichproben–t–Test

Annahme: Seien $x_1, \ldots, x_n, y_1, \ldots, y_m$ unabhängige, **normalverteilte** Grössen mit gleicher Varianz.
Dann gilt: Die Teststatistik t hat eine t–Verteilung mit $n + m - 2$ Freiheitsgraden.
Bei einer Datenanalyse sollten die Annahmen wenigstens grob geprüft werden, z. B. mittels Box–Plots.

Beispiel: Wir kehren zurück zum Vergleich von log–T_4–Zellanzahl für $n = 20$ Hodgkin– und $m = 20$ non–Hodgkin–Patienten.

Nach der obigen Formel erhalten wir ein $s = 0.67$ und damit als Testgrösse:

$$t = \frac{0.4}{0.67\sqrt{\frac{1}{20} + \frac{1}{20}}} = 1.88$$

$\mathrm{P}[t \geq 1.88]$ = einseitiges $p = 0.034$,
also ist $p \leq \alpha = 0.05 \Longrightarrow$ Hodgkin–Patienten haben signifikant mehr T_4–Zellen.
Wie bereits erwähnt, haben einseitige Tests eine höhere Trennschärfe als zweiseitige, was zu Misbrauch führen kann. Der zweiseitige Test hat den p–Wert $p = \mathrm{P}[t \leq -1.88$ oder $t \geq 1.88] = 0.068 \Longrightarrow p > \alpha = 0.05$
Es ergibt sich also kein signifikanter Unterschied zwischen den beiden Gruppen.

Wenn man mehr als zwei Gruppen miteinander vergleichen will, sollte eine Varianzanalyse als Verallgemeinerung des Zweistichproben–t–Tests durchgeführt werden (siehe Abschnitt 6.1).

4.2.3 Der gepaarte t–Test

(t–Test für „gepaarte“ oder „verbundene“ Stichproben)

Es ist ein relativ häufiger Fehler, dass der eben besprochene t–Test auch für den Vergleich von Daten benutzt wird, die am selben Menschen gewonnen werden, also nicht unabhängig sind. Dafür ist der gepaarte t–Test konzipiert.

Beispiele:

- prä–post–Vergleiche bei Therapiestudien
- Mehrfachuntersuchungen an denselben Patienten
- Vergleich EEG linke und rechte Hemisphäre

Beispiel: Es werden kardiologische Funktionen bei Typ I Diabetikern untersucht. Es ist bekannt, dass Diabetiker schlechtere kardiovaskuläre Werte im Vergleich zu Gesunden aufweisen. Die Frage, ob sich die Funktionen bei guter Glukose–Einstellung verbessern, wurde bei $n = 8$ Patienten untersucht. Hier wird die Herzrate analysiert.

$x_1, \ldots, x_n$ — Daten zum Zeitpunkt 1, bei schlechter Einstellung
$y_1, \ldots, y_n$ — Daten zum Zeitpunkt 2, bei guter Einstellung

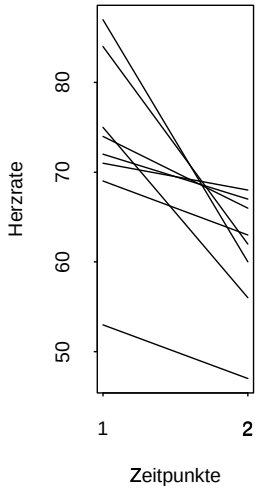
Man erwartet eine Erniedrigung der Herzrate (Alternativ–Hypothese H_1), was zu folgendem Hypothesenpaar führt:

$$H_0 : \mu_x = \mu_y$$
$$H_1 : \mu_x > \mu_y$$

Es wurden dazu die individuellen Veränderungen (post − prä Werte) $d_i = y_i - x_i$ in der Herzrate analysiert. Die entsprechende Formulierung der Hypothese ist dann mit $\delta = \mu_y - \mu_x$ wie folgt:

$$H_0 : \delta = 0$$
$$H_1 : \delta < 0$$

Nr	x	y	d
1	74	66	-8
2	72	67	-5
3	84	62	-22
4	53	47	-6
5	75	56	-19
6	87	60	-27
7	69	63	-6
8	71	68	-3
Mittelwert	73	61	-12
s	10	7	9.2



Wenn man bei den d_i angelangt ist, reduziert sich der gepaarte t–Test formal auf den Einstichproben–t–Test, d. h. folgende Teststatistik:

Teststatistik	$t = \dfrac{\bar{d}}{s_d/\sqrt{n}}$	Gepaarter t–Test

Unter der Annahme, dass die d_i normalverteilt sind, folgt die Teststatistik t unter der Nullhypothese einer t–Verteilung mit $(n-1)$ Freiheitsgraden.

$$t = \frac{-12}{9.2\,/\,\sqrt{8}} = -3.7$$

Die Wahrscheinlichkeit $(t \leq -3.7)$ zu erhalten ist 0.004. Man erhält demnach eine signifikante Verbesserung bei guter therapeutischer „Compliance“. Der Effekt ist deutlich und deshalb bereits für $n = 8$ Patienten nachweisbar.

Falsch wäre die Anwendung des Zweistichproben–t–Tests.

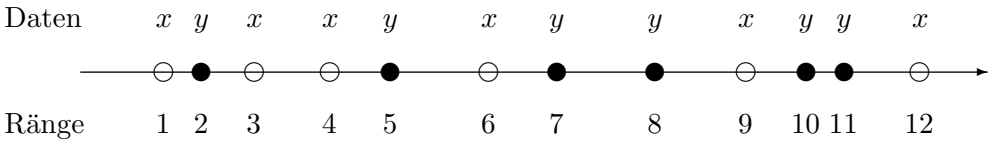
4.2.4 Rangtests: Mann–Whitney– und Wilcoxon–Test

In der Praxis ist bei der Anwendung der t–Tests oft problematisch, dass normalverteilte Daten vorausgesetzt werden, bzw. dass andernfalls eine Transformation zur Normalverteilung gesucht werden muss (z. B. logarithmieren der Anzahl T_4–Zellen). Wir wollen hier Tests vorstellen, die eine gute Trennschärfe haben, aber die Normalverteilungsannahme nicht erfordern. Die Idee ist dabei, nur die Rangordnung der Daten und nicht die Daten selbst zu benützen. Es besteht eine gewisse Ähnlichkeit zur Benutzung von Median und Quantilen als Kennwerte; diese beruhen auch auf der Rangordnung. Daraus ergibt sich auch die Unempfindlichkeit gegen Ausreisser und extreme Werte: Wie gross ein Wert numerisch auch sei, in der Rangordnung einer Stichprobe der Grösse n erhält er höchstens den Rang n.

Der Test zum Vergleich der Zentren von zwei unabhängigen Gruppen und damit das Analogon zum Zweistichproben–t–Test heisst **Mann–Whitney Test**. Er wird auch Wilcoxon–Test für unabhängige Stichproben oder Wilcoxon–Rangsummen–Test genannt. Das Analogon zum gepaarten t–Test ist der **Wilcoxon–Test für Paardifferenzen** (Wilcoxon signed rank test).
Wie bei den t–Tests setzt man voraus, dass die Beobachtungen bzw. Paare unabhängige Zufallsgrössen sind. Die Voraussetzung der Normalverteilung wird durch die schwächere Annahme einer stetigen Verteilung ersetzt.
Die Trennschärfe dieser Rangtests ist für ziemlich beliebige Situationen gut, auch dann, wenn die entsprechenden t–Tests nicht gültig sind. Die Rangtests haben aber auch bei Vorliegen einer Normalverteilung eine Trennschärfe nahe der der t–Tests (Effizienz 96%). Im Zweifelsfall sollte man also immer Rangtests anstelle von t–Tests verwenden.

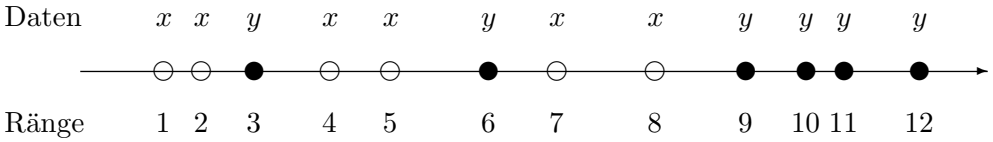
Zur Illustration des Vorgehens beim Mann–Whitney Test vergleichen wir zwei unabhängige Stichproben $x_1, \ldots, x_6$ und $y_1, \ldots, y_6$, indem wir sie in eine **gemeinsame** Rangordnung bringen.

Situation 1: $\mu_x \approx \mu_y$



Der mittlere Rang der x_i–Werte ist 5.8, derjenige der y_i ist 7.2, so dass sie annähernd gleich sind.

Situation 2: $\mu_y > \mu_x$



Der mittlere Rang der y_i ist mit 8.5 deutlich grösser als derjenige der x_i mit 4.5.

Das Vorgehen ist nun wie folgt:

1. Erstelle **gemeinsame** Rangordnung.
2. Berechne getrennt mittlere Ränge der x_i und der y_i.
3. Die mittleren Ränge dienen dem Programm dazu, p–Werte zu berechnen.

Beispiel: Wir wollen wieder prüfen, ob die T_4–Zellanzahl bei Hodgkin–Patienten erhöht ist, verglichen mit non–Hodgkin–Patienten.

Die T_4–Zellanzahl ist nicht normalverteilt, sondern rechtsschief, weshalb wir früher logarithmiert haben, um den t–Test anwenden zu können.

Hier ein Ausschnitt der geordneten Werte:

Gruppe	nH	nH	H	nH	nH	H	...
T_4–Zellen	116	151	171	192	208	257	...
Rang	1	2	3	4	5	6	...

Es fällt auf, dass die non–Hodgkin–Patienten bei kleinen Werten stark vertreten sind. SPSS liefert einen p–Wert von $p = 0.079$ bei zweiseitigem Testen, und damit keine signifikante Erhöhung der T_4–Zellanzahl bei remittierten Hodgkin–Patienten. Man erhält somit ein vergleichbares Ergebnis wie wenn man logarithmiert und den t–Test ($p = 0.068$) anwendet.

4.3 Tests für Proportionen oder Wahrscheinlichkeiten

4.3.1 Einstichprobenfall

Man möchte prüfen, ob eine berechnete Proportion oder relative Häufigkeit von einem vorgegebenen festen Wert signifikant abweicht.

$H_0 : p = p_0$, $\quad H_1 : p > p_0$ respektive $p < p_0$ (einseitig)
$p \neq p_0$ (zweiseitig)

Der Test erfolgt mittels Binomialverteilung. Beispiele von früher sind:

1. Das Standardmedikament heilt 40% der Patienten ($p_0 = 0.4$). Ist das neue Medikament besser, d. h. $H_1 : p_{neu} > p_0$?

2. Überwiegen Knabengeburten tatsächlich gegenüber Mädchengeburten?
   $H_0 : p = 0.5 = p_0$, $H_1 : p \neq 0.5 = p_0$
   Die Antwort ist „ja“, siehe Konfidenzintervalle im Abschnitt 4.6.

4.3.2 Zweistichprobenfall

Statistisch verglichen werden nun zwei empirische Proportionen oder Häufigkeiten.

Beispiel: In 34 von 113 Knaben und in 54 von 139 getesteten Mädchen erfolgt der Nachweis eines Grippevirus–Antikörpers. Gibt es einen Geschlechtsunterschied in der Häufigkeit?

Man kann — ähnlich zum Zweistichproben–t–Test — die Differenz der beiden Häufigkeiten geeignet normieren und die Testgrösse mit den Perzentilen der approximativ gültigen Standardnormalverteilung vergleichen. Für das Beispiel führt das auf ein zweiseitiges $p = 0.14$, d. h. der Geschlechtsunterschied ist nicht signifikant. Natürlicher ist es aber, die Resultate als Vierfelder–Tafel darzustellen und mit einem χ^2–Homogenitätstest statistisch zu prüfen (siehe Abschnitt 4.4.2).

Achtung: Auch hier muss man bei gepaarten Studien — z. B. wenn die Häufigkeit von Schmerzen vor und nach einer Behandlung bei denselben Patienten untersucht wird — anders vorgehen (McNemar–Test).

4.4 Der χ^2–Test

Der χ^2–Test — oder besser die χ^2–Statistik — eignet sich zur Beantwortung qualitativ unterschiedlicher Fragen. Gemeinsam ist ihnen, dass wir es mit kategoriellen Daten zu tun haben.

4.4.1 χ^2–Anpassungstest

Es geht hier darum, die empirische Verteilung kategorieller Daten mit einer vorgegebenen Verteilung statistisch zu vergleichen.

Beispiel: Die Genotypen A, B und C kommen nach einem Vererbungsmodell in den Häufigkeiten 1/4, 1/2, 1/4 vor. Hier ist die Verteilung durch ein Modell vorgegeben.

Um das Modell zu überprüfen, werden 100 Pflanzen gezüchtet, mit folgendem Ergebnis.

3 Zellen:

A	B	C
18	55	27

Die Übereinstimmung der Daten mit dem Modell soll nun statistisch geprüft werden.

$H_0 : p_A = 1/4, \quad p_B = 1/2, \quad p_C = 1/4$

Unter der Nullhypothese erwartet man Zellhäufigkeiten 25, 50, 25.

Diese Fragestellung entspricht für kategorielle Daten in etwa derjenigen eines Einstichprobentests bei kontinuierlichen Daten.

Idee: Vergleiche beobachtete („Obs“ für „observed“) und erwartete („Exp“ für „expected“) Zellhäufigkeiten:

$$(18-25)^2, \quad (55-50)^2, \quad (27-25)^2$$

Der Standardfehler dieser quadratischen Abweichungen ist jeweils die erwartete Zellhäufigkeit. Dies folgt aus einem mathematischen Argument mit der Poisson–Verteilung, dessen Details hier zu weit führen würden. Damit erhalten wir folgende Teststatistik X^2:

$$X^2 = \frac{(18-25)^2}{25} + \frac{(55-50)^2}{50} + \frac{(27-25)^2}{25} = 2.62$$

Man kann ähnlich das Problem mit k Zellen betrachten. Die Teststatistik sieht dann so aus:

Teststatistik	$X^2 = \sum_{\text{Zellen}} \frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$	χ^2–Anpassungstest

Die Prüfverteilung ist eine χ^2–Verteilung mit $(k-1)$ Freiheitsgraden (siehe Abschnitt 3.4). Der p–Wert ist aber nur approximativ gültig (für grosse n wird er exakt).

Beispiel: $X^2 = 2.62 \Longrightarrow \chi^2$ verteilt mit 2 Freiheitsgraden
oberes 5% Perzentil von $\chi^2_2 = 5.99$; $2.62 < 5.99 \Longrightarrow$ nicht signifikant

Der Versuch spricht also nicht gegen die Verteilung, die aus dem genetischen Modell abgeleitet wurde.

Beachten Sie, dass es beim χ^2–Test keine einseitige Alternative gibt, da die Vorzeichen der Abweichungen durch das Quadrieren verschwinden.

4.4.2 Testen auf Homogenität in Kontingenztafeln

Im Unterschied zu Abschnitt 4.4.1 vergleicht man nicht empirische Häufigkeiten mit theoretischen, sondern empirische Häufigkeiten von 2 oder mehr unabhängigen Gruppen miteinander.

Beispiel: Vergleich von Medikament A mit Medikament B an $n = 150$ Patienten. Die klinische Beurteilung des Gesundheitszustandes ist dreistufig als sehr gut, gut bzw. schlecht gegeben.

Nach Randomisierung erhalten 80 Patienten Medikament A und 70 Patienten Medikament B. Die Daten werden dann in einer Kontingenztafel (Kreuztabelle) angeordnet:

	sehr gut	gut	schlecht	n
A	37 (A_1)	24 (A_2)	19 (A_3)	80
B	17 (B_1)	33 (B_2)	20 (B_3)	70
Total	54	57	39	150

H_0: A und B sind gleich gut, d. h.

$$p_{A_1} = p_{B_1} = p_1, \quad p_{A_2} = p_{B_2} = p_2, \quad p_{A_3} = p_{B_3} = p_3$$

Das Problem ist ähnlich zu einem Zweistichproben–Problem bei kontinuierlichen Daten.

Das Testprinzip besteht wieder darin, in jeder Zelle die beobachtete (Obs) und erwartete (Exp) Anzahl zu vergleichen:

Teststatistik	$X^2 = \sum_{\text{Zellen}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$	χ^2–Test auf Homogenität

Die Prüfverteilung ist eine χ^2–Verteilung mit $(r-1) \times (c-1)$ Freiheitsgraden.
r = Anzahl Zeilen in der Kreuztabelle (r für „rows“)
c = Anzahl Kolonnen in der Kreuztabelle (c für „columns“)

Im Medikamenten Beispiel ist $r = 2$ (Medikament A, B) und $c = 3$ (Rating: sehr gut, gut, schlecht).

	sehr gut	gut	schlecht	n
A	37 (28.8)	24 (30.4)	19 (20.8)	80
B	17 (25.2)	33 (26.6)	20 (18.2)	70
Total	54	57	39	150

In Klammern stehen die erwarteten Häufigkeiten, falls beide Medikamente gleich gut sind. Sie werden berechnet, indem die Ergebnisse aus der gesamten Stichprobe (total) im Verhältnis 80:70 aufgeteilt werden.

Man erhält als Teststatistik $X^2 = 8.22$, und damit $p = \mathrm{P}[\chi^2_2 \geq 8.22] = 0.016 < 0.05$. Demnach sind die Medikamente A und B signifikant verschieden mit Irrtumswahrscheinlichkeit $\alpha = 0.05$.

Man sollte beachten, dass im Falle eines gepaarten Versuchsplans (Medikament A und B an derselben Stichprobe erprobt) eine andere Statistik anzuwenden ist (McNemar–Test).

Der p–Wert des χ^2–Tests ist nur für grosse n gültig. Allgemein sagt man, dass die Approximation für 2×2 Kontingenztafeln gut ist, wenn die erwarteten Häufigkeiten (Exp) in allen Zellen ≥ 5 sind. Ist diese Voraussetzung nicht gegeben, kann man **Fisher's exakten Test** anwenden. Für allgemeine $r \times c$ Tafeln genügt es für die Gültigkeit des χ^2–Tests, dass alle erwarteten Häufigkeiten ≥ 3 sind. Fisher's exakter Test ist für allgemeine $r \times c$ Tafeln nur in wenigen Statistikpaketen (z. B. in StatXact, SPSS ab Version 11 und in SAS) implementiert.

