

KLINISCHE STATISTIK

im Rahmen des Ökologischen Kursus

**Stoffkatalog
für den Schriftlichen Teil des Zweiten Abschnitts
der Ärztlichen Prüfung**

Wintersemester 2002/2003

Inhaltsverzeichnis

Teil I: Klinische Studien	1
Prinzipien der therapeutischen Prüfung	1
Kontrollierte klinische Studien	2
Prüfrichtlinien	5
Der Prüfplan	6
Der Befundbogen (Case Report Form CRF)	8
Auswertung	8
Beispiel eines Prüfplans:	9
Aufklärungsinhalte	14
Beispiel eines Prüfbogens	15
Zufällige Zuteilung der Behandlungen zu den Patientennummern	17
Teil II: Grundlagen der statistischen Auswertung	18
Beschreibende Statistik	18
Grundgesamtheit und Wahrscheinlichkeit	23
Schätzwerte (Statistiken) und Konfidenzintervalle	27
Grundlagen des statistischen Testens	31
Der verbundene (paarweise) t-Test	33
Der unverbundene t-Test zum Vergleich von zwei Mittelwerten	36
Vergleich von Wahrscheinlichkeiten (Chi ² -Test)	37
Vergleich von Zeiten (Überlebenszeit, Krankheitsdauer)	41
Teil III: Statistische Methoden in der Epidemiologie	44
Kohortenstudien	45
Fall-Kontrollstudien	49
Matched Pairs	53
Teil IV: Unterstützung von Diagnostik und Prognostik	54
Diagnostik und Klassifikation	54
Sensitivität und Spezifität	55
Prädiktive Werte	56

Teil 1: Klinische Studien

Prinzipien der therapeutischen Prüfung

Klinische Studien (klinische Prüfungen):

sind systematische Untersuchungen (Maßnahmen und Beobachtungen) **an Menschen (Patienten, Probanden)** mit dem Ziel, wissenschaftliche Erkenntnisse zu gewinnen. Bei den dabei systematisch angewandten Maßnahmen kann es sich um Behandlungen, diagnostische Verfahren, Vorsorgemaßnahmen oder Interventionen handeln.

In den **Grundsätzen für die ordnungsgemäße Durchführung der klinischen Prüfung von Arzneimitteln** (veröffentlicht vom Bundesminister für Jugend, Familie, Frauen und Gesundheit im Bundesanzeiger vom 30. Dezember 1987) heißt es:

"Klinische Prüfung im Sinne dieser Grundsätze ist die Anwendung eines Arzneimittels am Menschen zu dem Zweck, über den einzelnen Anwendungsfall hinaus Erkenntnisse über den therapeutischen oder diagnostischen Wert eines Arzneimittels, insbesondere über seine Wirksamkeit und Unbedenklichkeit, zu gewinnen; dies gilt unabhängig davon, ob die Prüfung in einer Klinik oder in der Praxis eines niedergelassenen Arztes durchgeführt wird."

Man unterscheidet:

- **Beobachtungsstudien:**

Die Maßnahmen werden nicht systematisch variiert.

Beispiele:

Beobachtung der Verträglichkeit eines neuen Medikaments,

Beobachtung von Schwangerschaftsverlauf und Mißbildung,

Beobachtung der Lebenserwartung und Lebensqualität von Tumorpatienten.

- **Kontrollierte Studien:**

Die Maßnahmen werden systematisch variiert.

Beispiel:

Wirksamkeitsnachweis eines neuen Mittels im Vergleich zur Standardbehandlung.

Phasen der klinischen Prüfung von Arzneimitteln:

Phase I: Untersuchung der Verträglichkeit und Kinetik (meist an Probanden)

Phase II: Auffinden der wirksamen Dosis bei Patienten

Phase III: Nachweis der Wirksamkeit und Verträglichkeit

Phase IV: Studien nach Zulassung des Arzneimittels

Kontrollierte klinische Studien

Kontrollierte klinische Studien dienen zur **klinischen Prüfung von Arzneimitteln auf Wirksamkeit und Verträglichkeit.**

"Grundsätzlich sollen klinische Prüfungen, wenn dies angemessen, d.h. dem therapeutischen Ziel nach sinnvoll und in der Durchführung auch möglich ist, **kontrolliert** durchgeführt werden. Dies schließt eine gleichzeitig beobachtete Kontrollgruppe und eine **randomisierte Zuteilung** der Patienten beziehungsweise Probanden zu den Behandlungsgruppen ein."