4.4.3 Test für Unabhängigkeit zweier Variablen

Im nächsten Kapitel (Kapitel 5) wird der Test auf Unabhängigkeit für kontinuierliche Daten (mittels Korrelationen) behandelt. Hier wollen wir dasselbe für diskrete Variable untersuchen.

Problemstellung: An einer Stichprobe der Grösse n werden 2 diskrete Variable erhoben und in einer Kontingenztafel angeordnet. Sind die Merkmale unabhängig?

H_0: $p_{ij} = p_i \times p_j$ für alle i, j

Sie erinnern sich: Unabhängigkeit ist so definiert, dass Wahrscheinlichkeiten multipliziert werden können.

Beispiel: An 400 Kindern wird deren Händigkeit geprüft, und es wird geprüft, ob Vater und Mutter links– oder rechtshändig sind. Frage: Vererbt sich die Händigkeit?

	Kind		
Vater × Mutter	rechts	links	total
rechts, rechts	303 (295.8)	37 (44.2)	340
rechts, links	29 (33.1)	9 (4.9)	38
links, links	16 (19.1)	6 (2.9)	22
total	348	52	400

( ) = erwartet, falls unabhängig

H_0: Es besteht kein Zusammenhang („Händigkeit ist nicht genetisch bedingt“)

Teststatistik: Obwohl wir eine andere Problemstellung haben, geht alles gleich wie beim Test auf Homogenität. Man vergleicht beobachtete und erwartete Zellhäufigkeiten und bildet die Summe der normierten Quadrate ($= X^2$). Diese ist dann wieder approximativ $\chi^2_{(r-1)\times(c-1)}$–verteilt.

Im Beispiel ist $X^2 = 9.15$, $p = \mathrm{P}[\chi^2_2 \geq 9.15] = 0.010 < \alpha = 0.05$. Wir gehen daher von einer Vererblichkeit der Händigkeit aus.

Da die erwartete Häufigkeit in 2 Zellen kleiner als 5 ist, sollte besser Fisher's exakter Test benutzt werden. SPSS liefert ebenfalls einen p–Wert von 0.010 für Fisher's exakten Test, das Ergebnis wird in diesem Beispiel also bestätigt.

4.5 Multiples Testen

Bisher sind wir davon ausgegangen, dass wir einen Test durchführen und dafür die Irrtumswahrscheinlichkeit α einhalten wollen. Oft werden in der Praxis mehrere Tests durchgeführt. Dabei werden aber die Irrtumswahrscheinlichkeiten nicht mehr eingehalten, wie das folgende Beispiel zeigt:

Wenn ich 20 mal unabhängig voneinander etwas teste, wird im Mittel bei $\alpha = 5\%$ ein Resultat signifikant, auch wenn in allen 20 Fällen die Nullhypothese richtig ist.

Beispiel: In einer Studie mit 4 diagnostischen Gruppen werden 20 Variable erhoben.
Naiv vergleicht man die 4 Gruppen paarweise untereinander, eine Variable nach der anderen.
$\Longrightarrow$ 120 $(= 6 \times 20)$ paarweise Vergleiche möglich
$\Longrightarrow$ 120 statistische Tests möglich

H_0 : Kein einziger Unterschied zwischen den Gruppen
H_1 : Unterschied in mindestens einer Variablen

Sei $\alpha = 0.05$ die Irrtumswahrscheinlichkeit für jeden einzelnen Test der paarweisen Vergleiche. Falls H_0 gilt, erhalten wir trotzdem im Mittel per Zufall $0.05 \times 120 = 6$ Ablehnungen von H_0. Das bedeutet, dass wegen der vielen Tests die Irrtumswahrscheinlichkeit α nicht für das multiple Testproblem gilt.

Allgemein: Wir führen k unabhängige Tests auf nominellem 5% Niveau durch. Wie gross ist dann die Wahrscheinlichkeit, bei Gültigkeit von H_0 mindestens einen p–Wert $p < \alpha$ zu erhalten („effektives α“)?

k	nominelles α	effektives α
1	0.05	0.05
2	0.05	0.10
3	0.05	0.14
5	0.05	0.23
10	0.05	0.40
20	0.05	0.64
50	0.05	0.92

Das viele Testen führt also auf eine „α–Inflation“.

Lösungen:

1. Multivariate statistische Verfahren wie Varianzanalyse (siehe Abschnitt 6.1)

2. Bonferroni–Korrektur (für kleine k!)

$$\Longrightarrow \quad \boxed{\alpha[\text{Einzeltest}] = \frac{\alpha}{k}}$$

Die Bonferroni–Korrektur ist konservativ, d. h. der Fehler 1. Art ist deutlich kleiner als α, womit die Trennschärfe sinkt.

3. In der Versuchsplanung werden wenige strikte Hypothesen zum Testen aufgestellt. Die Daten werden ansonsten deskriptiv ausgewertet.

4.6 Konfidenzintervalle (Vertrauensbereiche)

Bei Wiederholung einer Studie erhalten wir andere statistische Kennzahlen (vgl. Abschnitt 3.8). Dies ist erklärbar durch die unterschiedlichen Stichproben, was notwendigerweise einen Zufallseffekt mit sich bringt. Man möchte diese Zufallsschwankungen in den statistischen Kennzahlen quantitativ fassen.
Da der wahre Kennwert θ (zum Beispiel $\theta = \mu$, p) nicht bekannt ist, und die Schätzung mit statistischer Ungenauigkeit behaftet ist: Gibt es ein Intervall, in dem θ mit hoher Wahrscheinlichkeit liegt ? („Quantifizierung der Ungenauigkeit“)

Definition: Ein 95%-**Konfidenzintervall** $\left[\hat{\theta}_u, \hat{\theta}_o\right]$ ist ein zufälliges Intervall, das den unbekannten, wahren Wert θ mit Wahrscheinlichkeit 95% enthält.

In Formeln heisst dies: $\mathrm{P}\left[\hat{\theta}_u \leq \theta \leq \hat{\theta}_o\right] \geq 0.95$

Man kann auch allgemein $(1-\alpha)\times 100\%$ Konfidenzintervalle definieren, konventionell wird aber $\alpha = 0.05$ gesetzt.

Bei wiederholten Experimenten liegt man demnach in $\alpha \times 100\%$ der Fälle falsch. Es wird offensichtlich, dass Konfidenzintervalle etwas mit dem Konzept des Signifikanztests zu tun haben, weshalb wir sie hier einführen.

Nachfolgend ein Beispiel für die Wichtigkeit von Konfidenzbereichen für die biomedizinische Literatur. Publiziert wurde in Lancet eine multizentrische Studie zum akuten Herzinfarkt. Die Ergebnisse stützen sich stark auf Konfidenzintervalle ab, so dass diese bereits in der Zusammenfassung erscheinen.

Summary

In 1985 an overview of clinical trials confirmed that patients treated within 6 h of the onset of symptoms of myocardial infarction benefited from thrombolytic therapy. Doubt remained about treatment later than this and this uncertainty prompted further randomised studies. The South American multicentre trial EMERAS is one of these.

4534 patients entering hospital up to 24 h after the onset of suspected acute myocardial infarction were randomised between intravenous streptokinase (SK) 1·5 MU and placebo, during the period January, 1988, to January, 1991. Once the results of ISIS-2 were known, only patients presenting more than 6 h after symptom onset were randomised. There was no significant difference in mortality during the hospital stay (269/2257 [11·9%] deaths among SK patients *vs* 282/2277 [12·4%] in controls). Among the 2080 patients presenting 7–12 h from symptom onset there was a non-significant trend towards fewer deaths with SK (11·7% SK *vs* 13·2% control; 14% [SD 12] reduction with 95% confidence interval [CI] of 33% reduction to 12% increase), whereas there was little difference among the 1791 patients presenting after 13–24 h (11·4% *vs* 10·7%; 8% [16] increase with a 95% CI of 20% reduction to 45% increase). These 95% CIs are wide and are consistent with the results of previous studies among patients presenting late after symptom onset.

The EMERAS results, though not conclusive on their own, do contribute substantially to accumulating evidence on the question of whether fibrinolytic therapy really does produce any worthwhile improvement in survival among such patients.

Lancet 1993; **342:** 767–72

Wie man sieht, stehen Konfidenzintervalle in einem engen Zusammenhang zur Prüfung von Hypothesen. In der obigen Studie konnte kein Unterschied in der Sterblichkeit nach einem Herzinfarkt zwischen den Gruppen mit und ohne thrombolytischer Therapie nachgewiesen werden ("no significant difference"). Das heisst aber nicht, dass es keinen Unterschied gibt. Die Konfidenzintervalle zeigen, dass es möglich ist, dass die Therapie klinisch relevante Verbesserungen bis zu 33% bringt. Das muss dann allerdings erst noch durch neue Studien belegt werden, da natürlich auch die andere Grenze des Konfidenzintervalls (Verschlechterung um 12%) möglich wäre. Der Bezug zum Hypothesentesten ist genau gesagt so, dass ein Resultat mit $\alpha = 5\%$ signifikant ist, wenn der Wert der Nullhypothese ausserhalb des 95%–Konfidenzintervalls liegt.

4.6.1 Konfidenzintervall für μ bei bekanntem σ^2

Aus tutoriellen Gründen sei eine etwas vereinfachte Situation vorausgesetzt.

- Eine physikalische Grösse soll mit einem Messgerät mit **bekannter** Streuung $\sigma = \sigma_0$ bestimmt werden.
- Die Messungen $x_1, \ldots, x_n$ seien verteilt wie $\mathcal{N}(\mu, \sigma_0^2)$, d. h. sie sind normalverteilt mit Erwartungswert μ und Varianz σ_0^2.
- Gesucht ist das $(1-\alpha)$–Konfidenzintervall für μ.

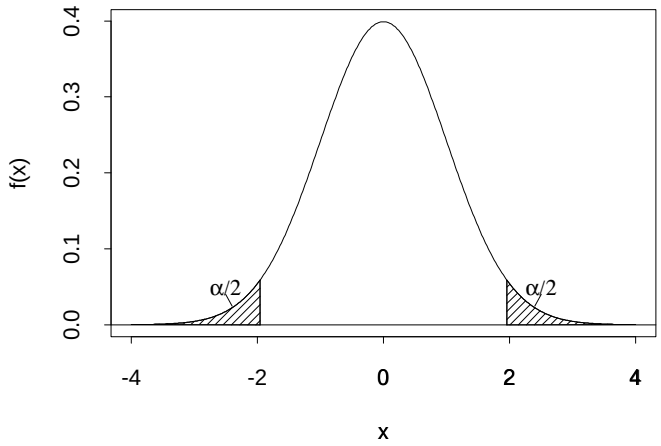
(i) Der Mittelwert $\bar{x}$ ist verteilt als $\mathcal{N}(\mu, \frac{\sigma_0^2}{n})$, da $\bar{x}$ die Varianz $\frac{\sigma_0^2}{n}$ hat.

(ii) Dann ist $\dfrac{\bar{x} - \mu}{\sigma_o \,/\sqrt{n}}$ verteilt als $\mathcal{N}(0,1)$.

(iii) $(1-\alpha)$–Konfidenzintervall für Erwartungswert μ bei bekanntem σ_0:

$$\bar{x} - z_{1-\alpha/2}\,\frac{\sigma_0}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\alpha/2}\,\frac{\sigma_0}{\sqrt{n}}$$

Begründung für die Formel des Konfidenzintervalls (für Interessierte):



In der Abbildung ist $\alpha = 0.05$ angenommen, mit $z_{\alpha/2} = -1.96$ und $z_{1-\alpha/2} = 1.96$.

- Nach Definition ist $\mathrm{P}\left[z_{\alpha/2} \le \dfrac{\bar{x} - \mu}{\sigma_0/\sqrt{n}} \le z_{1-\alpha/2}\right] = 1 - \alpha$.
- Da die Normalverteilung symmetrisch ist, folgt $z_{\alpha/2} = -z_{1-\alpha/2}$.
- Auflösen nach μ liefert das $(1-\alpha)$–Konfidenzintervall:

$$-z_{1-\alpha/2}\,\frac{\sigma_0}{\sqrt{n}} \le \bar{x} - \mu \le z_{1-\alpha/2}\,\frac{\sigma_0}{\sqrt{n}}$$

$$\Longrightarrow \bar{x} - z_{1-\alpha/2}\,\frac{\sigma_0}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\alpha/2}\,\frac{\sigma_0}{\sqrt{n}}$$

Bemerkungen:

1. Das Konfidenzintervall für μ liegt symmetrisch um $\bar{x}$, was plausibel ist. Seine Länge wird kleiner (man wird „sicherer“), wenn die Standardabweichung σ_0 sinkt oder die Stichprobe grösser wird.

2. Jedes Konfidenzintervall ist zufällig, in diesem Beispiel ist es zufällig durch seine Lage bei $\bar{x}$.

3. Die Annahme, dass die Standardabweichung der Stichprobe $\sigma = \sigma_0$ bekannt sei, ist im allgemeinen unrealistisch, σ muss geschätzt werden (siehe unten).

Numerisches Beispiel zur Illustration der Abhängigkeit des Konfidenzintervalls für μ von α und n: $\bar{x} = 0.2$, $\sigma_0 = 0.1$

	α		
n	0.05	0.01	0.001
10	[0.14, 0.26]	[0.12, 0.28]	[0.10, 0.30]
50	[0.17, 0.23]	[0.16, 0.24]	[0.15, 0.25]
200	[0.19, 0.21]	[0.18, 0.22]	[0.18, 0.22]

Wir sehen eine „Unschärferelation“: Je sicherer wir sein wollen, dass der wahre Wert im Intervall liegt, desto länger wird dann das Konfidenzintervall.

4.6.2 Konfidenzintervall für μ bei unbekanntem σ^2

Beispiel: Konfidenzintervall für den Erwartungswert μ der Anzahl T_4–Zellen auf der Basis von $n = 20$ Hodgkin–Patienten

Problem: Die Daten sind rechtsschief, deutlich nicht normalverteilt.

Lösung: 1. Daten logarithmieren
2. logarithmierte Daten als approximativ normalverteilt betrachten

log–T_4–Zellanzahl: $\bar{x} = 6.49$, $s = 0.71$

Die Idee ist nun, die Statistik $\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ von vorhin weiter zu verwenden, aber σ durch s zu ersetzen: $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

Die Statistik t wäre standardnormalverteilt, falls wir σ und nicht s im Nenner hätten. So gilt: t ist t–verteilt, (siehe Abschnitt 3.4), „wackelt draussen mehr“, da s ein Schätzer und keine feste Zahl wie σ ist („mehr Wahrscheinlichkeit in den Extrembereichen“).

Dann ergibt sich analog:

$(1-\alpha)$–Konfidenzintervall für μ bei unbekanntem σ:

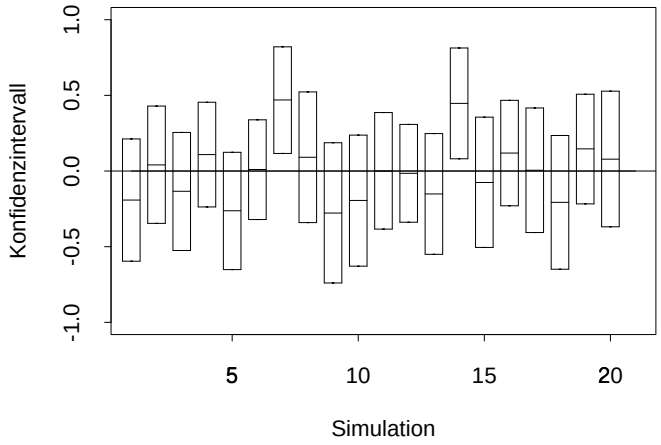
$$\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}$$

- Die Grösse $t_{1-\alpha/2}$ ist das $(1-\alpha/2)\times 100\%$–Perzentil einer t–Verteilung (mit $n-1$ Freiheitsgraden).
- Das Intervall ist symmetrisch um $\bar{x}$.
- Die Länge des Intervalls ist abhängig von n und s, d. h. die Lage und die Länge des Intervalls sind jetzt zufällig.

Für die log–T_4–Zellanzahl erhalten wir folgende $(1-\alpha)$–Konfidenzintervalle:

$$\begin{aligned} \alpha = 0.05 \quad & 6.14 \leq \mu \leq 6.84 \\ \alpha = 0.01 \quad & 5.90 \leq \mu \leq 6.99 \\ \alpha = 0.001 \quad & 5.76 \leq \mu \leq 7.22 \end{aligned}$$

Nach Konstruktion enthält ein Konfidenzintervall im Mittel in $(1-\alpha)\times 100\%$ der Fälle den wahren Wert. Auf dem Computer wurden 20 Stichproben von $n = 25$ normalverteilten „Pseudo–Zufallszahlen“ simuliert ($\mu = 0$, $\sigma^2 = 1$). Für jede Stichprobe wurde das 95%–Konfidenzintervall für μ nach obiger Formel konstruiert. Nicht nur die Lage, sondern auch die Länge ändert sich beträchtlich. Im Mittel würden wir auf 20 Stichproben in einem Fall erwarten, dass der wahre Wert nicht im Konfidenzintervall liegt. Hier enthalten zufälligerweise zweimal die Konfidenzintervalle den wahren Wert nicht. Deswegen müssen wichtige Ergebnisse repliziert werden, denn nach der Produktregel für Wahrscheinlichkeiten ist es äusserst unwahrscheinlich, in 2 unabhängigen Experimenten zufällig extreme Resultate zu erhalten.



4.6.3 Konfidenzintervall für relative Häufigkeiten

Man kann für praktisch alle interessierenden Grössen Konfidenzintervalle berechnen. Wichtig ist z. B. ein Konfidenzintervall für eine wahre relative Häufigkeit, z. B. eine Prävalenz. (Der Schätzwert für p ist $\hat{p} = k/n$, wenn z. B. bei einer Prävalenzschätzung k von n Personen eine Krankheit haben.) Die Grenzen eines Konfidenzintervalls werden relativ kompliziert berechnet , so dass wir nur Beispiele geben:

Beispiel: Man beobachtet $n = 20$ Geburten, $k = 7$ mal wird ein Knabe geboren.

$\hat{p} = 7/20 = 0.35$

Das 95%–Konfidenzintervall ist (0.15, 0.59).

- Der Vertrauensbereich ist **weit**, denn n ist klein.
- Der Vertrauensbereich schliesst 0.5 mit ein (gleiche Häufigkeit der Knaben- und Mädchengeburten mit $p = 0.5$ ist möglich).

Das folgende reale Beispiel zeigt, dass die Annahme gleicher Häufigkeit von Knaben– und Mädchengeburten doch nicht plausibel ist:

1950 – 1970: $1'944'700$ Geburten in der Schweiz, davon $997'600$ Knaben.

Ist es Zufall, dass der Schätzwert $\hat{p} = 0.5130$ für eine Knabengeburt von $p = 0.5$ abweicht? Das 99%–Konfidenzintervall ist (0.5121, 0.5139); d. h. das Konfidenzintervall ist bei diesem grossen n sehr eng. Es schliesst 0.5 (gleiche Wahrscheinlichkeit einer Knabengeburt) deutlich nicht mit ein. Damit ist mit $\alpha = 1\%$ nachgewiesen, dass Knabengeburten häufiger sind. Spekulativ könnte man einen Mechanismus der Natur postulieren, der für mehr Knabengeburten sorgt, um ihre höhere Mortalität im ersten Lebensjahr zu kompensieren.

4.6.4 Konfidenzintervalle und Tests

Es gibt einen engen Bezug von Konfidenzintervallen und Tests: Ein $(1 - \alpha)$–Konfidenzintervall enthält die Information eines zweiseitigen Tests zur Irrtumswahrscheinlichkeit α im folgenden Sinne:

(1) Wenn der hypothetische Wert (z. B. $\mu = \mu_0$) im Konfidenzintervall liegt, kann man H_0 nicht verwerfen.

(2) Wenn der hypothetische Wert ausserhalb des Konfidenzintervalls liegt — also bezogen auf die Daten unwahrscheinlich ist — dann wird die statistische Hypothese H_0 abgelehnt.

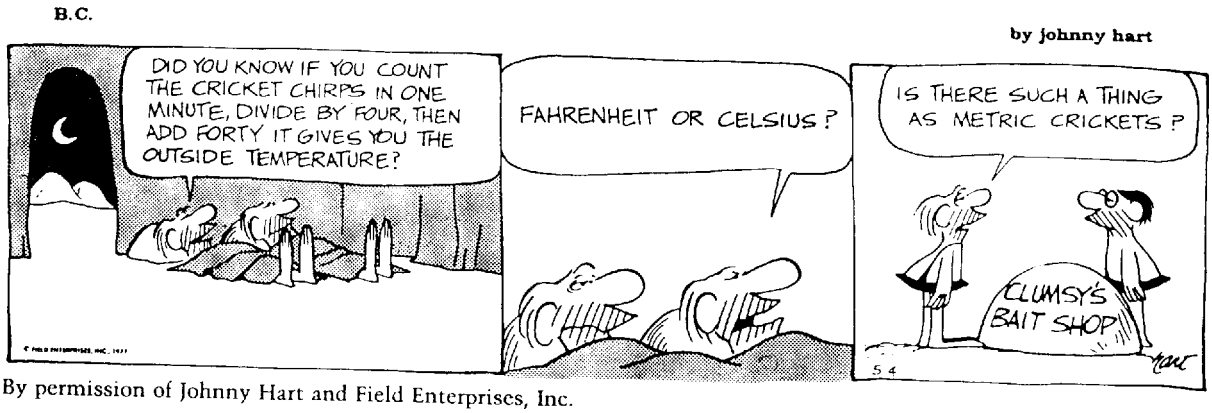
Konfidenzintervalle kombinieren Information von Schätzwerten und Tests.

5 Korrelation und Regression

Bisher wurde die statistische Analyse auf der Basis **einer** Messgrösse behandelt, d. h. univariat (Ausnahme: χ^2–Test). Jetzt geht es darum, Zusammenhänge zwischen **zwei** oder mehr stetigen Variablen (**bivariate, multivariate** Daten) zu untersuchen.

Mögliche Fragestellungen sind:

- Besteht eine Beziehung zwischen den Variablen ?
- Wie stark ist die Beziehung ?
- Welche Form hat die Beziehung ?
- Kann eine Variable von primärem Interesse aus der Beobachtung anderer Variablen vorhergesagt werden ?



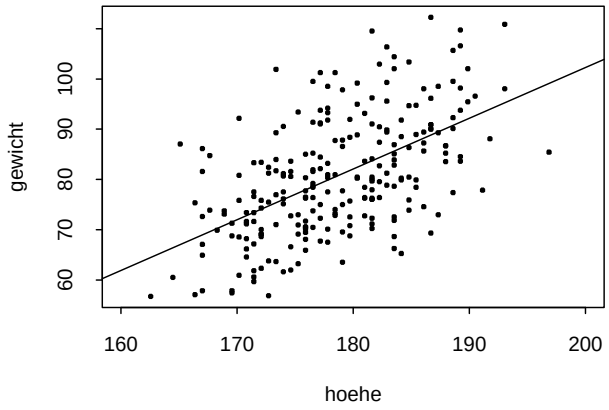
By permission of Johnny Hart and Field Enterprises, Inc.

5.1 Bivariate Daten

- Man beobachtet zwei **stetige** Variablen (x, y) an der selben Beobachtungseinheit, und erhält **paarweise** Beobachtungen: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
  Beispiel: Zusammenhang zwischen Gewicht und Grösse.
- Jede Korrelations– oder Regressionsanalyse sollte mit dem Ausdruck des **Scattergramms** (Streudiagramm) begonnen werden.

Die Daten, die wir in diesem Kapitel analysieren wollen, wurden von R. W. Johnson (Carleton College, Northfield, MN) zur Illustration der multiplen Regression und der Modellwahl zur Verfügung gestellt. In dieser Stichprobe wurden bei 252 Männern der Prozentsatz an Körperfett, Alter, Gewicht, Körpergrösse und 10 Körperumfangsmasse bestimmt. Dabei wurde die Körperdichte mittels einer Unterwasser–Wiege–Technik relativ aufwendig gemessen und mittels gewisser Eichformeln in prozentuales Körperfett umgerechnet. Ein hoher Anteil an Körperfett stellt ein Gesundheitsrisiko dar, aus praktischen Gründen möchte man

das Körperfett aus einfachen Körpermessungen approximativ bestimmen, wofür sich die Regressionsmethode eignet. Es erweist sich, dass man das Körperfett von Männern mittels multipler Regression bequem aus leicht zu bestimmenden Massen schätzen kann (siehe Abschnitt 5.6). Das folgende Scattergramm zeigt den Zusammenhang von Gewicht und Körpergrösse in der von Ausreissern befreiten Stichprobe von 241 Männern.



Man erhält so einen visuellen Eindruck vom Zusammenhang zwischen den Variablen. In diesem Fall erhärtet sich die Vermutung, dass Gewicht und Körpergrösse (positiv) zusammenhängen.

5.2 Korrelation: Definition und Eigenschaften

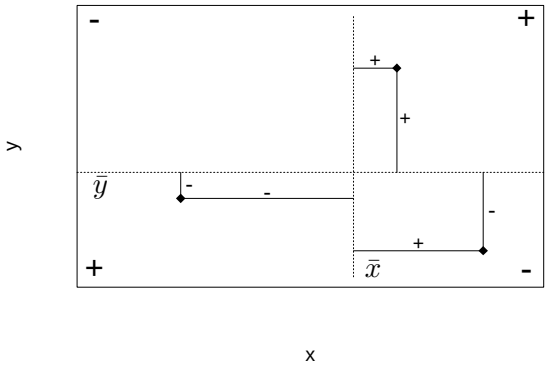
Eine Korrelation (auch Produkt–Moment Korrelation oder Pearson Korrelation genannt) misst, wie stark der **lineare** Zusammenhang, die lineare Übereinstimmung zwischen x und y ist.
Der Korrelationskoeffizient wird wie folgt berechnet:

Kovarianz: $\mathrm{Cov}(x,y) = s_{xy} = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$

Varianz: $s_x^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \qquad s_y^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$

Korrelation:
$$r = \frac{s_{xy}}{s_x\, s_y} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}}$$

Die Plausibilität des Zählers kann man anhand von 3 Datenpunkten graphisch verstehen:



Die beiden Mittelwerte $\bar{x}$ und $\bar{y}$ teilen die Ebene und damit die Daten in 4 Rechtecke ("Quadranten"). Rechts oben und links unten von den Mittelwerten der Variablen geben Beobachtungen einen positiven Beitrag zur Kovarianz und damit zur Korrelation; links oben und rechts unten einen negativen Beitrag. Damit ist die Korrelation ungefähr 0, wenn sich die Beobachtungen auf alle 4 Quadranten verteilen. Sie wird deutlich positiv, wenn sich die Beobachtungen um eine Gerade mit positiver Steigung gruppieren, d. h. im Quadranten I (rechts oben) und III (links unten). Bei Messungen vorwiegend im Quadranten II und IV wird sie negativ.

Plausibilität des Nenners:

Die Korrelation r wird durch diese Normierung mit den Standardabweichungen von den Masseinheiten unabhängig und ist damit besser interpretierbar.

Eigenschaften:

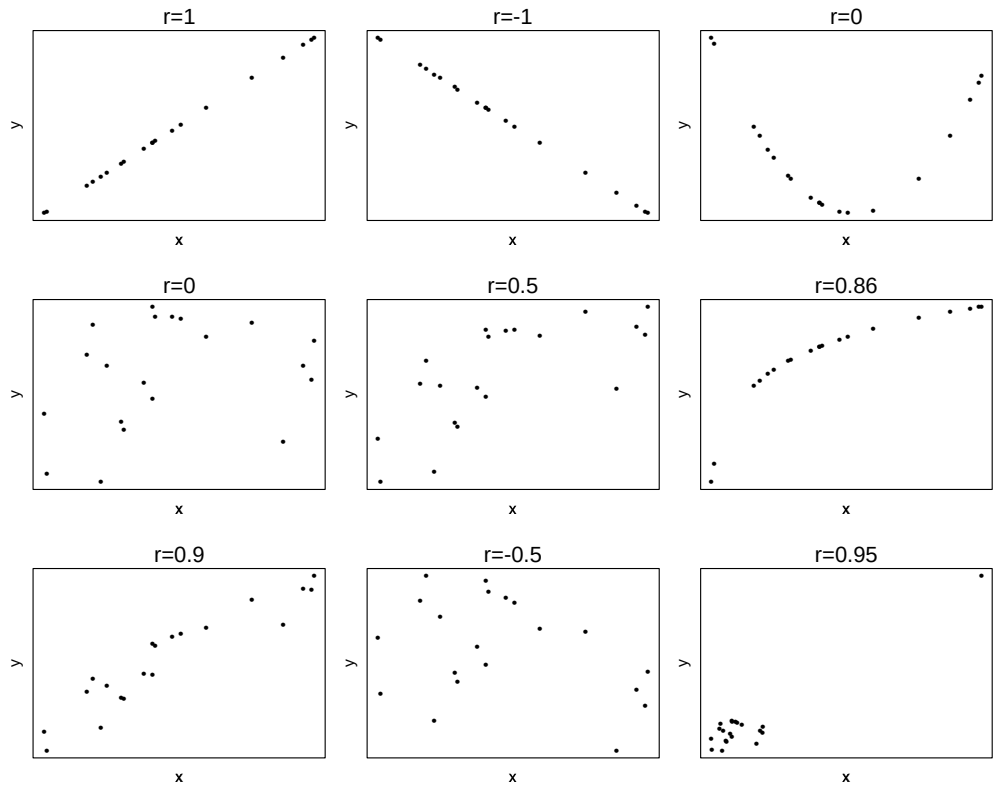
$-1 \leq r \leq 1$

$r = 1$	$\Longrightarrow$	deterministisch positiver linearer Zusammenhang zwischen x und y
$r = -1$	$\Longrightarrow$	deterministisch negativer linearer Zusammenhang zwischen x und y
$r = 0$	$\Longrightarrow$	kein linearer Zusammenhang

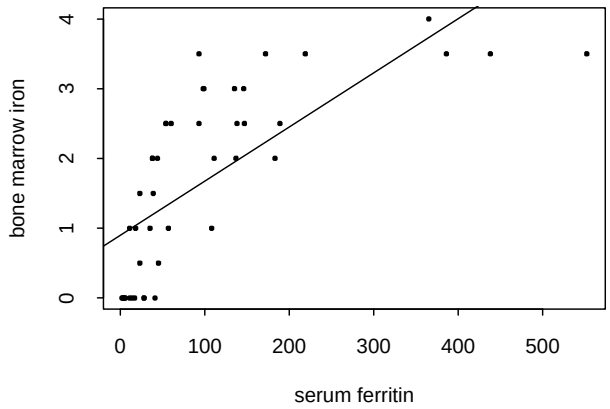
Allgemein gilt:

- Das Vorzeichen gibt die Richtung des Zusammenhangs an.
- Die Grösse gibt die Intensität des Zusammenhangs wieder.

Die folgenden Illustrationen zeigen Daten mit unterschiedlich starkem Zusammenhang. Man beachte, dass die Korrelation nur den **linearen** Zusammenhang misst, unter Umständen kann die Korrelation für gewisse deterministische nicht–lineare Zusammenhänge Null werden. Werte, die in der x– und y–Richtung extrem liegen, können eine starke Korrelation vortäuschen.



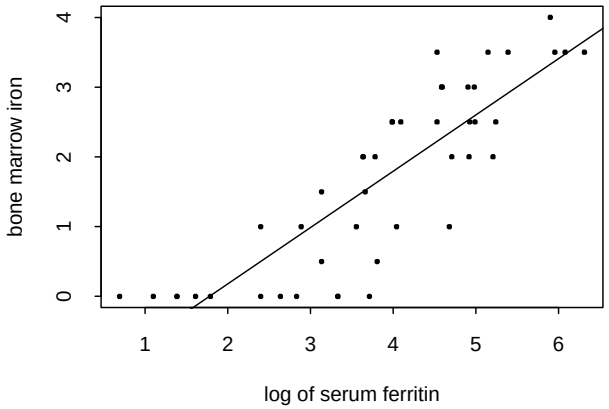
Beispiel: In einer Studie an 45 anämischen Patienten (Baumann Kurer et al., British J. Haematology, 1995) wurde untersucht, ob die invasive Messung des Eisengehaltes im Knochenmark durch eine einfache Blutprobe (Ferritingehalt im Blutserum) ersetzt werden kann.



Die Stichprobenkorrelation ist $r = 0.72$. Man sieht aber, dass die Werte nicht gleichmässig um eine Gerade streuen, eine Gerade nicht zu diesen Daten passt. Die Korrelation wird stark durch weit aussen liegende Werte oben rechts bestimmt.

- Da lineare Zusammenhänge am einfachsten zu behandeln sind, sollte man versuchen, durch Transformation auf einen linearen Zusammenhang zu kommen.
- Da weit vom Mittelwert entfernte Beobachtungen die Korrelation stark beeinflussen, sollte man versuchen, die beiden Variablen so zu transformieren, dass sie annähernd normalverteilt sind.

Häufig erfüllen Transformationen die beiden Forderungen gleichzeitig.



Im Beispiel ist die Stichprobenkorrelation nach der log–Transformation des Serum Ferritins $r = 0.85$, die Daten streuen gleichmässig um eine Gerade (diese heisst Regressionsgerade, siehe unten).

5.3 Testen auf linearen Zusammenhang und Konfidenzintervalle

Als nächstes soll geprüft werden, ob überhaupt ein linearer Zusammenhang zwischen den beiden Variablen statistisch nachzuweisen ist.

Nullhypothese: Die wahre Korrelation ρ ist gleich 0 („kein Zusammenhang“).

Annahme: (x, y) seien gemeinsam **normalverteilt**

Die folgende Testgrösse T ist mathematisch–statistisch begründet, sie folgt einer t–Verteilung mit $n - 2$ Freiheitsgraden:

Teststatistik:

$$T = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

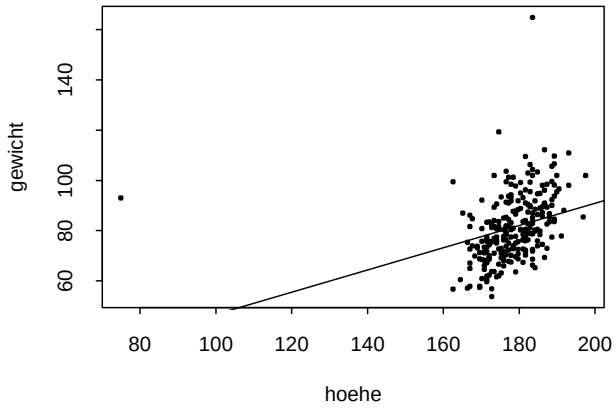
Beispiel: Zusammenhang von Gewicht und Körpergrösse bei Männern
$n = 241\,, \quad r = 0.55$
$\Longrightarrow T = 7.9 > t_{239;0.975} = 1.97,\ p < 0.0001$
Es besteht also ein signifikanter Zusammenhang.

Die Angabe eines Konfidenzintervalls ist für die Korrelation noch wichtiger als für die bisherigen statistischen Kennwerte. Mit etwas Erfahrung vermittelt die Angabe von n und $(\bar{x}, s)$ auch ein Gefühl für die Variabilität der Kennwerte, was aber bei der Korrelation nicht der Fall ist. Das **Konfidenzintervall** liefert einen Bereich, in dem die wahre Korrelation mit grosser Wahrscheinlichkeit liegt.
Die Berechnung des Konfidenzintervalls erfolgt approximativ, indem r durch eine sogenannte z–Transformation auf eine annähernd normalverteilte Grösse transformiert wird. Dabei zeigt sich auch, dass Konfidenzintervalle für Korrelation nahe bei 0 weiter sind als solche nahe bei ± 1.
Im Beispiel ist ρ im Intervall $(0.46, 0.64)$ mit Wahrscheinlichkeit $1 - \alpha = 0.95$ (95%–Konfidenzintervall für ρ). Konfidenzintervalle für die Korrelation können ziemlich gross werden, wenn die Stichprobe mässigen Umfang hat.