(Grundsätze für die ordnungsgemäße Durchführung der klinischen Prüfung von Arzneimitteln des BMJFFG, 30. Dezember 1987)

Charakteristika kontrollierter klinischer Studien sind:

- Vergleich der Prüfbehandlung mit einer Kontrollbehandlung,
- randomisierte (zufällige) Zuteilung der Patienten zu den Gruppen.

Mit der randomisierten Zuteilung soll die "Strukturgleichheit" der **Gruppen** bezüglich Ausgangsdaten, prognostischer Faktoren und Zusatzmaßnahmen erreicht werden. Die Patienten beider Gruppen sollen zusammen eine Zufallsstichprobe aus der Gesamtheit aller möglichen Anwender sein.

Ethische Voraussetzung: Deklaration von Helsinki

- Medizinischer Fortschritt beruht auf Forschung, die sich letztlich auch auf Versuche am Menschen stützen muß.
- Biomedizinische Forschung am Menschen muß den allgemein anerkannten wissenschaftlichen Grundsätzen entsprechen.
- Biomedizinische Forschung am Menschen ist nur zulässig, wenn die Bedeutung des Versuchsziels in einem angemessenen Verhältnis zum Risiko für die Versuchsperson steht.
- Das Recht der Versuchsperson auf Wahrung ihrer Unversehrtheit muß stets geachtet werden.
- Bei jedem Versuch am Menschen muß jede Versuchsperson ausreichend über Absicht, Durchführung, erwarteten Nutzen und Risiken des Versuches ... unterrichtet werden. Nach dieser Aufklärung sollte der Arzt die **freiwillige Zustimmung** der Versuchsperson einholen (informed consent).

Rechtliche Voraussetzungen: Arzneimittelgesetz (AMG)

- §22 Dem Antrag auf Zulassung müssen vom Antragsteller folgende Angaben in deutscher Sprache beigefügt werden:
(3)...die Ergebnisse der klinischen oder sonstigen ärztlichen...Erprobung (klinische Prüfung).
- §24 Den nach §22...erforderlichen Unterlagen sind Gutachten von Sachverständigen beizufügen... Im einzelnen muß aus den Gutachten insbesondere hervorgehen:
...aus dem klinischen Gutachten, ob das Arzneimittel bei den angegebenen Anwendungsgebieten angemessen wirksam ist, ob es verträglich ist, ob die vorgesehene Dosierung zweckmäßig ist und welche Gegenanzeigen und Nebenwirkungen bestehen.

Siebenter Abschnitt

Schutz des Menschen bei der klinischen Prüfung

§40 Allgemeine Voraussetzungen

Die klinische Prüfung eines Arzneimittels darf bei Menschen nur durchgeführt werden, wenn und solange

1. die Risiken, die mit ihr für die Person verbunden sind, bei der sie durchgeführt werden soll, gemessen an der voraussichtlichen Bedeutung des Arzneimittels für die Heilkunde ärztlich vertretbar sind,
2. die Person, bei der sie durchgeführt werden soll, ihre Einwilligung hierzu erteilt hat, nachdem sie durch den Arzt über Wesen, Bedeutung und Tragweite der klinischen Prüfung aufgeklärt worden ist, und mit der Einwilligung zugleich erklärt, daß sie mit der im Rahmen der klinischen Prüfung erfolgenden Aufzeichnung von Krankheitsdaten und ihrer Weitergabe zur Überprüfung an den Auftraggeber, an die zuständige Überwachungsbehörde oder die zuständige Bundesoberbehörde einverstanden ist,
3. die Person, bei der sie durchgeführt werden soll, nicht auf gerichtliche oder behördliche Anordnung in einer Anstalt untergebracht ist,
4. sie von einem Arzt geleitet wird, der mindestens eine zweijährige Erfahrung in der klinischen Prüfung von Arzneimitteln nachweisen kann,
5. eine dem jeweiligen Stand der wissenschaftlichen Erkenntnisse entsprechende pharmakologisch-toxikologische Prüfung durchgeführt worden ist,
6. die Unterlagen über die pharmakologisch-toxikologische Prüfung, der dem jeweiligen Stand der wissenschaftlichen Erkenntnisse entsprechende Prüfplan mit Angabe von Prüfern und Prüforten und die Voten der Ethik-Kommissionen bei der zuständigen Bundesoberbehörde vorgelegt worden sind,