5.4 Ausreisser und Gefahren der Korrelationsrechnung

- Wie wirken sich Ausreisser aus und wie behandle ich sie ?
- Wie kann man bei Nicht–Normalverteilung die Korrelation schätzen und testen ?

Beispiel: Gewicht und Körpergrösse von Männern
In der vollen Stichprobe von 252 Männern vor Elimination von Ausreissern sah das Scattergramm so aus:



Wir erhalten $r = 0.31$ und $p < 0.0001$. Die Ausreisser senken also die Korrelation von 0.55 auf 0.31, eine deutliche Verfälschung der Realität. Negativ wirkt sich vor allem der unsinnige Wert von knapp 80 cm für die Körpergrösse bei einem Gewicht von über 90 kg aus (vermutlich ein Tippfehler).

Spearmans Rangkorrelation

Ähnlich wie wir früher auf Rangverfahren zurückgegriffen haben, um von der Normalverteilungsannahme loszukommen (Beispiel Mann–Whitney–Test), stützen wir uns auch hier auf die Rangordnung der Daten ab, um robustere Resultate zu erhalten.

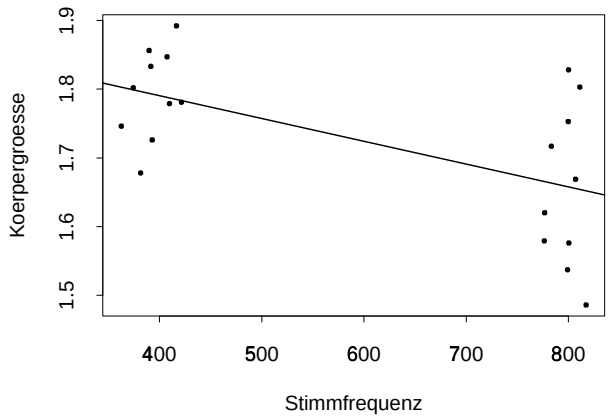
Vorgehen:

1. Man bringt $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ getrennt in Rangreihen.
2. Man korreliert die Ränge miteinander anstatt die Zahlen selber.

Dadurch wird der Einfluss von Ausreissern begrenzt. Im Beispiel ergibt sich jetzt Spearmans Rangkorrelationskoeffizient $r_S = 0.52$ und $p < 0.0001$ (bei den korrekten Daten: $r_S = 0.55$, $p < 0.0001$). Rangkorrelationen können die Daten nicht „reparieren“, aber die Auswirkungen falscher oder atypischer Werte mildern.

Gefahren der Korrelations–Rechnung

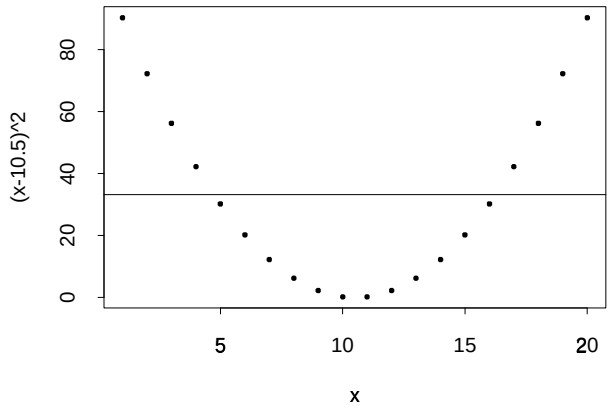
1. Bei 10 Variablen gibt es 45 mögliche Korrelationen. Man muss sich demnach davor hüten, einzelne signifikante Korrelationen überzubewerten (siehe Problem des multiplen Testens).
2. Allgemeine Trends führen zu Scheinkorrelationen über die Zeit: Der Preis von Benzin und die Scheidungsrate korrelieren, da beide einen Zeittrend aufweisen.
3. Heterogenitätskorrelation: Am Beispiel von Stimmfrequenz und Körpergrösse bei Männern und Frauen sieht man eine negative Korrelation ($r = -0.60$, $p = 0.006$), obwohl die Werte sowohl für Männer als auch für Frauen unkorreliert sind.



4. Trivialkorrelationen: Wenn man die Körpergrösse mit 12 ($= x$) und mit 20 Jahren ($= y$) misst, müssen die Werte gut korrelieren, da ja $y = x + z$ gilt, wobei z der Zuwachs von 12 bis 20 Jahren ist.

5. Konfundierung durch 3. Variable: Die Anzahl der Störche und Geburten in einem Kanton korreliert stark („Bringt der Storch die Kinder?“). Die konfundierende Variable, die beide Zahlen gleichsinnig beeinflusst, ist hier die Grösse des Kantons.

6. nichtlineare Zusammenhänge:



Hier ist $r = 0$, obwohl ein deterministischer aber quadratischer Zusammenhang deutlich ist.

7. Extreme Datenpunkte: Das frühere graphische Beispiel mit einem in der x– und der y–Richtung extrem verschobenen Wert zeigt, dass eine grosse Korrelation (im Beispiel 0.95) durch einzelne Werte vorgetäuscht werden kann, ohne dass ein allgemeiner Zusammenhang besteht. Aber auch extreme Datenpunkte in die x– oder y–Richtung allein können die Korrelation übermässig beeinflussen.

5.5 Einfache lineare Regression

- Die einfache Regressionsanalyse ist die statistische Analyse der Wirkung **einer** Variablen x auf eine andere stetige Variable y. Die Beziehung ist also im Unterschied zur ungerichteten Korrelationsanalyse gerichtet.

x = unabhängige Variable, erklärende Variable, Prädiktor (oft nicht zufällig: Zeit, Alter, Messpunkt)
y = abhängige Variable, erklärte Variable, Zielvariable, Outcome, Response

Ziel: Nicht nur die Stärke und Richtung ($\nearrow, \searrow$) des Zusammenhangs soll bestimmt werden, sondern es soll ein quantitatives Gesetz formuliert werden: Wie ändert sich y, wenn x sich ändert ? Wichtig ist auch, dass im Fall deterministischer x–Variablen eine Regression sinnvoll sein kann, wenn es die Korrelation nicht ist.

Der Name ist aufgrund eines **Missverständnisses** entstanden:

- „regression“ = Zurückentwicklung, Gegenteil von „progression“
- Galton, F. (1885). Regression toward mediocrity in hereditary stature.

- Galton bemerkte, dass die Söhne kleiner Eltern kleiner als der Durchschnitt sind, aber die Grösse **kehrt zum Mittelwert zurück**.

Beispiel: Das Körpergewicht ist ein naheliegendes Mass für Übergewicht. Wie wir aber gesehen haben (Scattergramm S. 72), hängt es von der Körpergrösse ab und ist demnach als Mass für Übergewicht nicht geeignet. Die folgende Regressionsgleichung quantifiziert dies:

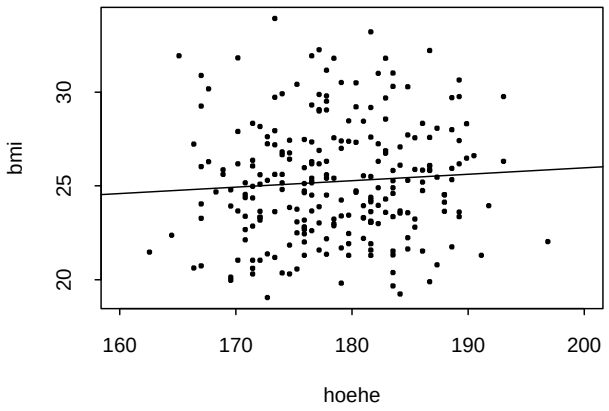
Regression: y = Gewicht, x = Höhe, $n = 241$

$$\widehat{y} = -99.7 + 1.01 \times x, \quad r^2 = 0.31, \quad p < 0.0001$$

Wie schwer sind Männer? Die beste Voraussage ist $\bar{y} = 80.7\,\text{kg}$. Die Standardabweichung der Messwerte um diesen Vorhersagewert beträgt SD $= s_y = 11.8\,\text{kg}$.
Wie schwer sind Männer von 175 cm? Die Zusatzinformation der Körpergrösse ändert die Vorhersage des Gewichts, indem der Wert der Regressionsgerade für $x = 175\,\text{cm}$ genommen wird: $\hat{y} = -99.7 + 1.01 \times 175 = 77.0\,\text{kg}$. Die Standardabweichung der Messwerte um diesen Vorhersagewert beträgt $s_e = 9.8\,\text{kg}$. Wir erhalten demnach exaktere Aussagen über das Gewicht, wenn wir die Körpergrösse berücksichtigen. Allgemein gilt, dass wir exaktere Ergebnisse erhalten, wenn wir über ein Regressionsmodell wichtige Einflussgrössen berücksichtigen.
In der Klinik wird oft der „body mass index“ (BMI = Gewicht/Höhe^2) als Mass für Übergewicht benutzt. Die Frage ist, ob der BMI tatsächlich von der Körpergrösse nicht beeinflusst ist.

Regression y = BMI = Gewicht / Höhe^2, x = Höhe

$$y = 19.2 + 0.034 \times x, \quad r^2 = 0.005, \quad p = 0.27$$



Der BMI ist nicht (oder vernachlässigbar) mit der Körpergrösse korreliert. Damit liefert er ein einfaches Mass für Übergewicht.

5.5.1 Statistisches Modell der Regression

$$y_i = f(x_i) + \varepsilon_i \quad i = 1, \ldots, n$$

f = Regressionsfunktion, „wahrer Verlauf“

ε_i = unbeobachtbare, zufällige Schwankungen (Fehler oder Rauschen). Die Residuen ε_i schwanken um 0 ($\mathrm{E}[\varepsilon_i] = 0$) und haben die Varianz σ^2.

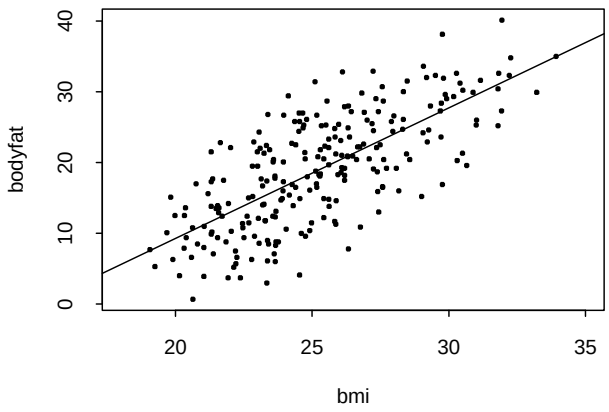
Das Problem der Bestimmung von f vereinfacht sich sehr, wenn wir f als eine lineare Funktion annehmen („lineare Regression“).

$$f(x) = a + b\,x$$

Damit sind nur noch der Achsenabschnitt (intercept) a und die Steigung (slope) b der Geraden $a + b\,x$ unbekannt und müssen aus den Daten bestimmt („geschätzt“) werden.

Beispiel: A priori sind sowohl prozentuales Körperfett als auch BMI als Masse für Übergewicht bei Männern von Interesse.

x = BMI (in kg/m^2)
y = Körperfett (in %)



Geradengleichung:

$$\text{bodyfat} = -27.6 + 1.84 \times \text{bmi}, \quad r^2 = 0.52, \quad p < 0.0001$$

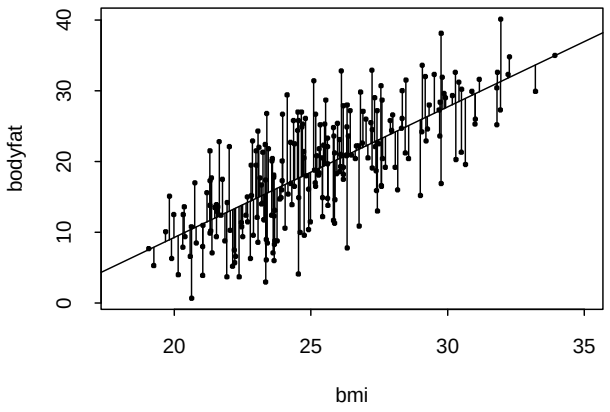
Interpretationen:

1. Männer mit einem BMI von 25 kg/m^2 haben im Mittel 18 % Körperfett.
2. Männer mit einem um 1 kg/m^2 erhöhten BMI haben im Mittel 2 % mehr Körperfett.

5.5.2 Methode der kleinsten Quadrate

Die Anpassung der Geraden an die Daten ist intuitiv dann am besten, wenn die Abstände der Messpunkte von der zu bestimmenden Geraden im Mittel klein sind. („method of least squares“)

- Übliche Methode zur Schätzung von a und b: es werden die vertikalen Abstände zur Geraden betrachtet, die zu bestimmen ist, und zwar — wie bei der Varianz — in der quadrierten Form.



Sei $\hat{y}_i$ der Wert der geschätzten Regressionsgerade ($= \hat{a} + \hat{b}\, x_i$) beim Wert x_i.

Wähle die Schätzung der Parameter so, dass die quadratische Abweichung

$$S(\hat{a}, \hat{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

minimal wird!

Die resultierenden Schätzwerte für a und b sind:

$$\text{Steigung: } \hat{b} \;=\; \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r\, \frac{s_y}{s_x}$$

$$\text{Achsenabschnitt: } \hat{a} \;=\; \bar{y} - \hat{b}\, \bar{x}$$

Man erhält diese Formeln über die Lösung eines Systems zweier linearer Gleichungen in $\hat{a}$ und $\hat{b}$.

5.5.3 Durch die Regression erklärte Varianz

Wir haben gesehen, dass eine Variable y um so besser durch die Variable x erklärt werden kann, je grösser die Korrelation r zwischen beiden Variablen dem Betrage nach ist. Wir können also einen Teil der Variabilität von y durch die Regression auf x erklären. Aus mathematischen Überlegungen ergibt sich, dass diese „erklärte Varianz“ als

$$s_{reg}^2 = r^2 s_y^2$$

berechnet werden kann. Die Grösse r^2 gibt also den Anteil der Varianz von y, der durch Kenntnis von x erklärt wird und ist in diesem Sinne bedeutsamer als r selber („Bestimmtheitsmass“). Im Spezialfall $r = \pm 1$ liegen alle Punkte auf einer Geraden, und es bleibt keine Variabilität bei y mehr übrig, wenn man x berücksichtigt. Die Varianz („Residualvarianz“)

$$s_{res}^2 = (1 - r^2) s_y^2$$

ist der Teil, der übrigbleibt, ein Schätzer der Varianz σ^2 der Residuen ε_i. In der Figur von Abschnitt 5.5.2 ist s_{res}^2 die Varianz der Abstände der Daten von der Regressionsgeraden.

5.5.4 Tests und Konfidenzintervalle in der linearen Regression

Hat die Variable x überhaupt einen Einfluss auf y, d. h. ist $b \neq 0$? Statistisch möchte man testen, ob sich y mit x systematisch ändert. Dabei benötigen wir die folgenden mathematischen Annahmen:

- Die Fehler ε_i sind unabhängig.
- Die Fehler sind normalverteilt $\mathcal{N}(0, \sigma^2)$ mit konstanter Varianz σ^2.

Nullhypothese: $b = 0$

Die Hypothese ist äquivalent dazu, dass die Korrelation Null ist (siehe oben). Unter der Annahme einer gemeinsamen Normalverteilung von (x, y) ergibt sich also der gleiche Test wie auf Korrelation $\rho = 0$ mittels der t–Verteilung.

In der Regressionsanalyse gilt:

- Alle Analysen werden **bedingt** für gegebene Werte $x_1, \ldots, x_n$ durchgeführt.

$\Longrightarrow$ Regressionsanalysen sind einfacher als Analysen der Korrelation.

$\Longrightarrow$ Die Verteilung der unabhängigen Variablen x wird nebensächlich.

Beispiel: Körperfett in Abhängigkeit vom BMI bei 241 Männern. Mit dem Programm SPSS erhält man den folgenden Ausdruck:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.718[a]	.516	.514	5.5472

a. Predictors: (Constant), BMI

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-27.617	2.939		-9.398	.000
	BMI	1.844	.116	.718	15.957	.000

a. Dependent Variable: BODYFAT

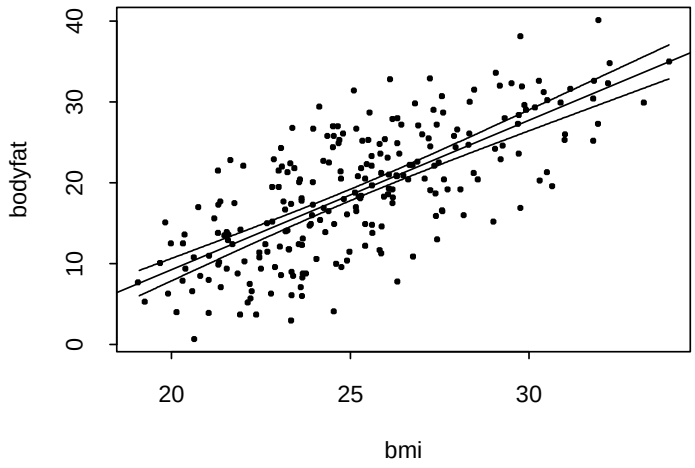
Die Regressionsgerade (Unstandardized Coefficients B) ist

$$\text{bodyfat} = -27.6 + 1.84 \times \text{bmi}$$

Das Bestimmtheitsmass r^2 („R Square“) hat den Wert 0.52, d. h. 52% der Variabilität des Körperfetts können durch die Regression auf den BMI erklärt werden. Die Schätzung der Residualvarianz (aus „Std. Error of the Estimate“ $= \hat{\sigma}$) ist $\hat{\sigma}^2 = 5.55^2 = 30.8$. „Sig.“ sind die p–Werte der statistischen Tests. Der Regressionskoeffizient ist signifikant ($p < 0.001$), also wird ein Zusammenhang bestätigt. Der Wert der Teststatistik („t“) und der p–Wert sind die gleichen wie bei der Prüfung auf Korrelation zwischen beiden Variablen (Daten nicht gezeigt).