7. der Leiter der klinischen Prüfung durch einen für die pharmakologisch-toxikologische Prüfung verantwortlichen Wissenschaftler über die Ergebnisse der pharmakologisch-toxikologischen Prüfung und die voraussichtlich mit der klinischen Prüfung verbundenen Risiken informiert worden ist und
8. für den Fall, daß bei der Durchführung der klinischen Prüfung ein Mensch getötet oder der Körper oder die Gesundheit eines Menschen verletzt wird, eine Versicherung nach Maßgabe des Absatzes 3 besteht, die auch Leistungen gewährt, wenn kein anderer für den Schaden haftet.

Die klinische Prüfung eines Arzneimittels darf bei Menschen vorbehaltlich des Satzes 3 nur begonnen werden, wenn diese zuvor von einer nach Landesrecht gebildeten unabhängigen Ethik-Kommission zustimmend bewertet worden ist; Voraussetzung einer zustimmenden Bewertung ist die Einhaltung der Bedingungen in Satz 1. Soweit keine zustimmende Bewertung der Ethik-Kommission vorliegt, darf mit der klinischen Prüfung erst begonnen werden, wenn die zuständige Bundesoberbehörde innerhalb von 60 Tagen nach Eingang der Unterlagen nach Satz 1 Nr. 6 nicht widersprochen hat. Über alle schwerwiegenden oder unerwarteten unerwünschten Ereignisse, die während der Studie auftreten und die Sicherheit der Studienteilnehmer oder die Durchführung der Studie beeinträchtigen könnten, muß die Ethik-Kommission unterrichtet werden.

§67 Allgemeine Anzeigepflicht

- (1) Betriebe und Einrichtungen, die Arzneimittel entwickeln, herstellen, **klinisch prüfen** oder einer Rückstandsprüfung unterziehen, prüfen, lagern, verpacken, in den Verkehr bringen oder sonst mit ihnen Handel treiben, haben dies vor Aufnahme der Tätigkeiten der zuständigen Behörde anzuzeigen
- (6) Der pharmazeutische Unternehmer hat Untersuchungen, die dazu bestimmt sind, Erkenntnisse bei der Anwendung zugelassener Arzneimittel zu sammeln, den kassenärztlichen Bundesvereinigungen sowie der zuständigen Bundesoberbehörde unverzüglich anzuzeigen.

Prüfrichtlinien

Grundsätze für die ordnungsgemäße Durchführung der klinischen Prüfung von Arzneimitteln

Bundesminister für Jugend, Familie, Frauen und Gesundheit
Bundesanzeiger Nr.243, Mittwoch, 30. 12. 1987

Arzneimittelprüfrichtlinien und Allgemeine Verwaltungsvorschrift zur Anwendung der Arzneimittelprüfrichtlinien

Bundesminister für Jugend, Familie, Frauen und Gesundheit
Bundesanzeiger Nr. 243 a, Freitag, 29. 12. 1989

Good Clinical Practice (GCP) for Trials on Medicinal Products in the European Community (Gute Klinische Praxis für die klinische Prüfung von Arzneimitteln in der Europäischen Gemeinschaft)

Empfehlungen des Spezialitätenausschusses der EG vom 11.7.1990, in Kraft gesetzt am 1.7.1991.