Konfidenzintervall für die Regressionsgerade

Statistikprogramme bieten $(1 - \alpha)$–Konfidenzintervalle für den Wert der Regressionsgeraden $a + bx$ für einen gegebenen Wert von x, und tragen diese Konfidenzintervalle in den Regressionsplot ein:



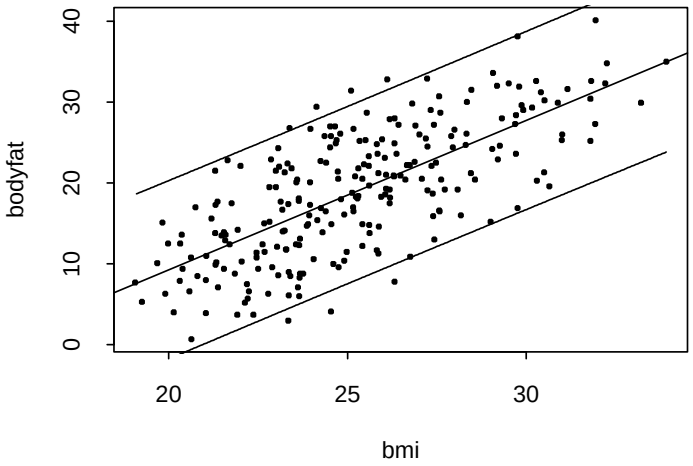
Für einen BMI von z. B. 25 kg/m^2 können wir aus der Graphik ablesen, dass der mittlere Körperfett–Wert mit 95%iger Sicherheit zwischen 17.8 und 19.2 % liegt.

Vorhersageintervall für zukünftige Beobachtungen

Ziel ist es vorauszusagen, in welches Intervall eine zukünftige Beobachtung für gegebenes x^* mit hoher Wahrscheinlichkeit zu liegen kommt.
Das Vorhersageintervall ist wesentlich breiter als das Konfidenzintervall, weil sich die Variabilität der geschätzten Regressionsgerade (Konfidenzintervall) und die Variabilität des Fehlers der zukünftigen Beobachtung (σ^2) addieren.

Achtung: Es besteht Verwechslungsgefahr mit dem Konfidenzintervall.

Die punktweisen Vorhersageintervalle werden üblicherweise in den Regressionsplot für alle x eingetragen:



5.6 Multiple Regression

Wenn man anstatt einer Einflussgrösse x deren k hat $(x_1, \ldots, x_k)$, so könnte man nach dem bisherigen Stoff k einfache (univariate) Regressionsanalysen durchführen. Damit kann man aber das Zusammenspiel der k Prädiktoren nicht erfassen. Deshalb möchte man eine Regressionsanalyse mit allen k Einflussgrössen in einem Modell durchführen. Damit stellt sich das Problem, passende Modelle zu finden und statistisch zu prüfen.
Es gibt eine Reihe von Gründen, anstelle von mehreren einfachen Regressionsanalysen eine multiple Regressionsanalyse durchzuführen:

1. Man möchte mögliche Effekte von zusätzlichen „Stör“–Variablen in einer Studie eliminieren, bei der grundsätzlich nur eine Einflussgrösse von Interesse ist.

   Beispiel: Häufige Störgrösse ist das Alter. y = Blutdruck, x_1 = Dosierung Hypertensivum, x_2 = Alter.

2. Man möchte mögliche Prognosefaktoren erforschen, von denen wir nicht wissen, ob sie alle wichtig oder zum Teil redundant sind.

Beispiel: y = Stenose, x_1 = HDL, x_2= LDL, x_3 = BMI, x_4 = Rauchen, x_5 = Triglyceride.

3. Man möchte eine Formel zur Vorhersage der abhängigen aus den erklärenden Variablen entwickeln.

   Beispiel: y = Erwachsenengrösse, x_1 = Grösse als Kind, x_2 = Grösse der Mutter, x_3 = Grösse des Vaters.

4. Man möchte die Wirkung einer Variablen x_1 auf eine andere Variable y studieren, wobei der Einfluss weiterer Variablen $x_2, \ldots, x_k$ berücksichtigt wird.

5.6.1 Beispiel: Prognostische Faktoren für Körperfett

Anzahl beobachteter Männer: $n = 241$
Abhängige Variable: bodyfat = prozentuales Körperfett
Wir interessieren uns für den Einfluss von 3 unabhängigen Variablen:
BMI in kg/m^2.
Bauchumfang (abdomen, engl. waist).
Quotient von Bauch– und Hüftumfang (waist/hip–ratio).

Wir haben eine multiple Regression des Körperfetts gegen die erklärenden Variablen bmi, abdomen und waist/hip–ratio gemacht:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.825[a]	.681	.677	4.5229

a. Predictors: (Constant), waist/hip-ratio, BMI, ABDOMEN

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-60.045	5.365		-11.192	.000
	BMI	.123	.236	.048	.519	.605
	ABDOMEN	.438	.105	.533	4.183	.000
	waist/hip-ratio	38.468	10.262	.280	3.749	.000

a. Dependent Variable: BODYFAT

Wir sehen, dass x_1 = bmi nicht mehr signifikant ist, wenn wir für x_2 = abdomen und x_3 = waist/hip–ratio kontrollieren. Die letzteren beiden enthalten demnach bereits die Information, die bmi zum Körperfett liefern kann. Wir lassen deshalb die Variable bmi weg. Dies stellt die einfachste Form der Variablenselektion dar. Die multiple Regression liefert jetzt:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.825[a]	.680	.678	4.5159

a. Predictors: (Constant), waist/hip-ratio, ABDOMEN

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-59.294	5.158		-11.496	.000
	ABDOMEN	.484	.057	.588	8.526	.000
	waist/hip-ratio	36.455	9.486	.265	3.843	.000

a. Dependent Variable: BODYFAT

Der Taille–Hüft–Quotient liefert eine signifikante Zusatzinformation zum Bauchumfang. Aus dem Bestimmtheitsmass „adjusted R squared“ sieht man aber, dass die zusätzliche Information klein ist. Eine Ursache besteht darin, dass die Prädiktoren stark zusammenhängen.

In der Regressionsgleichung der multiplen Regression gibt es statt einer mehrere unabhängige Variablen $x_1, \ldots, x_k$:

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$$

In unserem Beispiel sieht die geschätzte Regressionsgleichung für $k = 2$ so aus:

$$\text{bodyfat} = -59.3 + 0.484 \times \text{abdomen} + 36.5 \times \text{waist/hip–ratio}$$

6 Weiterführende Methoden

6.1 Varianzanalyse

6.1.1 Motivation

Die Varianzanalyse ist ein äusserst komplexes Gebiet der Statistik. Wir können in diesem Skript nur die einfachsten Beispiele behandeln, und das auch nur relativ oberflächlich. Um so wichtiger ist es, ein Gefühl für das Gebiet zu vermitteln, um Sie zu befähigen, im Bedarfsfall weitergehende Literatur zu studieren und Ausgaben von Statistik–Programmen zu interpretieren. Das Wort **ANOVA** ist eine Abkürzung für **AN**alysis **O**f **VA**riance.

Wir haben den Zweistichproben–t–Test kennengelernt, mit dem man Mittelwertsunterschiede zwischen zwei Gruppen nachweisen kann. Die Varianzanalyse stellt eine Verallgemeinerung dar, indem Mittelwertsunterschiede zwischen mehreren Gruppen analysiert werden. Trotz des Namens „Varianzanalyse“ geht es weiterhin primär um Mittelwertsunterschiede und nicht um Varianzen. Bei der t–Statistik wird ein Mittelwertsunterschied (im Zähler) mit seinem Standardfehler verglichen (im Nenner). Zur Erinnerung: der Standardfehler ist die Standardabweichung einer Kenngrösse, hier des Mittelwertunterschiedes. Die F–Statistik nimmt diese Idee auf, indem im Zähler **quadrierte** Mittelwertsunterschiede mit komplexeren Formen der Varianzschätzung im Nenner verglichen werden. ($F \approx t^2$).

Beispiel: 22 Bypass–Patienten wurden zufällig in 3 Behandlungsgruppen (unterschiedliche Beatmung) eingeteilt (Amess et al. 1978, Altman, S. 208). Die wissenschaftliche Frage lautete: Unterscheiden sich die Folsäurewerte in den roten Blutzellen?

group	red cell folate
1	243
1	251
1	275
1	291
1	347
1	354
1	380
1	392
2	206
2	210
2	226
2	249
2	255
2	273
2	285
2	295
2	309
3	241
3	258
3	270
3	293
3	328

Die Nullhypothese ist dann, dass die Erwartungswerte in den drei Gruppen identisch sind.

Die zentrale Ausgabe einer Varianzanalyse ist die **ANOVA–Tabelle**:

Tests of Between-Subjects Effects

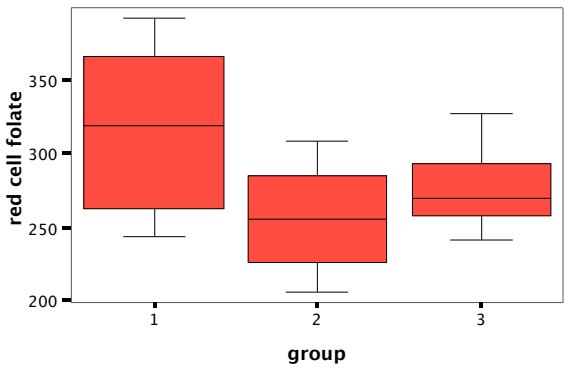
Dependent Variable: red cell folate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	15515.766[a]	2	7757.883	3.711	.044
Intercept	1660859.310	1	1660859.310	794.548	.000
GROUP	15515.766	2	7757.883	3.711	.044
Error	39716.097	19	2090.321		
Total	1820021.000	22			
Corrected Total	55231.864	21			

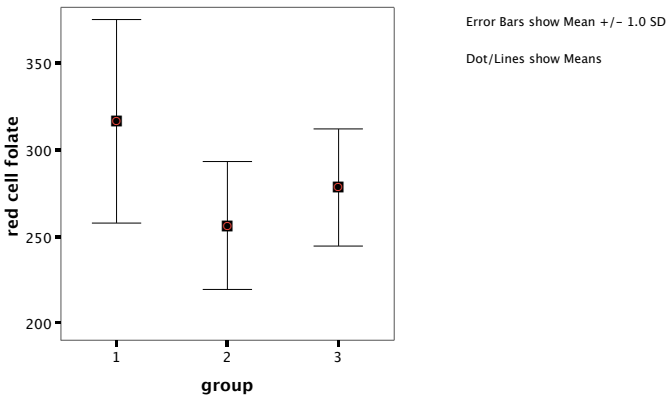
a. R Squared = .281 (Adjusted R Squared = .205)

Die wesentlichste Information ist der p–Wert „Sig.“ ($p = 0.044$) hinter GROUP, der aus der F–Verteilung berechnet wird. Es besagt hier, dass die Nullhypothese der Gleichheit aller Mittelwerte auf dem 5%–Niveau verworfen wird und die Behandlungen unterschiedliche Effekte haben. Der F–Wert — die Teststatistik der Varianzanalyse — wird durch Division der beiden Werte für „Mean square“ von GROUP und Error erhalten. Wie die Quadratsummen und die mittleren Quadrate berechnet werden, kann im Rahmen dieses Skripts nicht behandelt werden.

Als graphische Darstellungen von Daten in einer Varianzanalyse bieten sich gruppierte Box–Plots an. Man erhält so einen Eindruck, ob und wie die Lageparameter verschieden sind, ob Überlappung vorliegt, und zeigt, ob die Daten in den einzelnen Gruppen etwa normalverteilt sind und die gleiche Varianz haben.



Wenn man von der Normalverteilung ausgehen kann — was in der Varianzanalyse vorausgesetzt wird — bietet sich die Darstellung der gruppierten Mittelwerte mit Streuungsbalken an, je nach Verwendungszweck ± 1 Standardabweichung oder ± 1 Standardfehler des Mittelwertes oder 95% Konfidenzintervall des Mittelwertes.



Frage: Kann man die Gruppenunterschiede ebenso gut ohne Varianzanalyse nachweisen ?

Nach dem bisherigen Stoff der Biostatistik–Vorlesung wissen Sie, dass man Zweistichproben–t–Tests zum paarweisen Vergleich der Gruppen verwenden kann:

	Mean Diff.	DF	t-Value	P-Value
1 vs. 2	60.181	15	2.558	0.0218
1 vs. 3	38.625	11	1.327	0.2115
2 vs. 3	-21.556	12	-1.072	0.3046

- Man erhält zunächst einen signifikanten Unterschied zwischen den Gruppen 1 und 2. Es wurden jedoch 3 Hypothesen getestet. Damit stimmt das Gesamtsignifikanzniveau nicht mehr. Bei der Bonferroni–Korrektur wird das nominelle Signifikanzniveau durch 3 geteilt, um das Gesamtsignifikanzniveau von 5% einzuhalten. Jetzt sind keine Differenzen mehr signifikant (alle $p > 0.05/3 = 0.017$).

- Die ANOVA liefert einen p–Wert für die Frage: „Gibt es **überhaupt** einen Unterschied ?“ Bei mehr als 2 Gruppen ist es wissenschaftlich natürlich, diese Frage zuerst zu beantworten.

- Da die Varianzen in allen Gruppen als gleich angenommen werden, werden bei der Varianzanalyse **alle** Beobachtungen zur Schätzung der Varianz herangezogen (pooling). Dadurch erhält man eine bessere Trennschärfe, als wenn man die Gruppen paarwiese vergleichen würde.

Das Zweistichproben–Poblem ist eine ANOVA
Wenn das stimmt, müsste man das Zweistichproben–Problem statt mit dem ungepaarten t–Test auch mit einem Varianzanalyseprogramm lösen können. Um das zu überprüfen, vergleichen wir zuerst die beiden ersten Gruppen im Folsäurebeispiel mittels Zweistichproben–t–Test:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
red cell folate	Equal variances assumed	5.182	.038	2.558	15	.022	60.18	23.525	10.039	110.322
	Equal variances not assumed			2.490	11.579	.029	60.18	24.168	7.310	113.051

Dann geben wir die Daten in eine ANOVA:

Tests of Between-Subjects Effects

Dependent Variable: red cell folate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	15338.962[a]	1	15338.962	6.544	.022
Intercept	1390906.962	1	1390906.962	593.422	.000
GROUP	15338.962	1	15338.962	6.544	.022
Error	35158.097	15	2343.873		
Total	1429043.000	17			
Corrected Total	50497.059	16			

a. R Squared = .304 (Adjusted R Squared = .257)

Die p–Werte der beiden Verfahren stimmen tatsächlich überein und es ist $F = t^2$.

6.1.2 Einfache ANOVA

Mit der einfachen Varianzanalyse können Beobachtungen aus verschiedenen Gruppen verglichen werden. Sie ist die direkte Verallgemeinung des Zweistichproben–t–Tests von 2 auf m Gruppen.
Das statistische **Modell** heisst auch „vollständig randomisiertes Modell“:

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}\,, \quad i = 1, \ldots, m \quad j = 1, \ldots, n_i$$

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

Die Zerlegung $\mu_i = \mu + \alpha_i$ wird durch die Restriktion $\sum_{i=1}^{m} \alpha_i = 0$ eindeutig.

Kernstück der Varianzanalyse ist die **ANOVA–Tabelle**, hier noch einmal am Beispiel der Folsäure mit $m = 3$ Gruppen:

Tests of Between-Subjects Effects

Dependent Variable: red cell folate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	15515.766[a]	2	7757.883	3.711	.044
Intercept	1660859.310	1	1660859.310	794.548	.000
GROUP	15515.766	2	7757.883	3.711	.044
Error	39716.097	19	2090.321		
Total	1820021.000	22			
Corrected Total	55231.864	21			

a. R Squared = .281 (Adjusted R Squared = .205)

- Die ANOVA zerlegt die Varianz der Beobachtungen in die Variabilität der Gruppenmittelwerte um das Gesamtmittel („systematischer Anteil“) und die Variabilität der Beobachtungen **innerhalb** einer Gruppe („zufälliger Anteil“).

- Die Nullhypothese ist H_0: „alle Gruppenmittel sind gleich“.

- Unter der Alternative unterschiedlicher Erwartungswerte ist der systematische Anteil der Variabilität gross gegenüber dem zufälligen Anteil. Der ***F*–Test** bildet den Quotienten beider Anteile und lehnt die Nullhypothese gleicher Gruppenmittel ab, wenn dieser Quotient eine gewisse Schranke übersteigt.

Post–hoc Tests

Die ANOVA beantwortet nur die Frage, ob irgendein Mittelwertsunterschied vorliegt. Man interessiert sich aber oft dafür, welche Mittelwerte sich unterscheiden oder ob sich gewisse interessierende Paare von Mittelwerten unterscheiden. Wir erhalten dann ein multiples Testproblem (siehe Abschnitt 4.5).

- Diese Frage sollte nur untersucht werden, wenn die Varianzanalyse einen p–Wert < 0.05 ergeben hat (d. h. post–hoc).

- Man sollte von vornherein nur wesentliche Differenzen auswählen (je weniger, desto besser).

- Dann soll man modifizierte t–Tests mit der gepoolten Varianzschätzung aus der ANOVA berechnen und die p–Werte mit der Bonferroni–Methode korrigieren (**Bonferroni–Dunn**).

Multiple Comparisons

Dependent Variable: red cell folate

Bonferroni

(I) GROUP	(J) GROUP	Mean Difference (I–J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	60.18*	22.216	.042	1.86	118.50
	3	38.63	26.064	.464	-29.80	107.05
2	1	-60.18*	22.216	.042	-118.50	-1.86
	3	-21.56	25.501	1.000	-88.50	45.39
3	1	-38.63	26.064	.464	-107.05	29.80
	2	21.56	25.501	1.000	-45.39	88.50

Based on observed means.

*. The mean difference is significant at the .05 level.

Anders als im Abschnitt 4.5 besprochen, teilt SPSS das Signifikanzniveau α nicht durch die Anzahl k (hier k=3) durchgeführter Tests, sondern multipliziert die p–Werte mit k. Also ist hier weiterhin $p < 0.05$ signifikant.
Im multiplen Vergleich unterscheiden sich die Gruppen 1 und 2 signifikant, während alle anderen Vergleiche nicht signifikant werden.

Nichtparametrische Varianzanalyse: Kruskal–Wallis Test

Wenn die Daten nicht normalverteilt sind, und auch nicht ohne weiteres auf Normalverteilung transformiert werden können, dann bietet sich eine nichtparametrische oder Rang–Varianzanalyse an.
Der Kruskal–Wallis Test ist eine Verallgemeinerung des Mann–Whitney Tests für den Vergleich von 2 auf m Gruppen.

6.1.3 Zweifache ANOVA

Anstatt nur bezüglich eines (Einfluss–) Faktors (im vorangegangenen Beispiel war das „Behandlungsgruppe“) möchte man Mittelwerte bezüglich zwei Einflussfaktoren vergleichen.

Beispiel: Ausatemdruck bei zystischer Fibrose (O'Neill et al. 1983, *Am. Rev. Respir. Dis.*)

Wir wollen den Ausatemdruck PEmax (Index für Atemmuskelkraft) in Abhängigkeit von den Faktoren BMP (BMI als % des altersspezifischen Medians Gesunder — Faktor A) und Geschlecht (Faktor B) untersuchen.Wenn wir BMP als stetige Variable analysieren, handelt es sich um eine **Kovarianzanalyse**. Wir zerlegen BMP hier aus tutoriellen Gründen in 2 Klassen:

1. Faktor A: BMP eingeteilt in untergewichtig ($< 80\%$ der Norm) und normalgewichtig ($\geq 80\%$ der Norm) als Gewichtsstatus

2. Faktor B: Geschlecht

Modell: „Zweiweg–Kreuzklassifikation“
Das Modell heisst auch „vollständig randomisiertes Block–Modell“.

$$\begin{aligned} y_{ijk} &= \mu_{ij} + \varepsilon_{ijk} \\ i &= 1, \ldots, m_1 && \text{— Stufen von A} \\ j &= 1, \ldots, m_2 && \text{— Stufen von B} \\ k &= 1, \ldots, n_{ij} \geq 0 && \text{— Wiederholungen} \\ \varepsilon_{ijk} &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

Die Kreuzklassifikation bezieht sich darauf, dass alle Stufen von A mit allen Stufen von B kombiniert werden können (über Kreuz). Dabei dürfen Zellen auch unbesetzt sein (d. h. $n_{ij} = 0$). Die vollständige Randomisation bezieht sich auf die Versuchsdurchführung.
Die Analyse beruht auf der folgenden Zerlegung der Populationsmittelwerte:

$$\begin{aligned} \mu_{ij} &= \mu + (\mu_i - \mu) + (\mu_j - \mu) + (\mu_{ij} - \mu_i - \mu_j + \mu) \\ &= \mu + \alpha_i + \beta_j + \gamma_{ij} \\ &= \text{„Gesamtmittel“} + \text{Haupteffekt von A} + \text{Haupteffekt von B} \\ &\quad + \text{Wechselwirkung von A und B} \end{aligned}$$

Das wesentlich **Neue** gegenüber der einfachen Varianzanalyse sind die Wechselwirkungen γ_{ij}. Je nachdem, ob Wechselwirkungen in das Modell aufgenommen werden, unterscheiden wir 2 Modelle:

- Additives Modell: $\mu_{ij} = \mu + \alpha_i + \beta_j$
- Modell mit Wechselwirkungen: $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$

In der Zweiweg–Kreuzklassifikation werden gleichzeitig 3 wissenschaftliche Hypothesen geprüft (hier im Beispiel Ausatemdruck):

1. Gibt es einen Geschlechtsunterschied beim Ausatemdruck (Faktor „sex“)?
2. Unterscheidet sich der Ausatemdruck von leichten und normalgewichtigen Patienten (Faktor „status“)?
3. Ist der Unterschied zwischen leichten und normalgewichtigen Patienten geschlechtsspezifisch (Wechselwirkung „sex * status“)?

So sehen die Daten des Beispiels aus. SPSS akzeptiert als Bezeichnung auch Kategorien (light, normal) anstelle von Stufen $(0, 1)$.

PEmax	BMP	sex	status
95	68	0	light
85	65	1	light
100	64	0	light
85	67	1	light
95	93	0	normal
80	68	1	light
65	89	1	normal
110	69	1	light
70	67	0	light
95	68	1	light
110	89	0	normal
90	90	1	normal
100	93	0	normal
80	93	1	normal
134	66	1	light
134	70	1	light
165	70	0	light
120	92	1	normal
130	69	0	light
85	72	1	light
85	86	0	normal
160	86	0	normal
165	97	0	normal
95	71	0	light
195	95	0	normal

So sieht der Ausdruck der Varianzanalyse in SPSS aus:

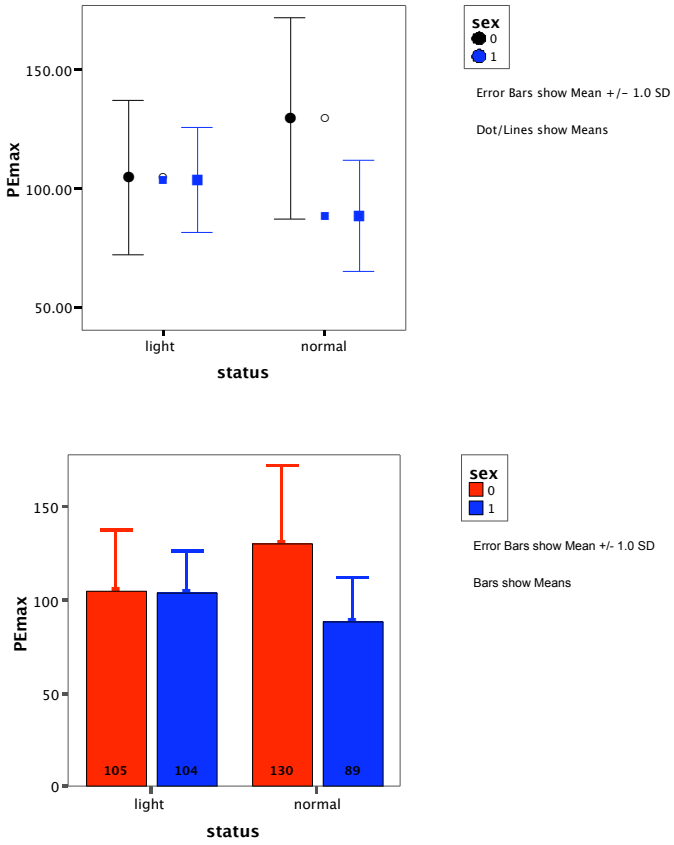
Tests of Between-Subjects Effects

Dependent Variable: PEmax

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5013.890[a]	3	1671.297	1.609	.217
Intercept	269640.092	1	269640.092	259.522	.000
SEX	2630.618	1	2630.618	2.532	.127
STATUS	140.092	1	140.092	.135	.717
SEX * STATUS	2387.461	1	2387.461	2.298	.144
Error	21818.750	21	1038.988		
Total	324512.000	25			
Corrected Total	26832.640	24			

a. R Squared = .187 (Adjusted R Squared = .071)

Wichtig ist hier die letzte Spalte. Die p–Werte für die Effekte „sex“, „status“ und die Wechselwirkung „sex * status“ sind sämtlich grösser als 0.05. Somit lassen sich keine Unterschiede nachweisen.
Die Wechselwirkungen werden im Beispiel mit Vorteil durch ein Punktdiagramm oder ein Balkendiagramm dargestellt:



Die Graphiken zeigen einen Geschlechtsunterschied in PEmax für Normalgewichtige, aber nicht für Untergewichtige. Diese angedeutete Wechselwirkung ist aber nicht signifikant.

6.1.4 ANOVA für wiederholte Messungen

Die Analyse von wiederholten Messungen („repeated measures“) gehört zu den Problemen, bei denen die meisten Fehler gemacht werden. Bei einem Messwiederholungs–Experiment wird typischerweise eine Grösse — z. B. der Blutdruck — wiederholt an denselben Patienten gemessen. Es ist dies eine Verallgemeinerung der Situation, die wir beim **gepaarten** t–Test angetroffen haben, bei dem Patienten zu zwei Zeitpunkten untersucht wurden, z. B. vor und nach Medikation. Es gibt aber auch andere Typen von Messwiederholungen, wie Messungen an verschiedenen Stellen des Kopfes beim EEG oder die Verabreichung unterschiedlicher Medikamente an denselben Patienten. Ziel der Analyse eines Messwiederholungsfaktors („repeated factor“) ist der Nachweis einer Veränderung z. B. im Verlauf. Die Nullhypothese ist demnach, dass es keine Veränderung gibt. Im Gegensatz zu den Messdaten verschiedener Patienten sind wiederholte Messungen innerhalb eines Patienten korreliert.

Merke: Die ANOVA für wiederholte Messungen ist eine Verallgemeinerung des gepaarten t–Tests.

Beispiel: Maskin et al. (1985) (Altman, S. 327) untersuchten die Kurzzeitwirkung eines Medikamentes auf die Herzrate bei 9 Herzpatienten.

	Time (mins)			
Subject	0	30	60	120
1	96	92	86	92
2	110	106	108	114
3	89	86	85	83
4	95	78	78	83
5	128	124	118	118
6	100	98	100	94
7	72	68	67	71
8	79	75	74	74
9	100	106	104	102

Modell: „repeated measures ANOVA“

$$
\begin{aligned}
y_{ij} &= \mu + \alpha_i + b_j(t_i) + \varepsilon_{ij} \\
t_i &\text{ — Zeitpunkte, Messstellen, } i = 1, \ldots, m \\
j &= 1, \ldots, J \text{ — Personen} \\
b_j(t_i) &\text{ — individueller (zufälliger) Effekt der Person } j \text{ zum Zeitpunkt } t_i \\
\varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

Wir gehen also davon aus, dass es einen mittleren Verlauf (oder Pattern) $\mu + \alpha_i$ gibt, von dem die individuellen Verläufe um $b_j(t_i)$ abweichen. Im Beispiel ist zu testen, ob wir im Verlauf Konstanz haben ($\alpha_i = 0$) oder eine Veränderung stattfindet (ein $\alpha_i \neq 0$) .
Die verschiedenen korrekten Analysemethoden unterscheiden sich durch die Voraussetzung an diese individuellen Abweichungen. Ganz falsch wäre es hingegen, eine normale einfache ANOVA durchzuführen.

1. multivariates Einweg–Modell (MANOVA): Hier nimmt man nur an, dass die individuellen Abweichungen irgendwelche Kurven sind, die für alle Personen die gleiche multivariate Normalverteilung (mit Erwartungswert 0) haben, ansonsten aber beliebig sind. (unstrukturiert)

2. univariates Modell der Varianzanalyse mit Messwiederholungen: Hier nimmt man an, dass die Korrelationen zwischen den Messwerten zu verschiedenen Zeitpunkten immer gleich sind, egal wie weit die Zeitpunkte voneinander entfernt sind. (compound symmetry)

Die einschränkende Voraussetzung der „compound symmetry“ ist bei mehr als 2 (Zeit–, Mess–) Punkten selten erfüllt.

Ausweg: Es gibt eine sogenannte **Greenhouse–Geisser Korrektur** für Abweichungen von der „compound symmetry“, die die univariate Methode approximativ gültig macht.

So sieht der Ausdruck der Varianzanalyse in SPSS aus:

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Sphericity Assumed	150.972	3	50.324	4.070	.018
	Greenhouse-Geisser	150.972	2.120	71.226	4.070	.034
	Huynh-Feldt	150.972	2.904	51.988	4.070	.019
	Lower-bound	150.972	1.000	150.972	4.070	.078
Error(TIME)	Sphericity Assumed	296.778	24	12.366		
	Greenhouse-Geisser	296.778	16.957	17.502		
	Huynh-Feldt	296.778	23.232	12.775		
	Lower-bound	296.778	8.000	37.097		

Es gibt also einen signifikanten Trend: der Wert ist $p = 0.034$, und zwar nach Greenhouse–Geisser Korrektur.

Im Beispiel haben wir nur einen Messwiederholungsfaktor gehabt. Oft hat man es in der Praxis aber mit komplexen Experimenten zu tun, wo die Verläufe in verschiedenen Gruppen verglichen werden. Verlauf ist dann ein Messwiederholungsfaktor, Gruppe aber nicht.

6.2 Logistische Regression

Die logistische Regression hat als Analysemethode in der medizinischen Forschung zunehmend an Bedeutung gewonnen. Sie wird deshalb hier behandelt, wenn auch nicht umfassend und ohne den relativ aufwendigen mathematischen Hintergrund. Wie bei der „gewöhnlichen" Regression (Kapitel 5) wird versucht, eine Ziel– („outcome") Variable y möglichst gut durch andere Variablen $x_1, \ldots, x_k$ zu erklären.

Neu an der logistischen Regression ist, dass der **„Outcome"** y nicht als kontinuierlich, sondern **als binär** angenommen wird.

Beispiele:

A. Therapiestudie
   y = Patienten überleben ($y = 0$) oder sterben ($y = 1$), x_1 = Therapieform ($x_1 = A, B$; nominal), x_2 = Alter (in Jahren; stetig), $x_3 \ldots$ = Laborparameter und andere erklärende Variable

B. Fall–Kontroll–Studie (Epidemiologie)
   y = Fall ($y = 1$) oder Kontrolle ($y = 0$), x_1 = exponiert ($x_1 = 1$) oder nicht ($x_1 = 0$), x_2 = Confounder (ein Confounder ist eine Variable, die das Risiko „Fall" zu werden beeinflussen kann, aber nicht zur Exposition gehört).

Für univariate Analysen mit nur einer erklärenden Variablen x ist die logistische Regression nicht unbedingt notwendig. Bei einer stetigen erklärenden Variablen vergleicht man üblicherweise die x–Werte der beiden Gruppen mit $y = 0$ und $y = 1$ mittels Mann–Whitney Test (oder bei normalverteilten Daten mittels ungepaartem t–Test). Bei einer kategoriellen erklärenden Variablen nimmt man Fisher's exakten Test (oder den χ^2–Test). Signifikante Ergebnisse bedeuten jeweils einen signifikanten Zusammenhang zwischen x und y.

Beispiel: In diesem Kapitel wollen wir als Beispiel Risikofaktoren für den Befall von benachbarten Lymphknoten beim Prostata–Krebs untersuchen (Brown, 1980). Folgende Variablen wurden bei $n = 52$ Patienten erhoben:
y = Befall von Knoten (0 = Keiner, 1 = Befall)
x = Alter, Phosphatase, Röntgenbefund, Tumorgrösse, Tumorstadium.
Die ersten beiden x–Variablen sind stetig, die restlichen binär.
Die Kontingenztafel für den Zusammenhang von Befall von Knoten und Röntgenbefund sieht so aus:

	Röntgenbefund		
	$x = 0$	$x = 1$	
kein Befall ($y = 0$)	28	4	32
Befall ($y = 1$)	9	11	20
	37	15	52

Die Vorhersagekraft des Röntgenbefundes wird mittels **Sensitivität** = 11/20 = 55% und **Spezifität** = 28/32 = 87% bewertet. Sensitivität und Spezifität werden jeweils in den

Gruppen „kranker“ und „gesunder“ Patienten berechnet. Der Zusammenhang ist signifikant (Fisher's exakter Test $p = 0.002$).
Aus der Epidemiologie stammen zwei Masse für den Zusammenhang der binären Variablen x und y: das **relative Risiko** (RR) und der **Odds ratio** (OR).

„Risiko“ ist definiert als

$$\mathrm{P}(y = 1|x) = p(x)\,,$$

also $p(0) = 9/37 = 24\%$ und $p(1) = 11/15 = 73\%$. Das relative Risiko ist dann

$$p(1)/p(0) = \frac{11 \times 37}{15 \times 9} = 3.0\,,$$

das heisst, das Risiko für den Befall eines Knotens ist bei Patienten mit positivem Röntgenbefund gegenüber denen mit negativem Befund dreifach erhöht. Das Risiko und auch das relative Risiko sind nur dann gültig, wenn eine repräsentative Stichprobe aus der Grundgesamtheit der Patienten mit Prostatakrebs gezogen wurde. Bei Fall–Kontroll Studien ist das nicht der Fall, da hier das Verhältnis der Zahlen von Patienten mit und ohne Befall willkürlich gewählt wird.

Aus der Wettsprache kennen wir die „Odds“ (Chance, Kurs) definiert als

$$\frac{\mathrm{P}(y = 1|x)}{\mathrm{P}(y = 0|x)} = \frac{p(x)}{1 - p(x)}$$

In der Epidemiologie wird der **„Odds ratio“** als Mass für relatives Risiko benutzt:

	$x = 0$	$x = 1$
$y = 0$	28	4
$y = 1$	9	11

$$\mathrm{OR} = \frac{\mathrm{P}(y = 1|x = 1)}{1 - \mathrm{P}(y = 1|x = 1)} \Big/ \frac{\mathrm{P}(y = 1|x = 0)}{1 - \mathrm{P}(y = 1|x = 0)} = \frac{28 \times 11}{9 \times 4} = 8.6$$

Der odds ratio für positive Röntgenbefunde gegenüber negativen Röntgenbefunden ist 8.6, d. h., der odds („Risiko“) für den Befall von Knoten ist bei positivem Befund um den Faktor 8.6 höher als bei negativem Befund.
Das Verhältnis der Patientenzahlen mit und ohne Befall kürzt sich bei der Berechnung des odds ratios heraus, so dass er auch bei Case–Control Studien gültig ist.
Falls eine Krankheit selten ist, sind der OR und das relative Risiko annähernd gleich:

$$\mathrm{OR} = \frac{p(1)}{1 - p(1)} \Big/ \frac{p(0)}{1 - p(0)} \approx \frac{p(1)}{p(0)}$$

Dadurch sind Odds ratios und logistische Regression sehr populär in der Epidemiologie. In unserem Beispiel sieht man aber, dass sie quantitativ auch sehr verschieden sein können.

6.2.1 Modellierung mittels logistischer Regression

Bei der Verallgemeinerung von der gewöhnlichen linearen Regression auf die logistische Regression fragen wir uns: Was ist fundamental an einer einfachen Regression mit einer x–Variablen? Das allgemeine Modell ist:

$$y_i = f(x_i, \beta) + \varepsilon_i, \quad i = 1, \ldots, n$$

wo: f = vorher spezifizierte Funktion, z. B. linear $f(x_i, \beta_0, \beta_1) = \beta_0 + \beta_1 x_i$

Die Regressionsfunktion $f(x, \beta)$ ist mathematisch der bedingte Erwartungswert von y bei gegebenem Wert von x, d. h.

$$\mathrm{E}(y \mid x) = f(x, \beta)$$

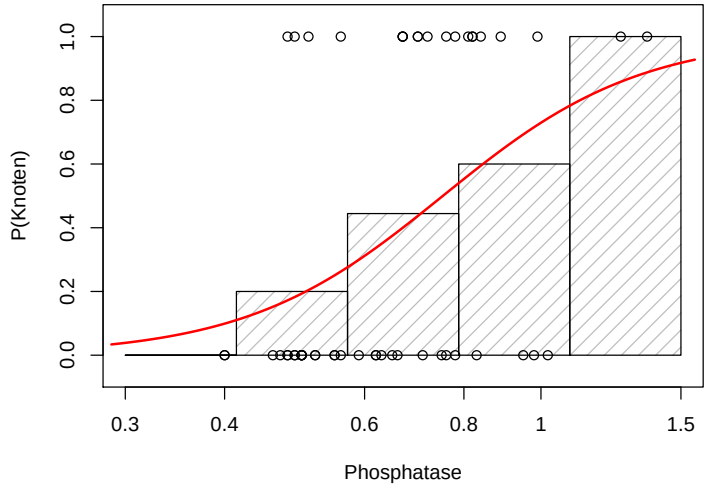
Mathematische Überlegungen zeigen, dass in der logistischen Regression — d. h. bei binärem y — eine Wahrscheinlichkeit $p(x)$ modelliert werden muss.

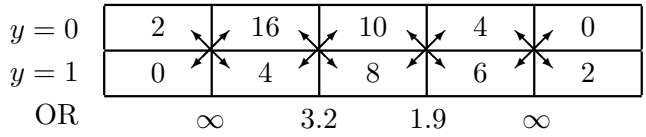
- Der Outcome ist eine binäre Variable: Ereignis ($y = 1$), kein Ereignis ($y = 0$).
- Die Ereigniswahrscheinlichkeit sei $p = p(x) = \mathrm{P}(y = 1|x)$.
  Also: $\mathrm{E}(y|x) = 0 \times \mathrm{P}(y = 0|x) + 1 \times \mathrm{P}(y = 1|x) = p(x)$

Weshalb funktioniert die gewöhnliche Regressionsrechnung nicht mehr bei binärem y?

Es wird eine Wahrscheinlichkeit modelliert, die bekanntlich zwischen 0 und 1 liegt. Also liegt man mit einer linearen Beziehung falsch.

Beispiel: Befall von Knoten beim Prostata–Krebs
Die nachstehende Abbildung zeigt ein (x,y) Streudiagramm mit x = Phosphatase, auf einer logarithmischen Skala. Intuitiv erscheint eine lineare Regression nicht angemessen.
Wenn man die Häufigkeit für „Befall von Knoten“ ($y = 1$) in einem Balkendiagramm aufträgt, sieht man einen Trend: Befallene Knoten mit höheren Phosphatasewerten treten häufiger auf.





Auch hier kann man wieder odds ratios für benachbarte Säulen berechnen; z. B. ist der OR für Phosphatase in [0.58–0.79] vs. [0.41–0.57] $= \dfrac{16 \times 8}{10 \times 4} = 3.2\,.$
Bei Änderung um mehrere Klassen ist der OR **multiplikativ**, wie die obere Tabelle verdeutlicht.
Aufgrund des Bildes liegt es nahe, den bedingten Erwartungswert $p(x)$ durch eine Verteilungsfunktion zu modellieren.
Annahme: Der wahre odds ratio sei für benachbarte Klassen konstant (ähnlich zur Annahme des konstanten Anstiegs der Regressionsfunktion in der linearen Regression).
Aufgrund der Multiplikativität des OR und diesen Annahmen kann man mathematisch zeigen, dass der Logarithmus des odds („logit“) linear in x ist.

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

Folglich ist $p(x)$ die logistische Verteilungsfunktion (daher der Name):

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Die Funktion $p(x)$ modelliert die Wahrscheinlichkeit, dass $y = 1$ ist (im Beispiel: dass ein Befall auftritt) in Abhängigkeit von x (im Beispiel: der gemessenen Phosphatase).

Die **logit Transformation** $\log(p/(1-p))$ führt auf eine lineare Beziehung zwischen x–Variable und „outcome“ y:

$$\boxed{g(x) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x}$$

In der vorangegangenen Abbildung sehen wir, dass die logistische Verteilungsfunktion (durchgezogene Linie) den Trend ganz gut beschreibt.

6.2.2 Schätzen und Testen in der logistischen Regression

Wie früher, bei der gewöhnlichen Regression, stellen sich die folgenden Probleme:

A. Wie bestimmt (schätzt) man β_0, β_1 am besten?

B. Wie prüft (testet) man, ob der Einfluss von x auf $\mathrm{P}(y = 1|x)$ signifikant ist? Die naheliegende Nullhypothese ist $\mathrm{H}_0 :\ \beta_1 = 0$, und die zweiseitige Alternativhypothese ist $\beta_1 \neq 0$.

Die Methode der kleinsten Quadrate ist inadäquat, und die gebräuchliche Methode zum Schätzen ist die Maximum–Likelihood–Methode. Dasselbe Prinzip liefert auch statistische Tests und Konfidenzintervalle, insbesondere für den Koeffizienten β_1.

Der gebräuchliche Test für einen einzelnen Prädiktor x ist der Wald–Test (nach dem bedeutenden Statistiker Abraham Wald):

Wald–Teststatistik	$W = \dfrac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)}$

Der p–Wert basiert dann auf der approximativen Standardnormalverteilung von W.
Aufgrund der nachfolgenden Tabelle erhält man für $x = \log_2(\text{Phosphatase})$ den Schätzwert $\hat{\beta}_1 = 2.42$ mit $\widehat{\mathrm{SE}}(\hat{\beta}_1) = 0.88$. Damit ergibt sich (in der Spalte „Coef/SE“):

$$W = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} = \frac{2.42}{0.88} = 2.76$$

Die Normalverteilungsapproximation ergibt $\mathrm{P}(|Z| > 2.76) = 0.006$, d. h. die Phosphatase beeinflusst den Befall von Knoten signifikant und zwar klinisch negativ.

	Coef	Std. Error	Coef/SE	Chi–Square	P–Value	Exp(Coef)
constant	.99	.60	1.64	2.71	.1000	2.70
$\log_2$(phosph.)	2.42	.88	2.76	7.60	.0058	11.26

6.2.3 Interpretation der Koeffizienten

Die Interpretation der Koeffizienten ist anders als bei der gewöhnlichen Regression: Wenn sich dort x um eine Einheit ändert, so ändert sich y um β_1 Einheiten. Die Beziehung zwischen $p(x)$ und x ist aber linear in logits:

$$g(x) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

Also: Bei einer Änderung von x um eine Einheit erfolgt eine Änderung in $p(x)$ um β_1 logits. Was bedeutet dies?

Binäre x–Variable

Die beiden Ausprägungen der erklärenden Variablen x werden intern mit $x = 0$ und $x = 1$ codiert. Die logistische Regression erlaubt die Bestimmung des Odds ratio, weil die folgenden Bestimmungen gelten:

$$\begin{aligned} \mathrm{OR} &= \exp(\beta_1) \\ \log(\mathrm{OR}) &= \beta_1 = g(1) - g(0) \end{aligned}$$

Der OR ist ein Masss dafür, wie stark die Wahrscheinlichkeit krank zu werden durch eine Exposition erhöht ist. Ein $\beta_1 > 0$ und damit OR > 1 bedeutet ein erhöhtes, $\beta_1 < 0$ und damit OR < 1 ein verringertes Risiko. Wenn die beiden Ausprägungen der erklärenden Variablen x vertauscht werden, dreht sich das Vorzeichen von β_1 um, und

$$\mathrm{OR}_{B \text{ vs. } A} = \exp(-\beta_1) = 1/\mathrm{OR}_{A \text{ vs. } B}$$

Also nicht der Schätzwert $\hat{\beta}_1$, sondern $\exp(\hat{\beta}_1)$ als OR werden in Publikationen veröffentlicht und als erhöhtes oder erniedrigtes Risiko interpretiert.

Kontinuierliche x–Variable

Der logit g ist linear in x:

$$g(x) = \beta_0 + \beta_1 x$$

Wenn x sich um eine Einheit ändert, ändert sich logit = log(odds) um β_1 Einheiten. Der Odds ratio OR $= \exp(\beta_1)$ ist also ein quantitatives Mass für die Risikoerhöhung bei Änderung von x um eine Einheit. Der Odds ratio für eine Änderung von x um k Einheiten ist

$$\exp(k\,\beta_1) = (\exp(\beta_1))^k = \mathrm{OR}^k\,.$$

Der OR ist also multiplikativ.

Beispiel: Der Odds ratio für die Erhöhung der Wahrscheinlichkeit des Auftretens von Knoten bei Erhöhung der $\log_2$(Phosphatase) um eine Einheit, d. h. Erhöhung der Phosphatase um den Faktor 2 ist OR $= \exp(2.42) = 11.3$. Bei einer Erhöhung der Phosphatase um 50% erhöht sich $\log_2$(Phosphatase) um $\log_2(1.5)$ Einheiten, also ist der entsprechende OR $= 11.3^{\log_2(1.5)} = 4.1$.

6.2.4 Multiple logistische Regression

Wenn $k > 1$ erklärende Variablen $x_1, \ldots, x_k$ vorhanden sind, bietet sich eine multiple logistische Regression an. Die Gründe sind die gleichen wie für eine multiple lineare Regression:

1. Man möchte mögliche Effekte von „Stör“–Variablen in einer Studie mit einer Einflussgrösse eliminieren.

2. Man möchte mögliche Prognosefaktoren erforschen, von denen wir nicht wissen, ob sie wichtig oder redundant sind.

3. Man möchte eine Formel zur besseren Vorhersage des individuellen Risikos aus den erklärenden Variablen entwickeln.

Das multiple Problem wird analog mit dem Maximum–Likelihood–Prinzip gelöst. Die $k+1$ nicht–linearen Gleichungen führen allerdings manchmal zu numerischen Problemen. Als Faustregel gilt, dass man mindestens 20 Ereignisse und 20 Nichtereignisse pro erklärender Variabler haben sollte. Unter Umständen können aber auch schon 10 Ereignisse tolerierbar sein.
Die nachstehende Tabelle gibt die Resultate einer univariaten Analyse für das Prostata–Krebs–Beispiel. Dabei wird jede x–Variable vorerst als einzelner Prädiktor analysiert.

Variable	$\hat{\beta}$	$\widehat{\mathrm{SE}}(\hat{\beta})$	W	p–Wert
$\log_2$(Phosphatase)	2.42	0.88	2.76	0.006
Alter	−0.05	0.05	−0.96	0.337
Röntgen	2.15	0.70	3.07	0.002
Grösse	1.61	0.63	2.55	0.011
Stadium	1.14	0.60	1.91	0.056

Offensichtlich sind Phosphatase, Röntgenbefund und Tumorgrösse signifikant, während das Alter deutlich nicht signifikant ist. Das Tumorstadium ist ein Grenzfall. Deshalb wird das Alter weggelassen, und für den Rest wird eine multiple logistische Regression gerechnet. Dann ist das Stadium nichtsignifikant und wird aus der Analyse entfernt (dies ist auch sinnvoll wegen obiger Faustregel über das maximale k).

Variable	$\hat{\beta}$	$\widehat{\mathrm{SE}}(\hat{\beta})$	W	p–Wert
(Intercept)	−0.54	0.83	−0.65	0.515
$\log_2$(Phosphatase)	2.37	1.03	2.31	0.021
Röntgen	1.97	0.82	2.40	0.016
Grösse	1.62	0.75	2.15	0.032

Interpretation:

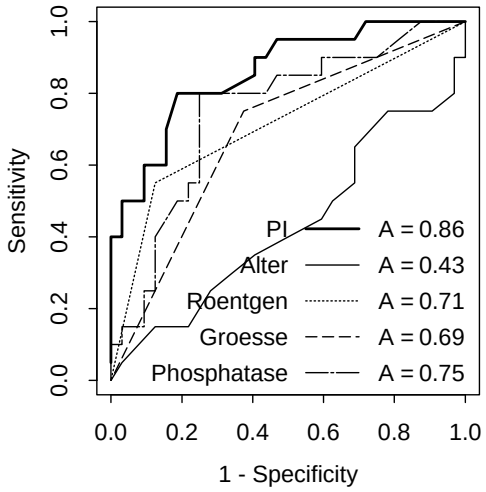
- $\hat{\beta}_i$ — Einfluss von x_i bei festgehaltenen übrigen Variablen.
- p–Werte — Liefert x_i bei bekannten übrigen Variablen eine zusätzliche Information für $\mathrm{P}(y = 1)$? In diesem Beispiel sind Phosphatase, Röntgenbefund und Tumorgrösse signifikant. Sie werden deshalb auch als unabhängige Risikofaktoren bezeichnet. Man muss aber beachten, dass wir nur 52 Patienten mit 20 Ereignissen haben, so dass diese Ergebnisse mit der gegebenen Vorsicht interpretiert werden müssen.
- $\exp(\hat{\beta}_i)$ — Odds ratio bei festgehaltenen übrigen Variablen, d. h., OR eines Patienten mit Röntgen = 1, Grösse = 1, dafür um den Faktor 2 verringerter Phosphatase gegenüber einem Patienten mit Röntgen = Grösse = 0 :
  $\mathrm{OR} = 5.0 \times 7.2/10.7 = 3.4$
- Wie kann man die Information von mehreren signifikanten Einflussfaktoren kombinieren ?

  $$\mathrm{PI} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

  ist ein prognostischer Index (Score). Wenn PI gross ist (> Cut-Punkt), sagen wir „$y = 1$“ voraus.

Güte der Vorhersage
Eine graphische Methode zur Beurteilung der logistischen Regression ist die ROC– (receiver operating characteristic) Kurve. Dabei wird für sämtliche Cut–Punkte c des prognostischen Indexes PI untersucht, wie gut sich die binäre Variable $\{\text{PI} > c\}$ zur Vorhersage des Outcomes y eignet. Dazu wird die Sensitivität gegen $1 -$ Spezifität abgetragen. Die Fläche unter der ROC–Kurve ist ein Mass für die Güte der Vorhersage; eine Fläche von 0.5 entspricht absolutem Unwissen.

Beispiel: Auftreten von Knoten beim Prostata–Krebs.



Der prognostische Index ist

$$\text{PI} = 2.4 \times \log_2(\text{Phosphatase}) + 2.0 \times \text{Röntgen} + 1.6 \times \text{Grösse}$$

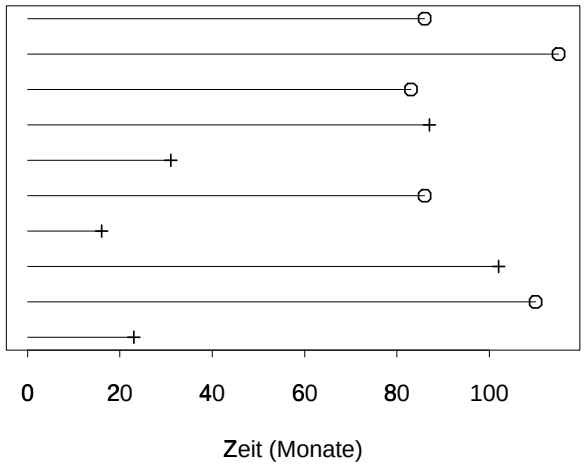
Die Fläche unter der ROC–Kurve ist 0.86 verglichen mit $A = 0.75$ für den besten univariaten Prädiktor. Das bedeutet, dass die Diagnose mittels mehrerer Variabler sicherer gestellt werden kann. Die Fläche unter der ROC–Kurve für Alter ist 0.43; das Alter des Patienten ist also für die Vorhersage von Knoten absolut unbrauchbar.

6.3 Survivalanalyse

6.3.1 Zensierte Beobachtungen

Bei der Survivalanalyse bestehen die zentralen Daten aus einer Zeitdauer. Dabei steht in der Medizin die Zeit von Therapiebeginn (oder Operation) bis zum Tod im Vordergrund, die Überlebenszeit. Es können aber auch andere Ereignisse am Schluss stehen, wie z. B. Wiederaufnahme in die Klinik oder Rezidiv (allgemein ist die Variable „time to event“). Ziel ist, Überlebenszeiten statistisch zu beschreiben und für verschiedene Gruppen zu vergleichen (Beispiel: längere Überlebenszeit = bessere Therapie). Oft ist es auch wichtig, Zusammenhänge zwischen der Überlebenszeit und erklärenden Variablen wie Alter, Art der Therapie oder Schwere der Krankheit (Risikofaktoren, x–Variablen) herzustellen.

Ein wesentlicher Unterschied zwischen der Survivalanalyse und den bisher behandelten Methoden ist die **Zensierung** der Daten („incomplete follow up“). Die natürliche Zensierung entsteht dadurch, dass die Studie beendet wird, bevor alle Patienten gestorben sind („withdrawn alive“). Um einen vollständigen Überblick über die Überlebenszeiten nach einer perinatalen Behandlung zu erhalten, müssten wir sonst eventuell viele Jahre warten. Eine weitere Zensierung entsteht dadurch, dass Patienten unabhängig vom Studiengegenstand ausscheiden („lost to follow up“, Beispiel: Emigration, Unfalltod). Als Ergebnis dieser Versuchsdurchführung erhalten wir **zensierte Überlebenszeiten**, d. h. die Überlebenszeit ist für einen Teil der Patienten bekannt, für andere wissen wir nur, dass sie bis zu einem gewissen Zeitpunkt gelebt haben. Die Patienten treten oft zu unterschiedlichen Zeiten in die Studie ein (z. B. Zeitpunkt der Operation), was in der nachfolgenden Graphik durch den Nullpunkt symbolisiert wird.



Es handelt sich um eine Teilstichprobe von 10 Patienten einer Melanomstudie, die nachstehend besprochen wird. „+“ bedeutet, dass das Ereignis (hier Tod) nach dieser Zeitspanne eingetreten ist, „∘“ bedeutet, dass die Beobachtung nach dieser Zeitspanne zensiert ist, d. h. der Patient lebte zu diesem Zeitpunkt noch, danach wurde die Studie beendet (withdrawn alive), oder der Patient konnte aus irgend einem Grund nicht länger beobachtet werden (lost to follow–up). Die Daten dieser Teilstichprobe sehen in Excel wie folgt aus:

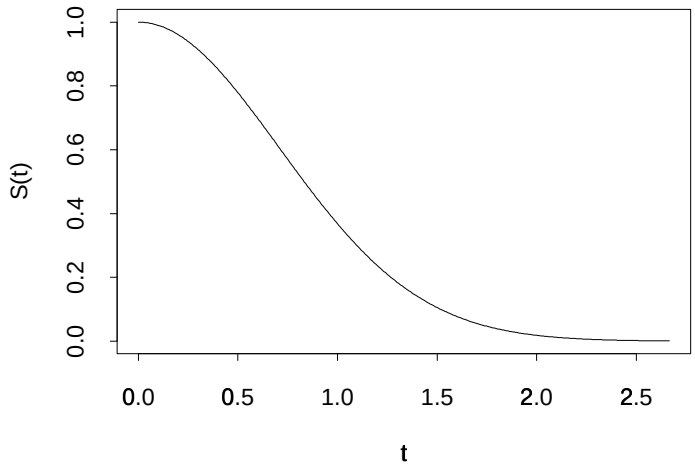
	A	B	C	D	E
1	nr	sex	Breslow	zeit-total	status
2	2	1	2	87	1
3	3	2	2.13	102	1
4	4	1	2.07	23	1
5	6	2	2.23	16	1
6	7	1	1.87	31	1
7	22	1	3.4	110	0
8	23	1	2.68	83	0
9	24	1	0.46	115	0
10	28	1	1.11	86	0
11	30	2	0.64	86	0

Die Variable „status“ beschreibt, ob der Patient gestorben ist (status = 1) oder am Ende lebte (status = 0). Die Variable „zeit–total“ ist die Zeitspanne von der Diagnose bis zum Tod oder der Zensierung.

6.3.2 Überlebensfunktion und Kaplan–Meier Schätzer

Sei $F(t)$ die Verteilungsfunktion (siehe Kapitel 3) der Überlebenszeiten. Dann ist die **Überlebensfunktion** $S(t)$ definiert als

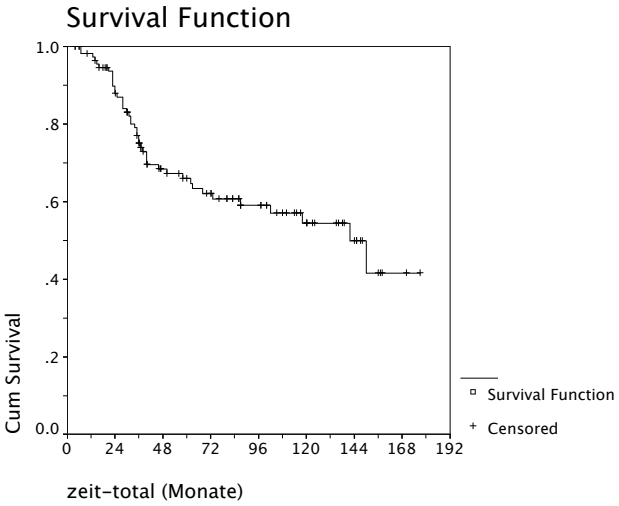
$$S(t) = 1 - F(t)$$



Sie gibt also den Prozentsatz der Patienten an, die nach einer gewissen Zeit t überlebt haben. Die Frage ist nun, wie wir die Überlebensfunktion $S(t)$ aus den Daten schätzen. Wir wollen als Beispiel die Überlebenszeiten von $n = 116$ Melanompatienten im Stadium 1 analysieren. Diese Studie wurde am Universitäts–Spital Zürich durchgeführt, um Risikofaktoren für die Mortalität zu untersuchen, siehe unten. Die folgende Tabelle ist nach Geschlechtern getrennt, „+“ bedeutet gestorben, und „∘“ bedeutet eine zensierte Überlebenszeit:

Geschlecht	Beobachtungszeit (Monate)
männlich	4○, 6○, 7+, 13+, 14○, 16+, 16○, 20○, 21+, 23+, 24+, 25+, 28+, 28+, 30+, 30○, 32+, 34+, 36+, 36○, 36○, 37+, 37○, 38+, 38○, 40+, 40+, 40○, 40+, 47○, 50+, 58+, 62+, 68+, 73+, 80○, 83○, 86○, 87○, 87○, 97○, 102+, 114○, 117○, 123○, 139○, 147○, 170○
weiblich	4○, 6○, 7+, 10○, 14+, 15+, 18○, 19○, 20○, 23+, 23+, 23+, 24+, 24○, 28+, 30○, 30○, 31+, 32+, 35+, 35+, 35○, 36○, 36+, 37○, 40○, 46○, 46+, 47○, 50○, 56○, 58○, 60○, 63+, 70○, 72○, 72○, 76○, 80○, 83○, 86○, 87+, 87○, 97○, 100○, 100○, 105○, 108○, 110○, 115○, 118+, 120○, 120○, 120○, 124○, 135○, 136○, 138○, 142+, 144○, 145○, 145○, 148○, 150+, 156○, 157○, 158○, 177○

In der folgenden Graphik von SPSS sehen Sie die Schätzung der Überlebensfunktion $S(t)$ durch den **Kaplan–Meier–Schätzer**, der die Methode der Wahl ist.

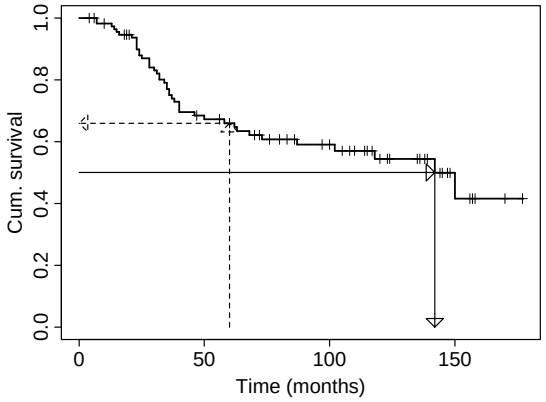


Jede Treppenstufe „¬“ bedeutet ein Ereignis (Todesfall), Striche „+“ zeigen die zensierten Fälle.
Der Kaplan–Meier–Schätzer ist eine Treppenfunktion mit Stufen bei jedem Todesfall. Die Stufenhöhe wird nicht nur durch die Anzahl Todesfälle zu einem Zeitpunkt bestimmt, sondern auch dadurch, wieviele Patienten zu diesem Zeitpunkt noch leben („patients at risk“).

Neben Kaplan–Meier–Schätzungen sind auch aktuarische „Life Table“ Schätzungen in Gebrauch. Wir empfehlen aber den Kaplan–Meier–Schätzer.

6.3.3 Beschreibung von Überlebenszeiten

Bei der Beschreibung von zensierten Daten werden häufig Fehler gemacht. Zunächst ist es sinnvoll, die mittlere und Median–follow–up Zeit anzugeben. Dazu gibt man die Anzahl der Patienten unter Risiko für einige wenige Zeitpunkte, z. B. nach 1, 5 und 10 Jahren an. Man erhält so einen Eindruck von der Qualität der Studie.



Median der Überlebenszeit

Der Median der Überlebenszeit als charakteristische Kennzahl ist wie folgt definiert:

50% der Überlebenszeiten sind länger als der Median
50% der Überlebenszeiten sind kürzer als der Median

Der Median ist aus der Kaplan–Meier–Kurve ablesbar, indem man vom 50%–Punkt auf der y–Achse horizontal nach rechts geht und den Schnittpunkt bestimmt (falls möglich — durchgezogener Pfeil in der obigen Graphik).

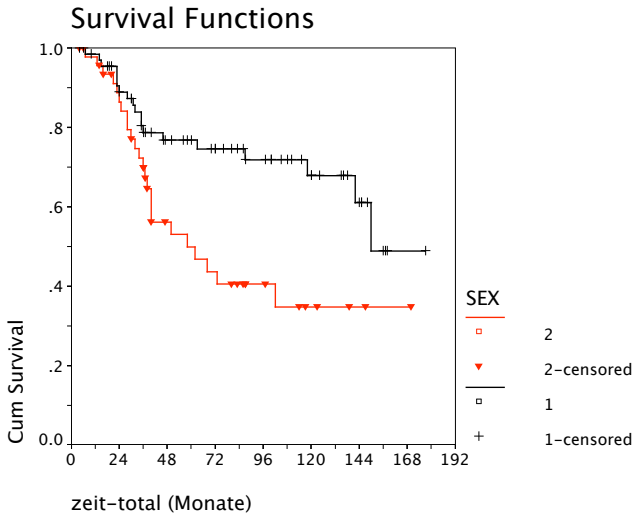
Wert der Kaplan–Meier Kurve

Es ist sinnvoll, einige wenige Werte der Kaplan–Meier Kurve anzugeben, z. B. nach 1, 5 (gestrichelter Pfeil) und 10 Jahren, natürlich nur, wenn noch genügend Patienten unter Risiko sind. Häufig wird auch der Standardfehler angegeben, also $\hat{S}(t) \pm \mathrm{SE}\left(\hat{S}(t)\right)$.

Bloss nicht !

Die durchschnittliche Überlebenszeit ist die Fläche unter der Überlebensfunktion. Wenn die letzte Beobachtung zensiert ist, erreicht die Kurve die Null nicht, und die Schätzung der mittleren Überlebenszeit ist verfälscht. Aber auch, wenn der letzte Patient gestorben ist, hängt der Wert stark von wenigen extremen Überlebenszeiten ab, und sollte deshalb nicht verwendet werden.
Generell falsch ist es, einfache deskriptive Statistiken von zensierten Daten zu präsentieren. Häufig findet man den Anteil (%) oder die mittlere Überlebenszeit der Gestorbenen. Diese Zahlen machen aber keinen Sinn. Falsch sind natürlich auch die statistischen Tests, die auf diesen Statistiken basieren.

Vergleich von Überlebensfunktionen
Zum Vergleich der Überlebensfunktionen von verschiedenen Patientengruppen können die Schätzer für beide Gruppen getrennt berechnet werden. Im Beispiel wird die Überlebenszeit von weiblichen (sex = 1) und männlichen Melanompatienten (sex = 2) verglichen.



Der Test der Nullhypothese

$$H_0: \text{ die Überlebensfunktionen der beiden Gruppen sind gleich}$$

heisst **Logrank–Test**. Das Ergebnis für das Beispiel (in SPSS) ergibt einen signifikanten Geschlechtsunterschied.

Test Statistics for Equality of Survival Distributions for SEX

	Statistic	df	Significance
Log Rank	7.79	1	.0052

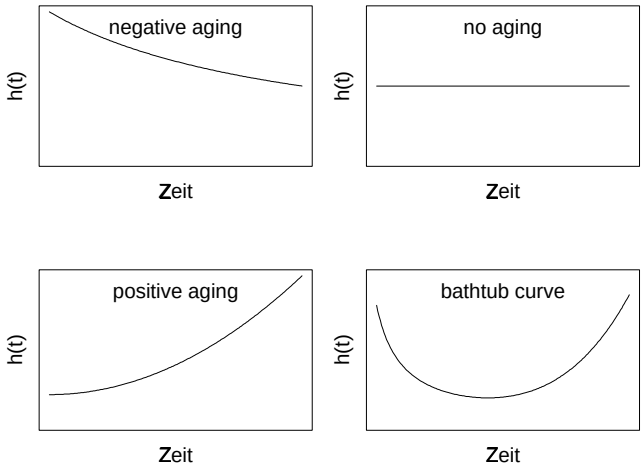
6.3.4 Cox–Regression

Die Hazardfunktion
Noch wichtiger als die Überlebensfunktion ist die Hazardfunktion, um die Wahrscheinlichkeit des Versterbens (bzw. des Überlebens) zu analysieren. Grob gesprochen gibt die Hazardfunktion („hazard rate“) $h(t)$ die Wahrscheinlichkeit an, im nächsten Moment zu sterben, wenn man die Zeit t erreicht hat. Formal ist sie definiert durch:

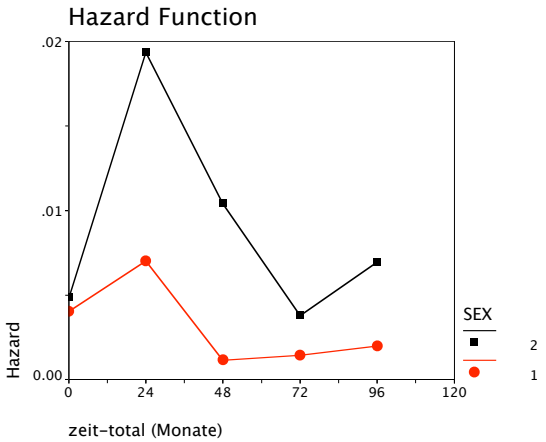
Definition **Hazardfunktion**: $h(t) = f(t)/S(t)$,
wo $f = F'$, F = Verteilungsfunktion der Überlebenszeiten, und $S(t) = 1 - F(t)$.

Sie wird auch als „failure rate“ (wegen der Anwendung in der Qualitätskontrolle) oder „force of mortality“ bezeichnet. Die Information, die in h, S, F und f steckt, ist äquivalent,

aber h ist an sich informativer, leider auch schwieriger zu schätzen. Es gibt einige typische Hazardverläufe:



Das „negative aging“ könnte man nach einer Operation finden, wo das Risiko unmittelbar danach am grössten ist. Das „positive aging“ findet man bei einem reinen Verschleissprozess. Die „bathtub curve“ entspricht in etwa der lebenslangen Hazardfunktion mit einem erhöhten Risiko nach der Geburt und im Alter. Für das Melanombeispiel erhalten wir für weibliche und männliche Patienten deutlich unterschiedliche Hazardfunktionen (mit der Life–table Methode in SPSS).



Die Mortalität insgesamt ist bei Frauen deutlich niedriger. Männer zeigen einen Anstieg des Risikos über 3–5 Jahre und danach einen Abfall.

Das Modell von Cox

Beim Cox–Modell handelt sich wieder um einen Typus der Regressionsanalyse. Wie bei der logistischen Regression ist die y–Variable das Ereignis, z. B. Tod. Eine besondere x–Variable ist die Zeit bis zum Event, die Überlebenszeit t. Das Cox–Modell erlaubt es, die Hazardfunktion $h(t)$ zusätzlich zur Abhängigkeit von der Überlebenszeit in Abhängigkeit von Einflussfaktoren (Risiko– oder protektive Faktoren) $x_1, \ldots, x_k$ zu untersuchen. Das Modell wird auch „proportional hazards model“ genannt. Es ist in der Krebsforschung und auch sonst in der klinischen Forschung relativ verbreitet.

Es soll hier nicht im Detail besprochen werden, da es mathematisch anspruchsvoll ist.

Das Vorgehen wird jetzt am Melanom–Beispiel illustriert. Als mögliche Risikofaktoren untersuchen wir das Geschlecht und die Eindringtiefe des Tumors (Breslow–Index).

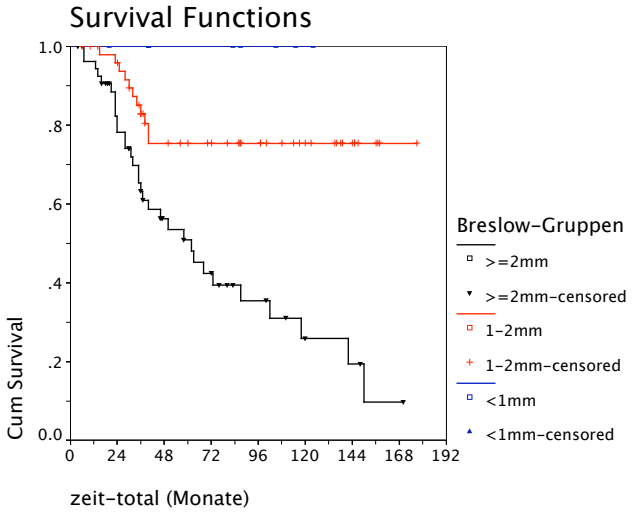
Obwohl die Survivalanalyse die Normalverteilung der Prädiktoren nicht voraussetzt, erweist sich eine Transformation zur approximativen Normalverteilung hin häufig als sinnvoll.

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
SEX	.640	.316	4.110	1	.043	1.897
V13	1.203	.217	30.817	1	.000	3.330

Die beiden Einflussgrössen sex und log(Breslow) = V13 sind beide im multiplen Cox–Modell signifikant (siehe „Sig.”). Der Hazard bei gegebenem Breslow–Index ist für einen Mann 1.9 mal so hoch wie für eine Frau (relativer Hazard in „Exp(B)”). Bei vorgegebenem Geschlecht erhöht sich der Hazard bei einer Erhöhung des log(Breslow) um eine Einheit um den Faktor 3.3. Man kann mit dem relativen Hazard ganz analog zum Odds ratio rechnen (siehe Abschnitt 6.2.3).

Um die Bedeutung der Eindringtiefe graphisch und quantitativ zu illustrieren, wird die Stichprobe in Tumore mit einem Durchmesser < 1mm, zwischen 1 und 2 mm und ≥ 2 mm geteilt.



Der Vergleich der Kaplan–Meier–Kurven zeigt eindrücklich die schlechtere Prognose für dickere Tumore.

Literatur:

Altman, D. G. (1991). *Practical statistics for medical research.* Chapman and Hall.
Kap. 13: Analysis of Survival Times

Armitage, P. & G. Berry (1987). *Statistical Methods in Medical Research.* Blackwell.
Kap. 14: Survivalanalyse

Matthews, D. E. & V. T. Farewell (1988). *Using and Understanding Medical Statistics.* Karger.
Kap. 6: Kaplan–Meier or 'Actuarial' Survival Curves
Kap. 7: The Log–Rank or Mantel–Haenszel Test for Comparison of Survival Curves
Kap. 12: Proportional Hazard Regression

Statistikprogramme

Survivalanalyse kann mit den folgenden Statistikprogrammen durchgeführt werden: BMDP, R, SAS (LIFETEST, LIFEREG, PHREG), SPSS, STATA (ltable, survival), StatView, Splus.

Index