Biostatistical methodology in clinical trials in applications for marketing authorizations for medical products

Note for Guidance; Committee for Proprietary Medicinal Products
Efficacy Working Party, Brussels, December 1994, III/3630/92-EN

Structure and Content of Clinical Study Reports

ICH-Efficacy Topic 3, Draft 10 -Oct. 27, 1994

Der Prüfplan

soll folgendermaßen gegliedert sein:

- 1. Allgemeine Angaben**
 - Titel der Studie
 - Name und Adresse der Verantwortlichen
 - Einrichtungen und Orte

- 2. Begründung und Ziele, ethische Aspekte**
 - Grund für die Durchführung
 - Voraussetzungen und Hintergrundinformation
 - Ziele
 - Nutzen und Risiken

- 3. Patienten oder Probanden, Ein- und Ausschlußkriterien**
 - Untersuchungspopulation
 - Auswahlverfahren
 - Ein- und Ausschlußkriterien
 - Kontrolle der Kriterien
 - Aufklärung und Zustimmung
 - Zahl der Patienten oder Probanden

- 4. Beabsichtigte Maßnahmen**
 - Prüf-, Vergleichs- und Kontrollmaßnahmen
 - Art, Dosierung und Dauer der Anwendung
 - Begründung für Art, Dosierung und Dauer
 - Risiken und Belästigungen
 - Maßnahmen zur Reduktion von Risiken und Belästigungen
 - Kontrolle der Durchführung

- 5. Studientyp**
 - Beobachtungsstudie, kontrollierte Studie
 - Randomisierung, Verblindung u.ä.
 - Zuteilung zu den Patienten oder Probanden
 - Interindividuell oder intraindividuell
 - Uni- oder multizentrisch, Stratifikation
 - Kontrolle der Zuteilung und Verblindung

- 6. Messungen, Befundungen, Beobachtungen**
 - vorgesehene Messungen und Befundungen
 - Zeitpunkte der Messung und Befundung
 - Validität der Verfahren
 - Risiken und Risikoreduktion

- 7. Ausführliche Beschreibung des Studienablaufs**
 - Studienphasen (Vorphase, Hauptphase, Nachbeobachtung)
 - zulässige und unzulässige Begleitmaßnahmen
 - Überwachung der Studie

- 8. Zielgrößen**
 - Bewertung der Studienziele
 - Kriterien für Wirksamkeit und Verträglichkeit
 - Stör- und Begleitgrößen

- 9. Datenerfassung und Dokumentation**
 - Ausfüllen der Dokumentationsbogen
 - Kontrolle der Richtigkeit (Monitor)
 - Computereingabe und Kontrolle
 - Dokumentation und Datensicherheit
 - Datenschutz

- 10. Auswertung**
 - Auswertungsplan (statistische Modelle, zu prüfende Hypothesen, statistische Verfahren)
 - Statistische Auswertungssysteme
 - Qualitätskontrolle der Auswertung
 - Ein- und Ausschluß von Befunden
 - Konfirmatorische und explorative Analyse
 - Behandlung der Stör- und Begleitgrößen
 - Statistische Begründung der Patientenzahl

- 11. Patientensicherheit und Compliance**
 - Überwachung der Sicherheit
 - Erfassung, Bewertung und Dokumentation unerwünschter Ereignisse (UAW)
 - Maßnahmen bei unerwünschten Ereignissen
 - Kriterien für individuellen Studienabbruch
 - Kriterien für Abbruch der gesamten Studie
 - Versicherungsschutz
 - Überwachung der Patientencompliance

Der Befundbogen (Case Report Form CRF)

soll Vorgaben zur standardisierten Erfassung folgender Daten haben:

1. Identifikation von
 Studie, Ort, Patientenummer (anonymisiert), (Behandlungsgruppe)
2. Basisdaten
 Alter, Geschlecht, Größe, Gewicht
3. Ein- und Ausschlusskriterien
4. Diagnose und Dauer der Erkrankung
5. Anamnese, Vorbehandlungen und Risikofaktoren
6. Zusatzerkrankungen und Zusatzmedikation
7. Ausgangsbefunde
8. Verlaufsbefunde, Kontrollbefunde, zusätzlich durchgeführte Maßnahmen
9. Unerwünschte Ereignisse und Besonderheiten (Krankheit, Änderung der
 Behandlung, Abbruch)
10. Abschlußbeurteilung

Auswertung

Die Auswertung umfaßt folgende Schritte:

1. Beschreibung der Befunde, getrennt nach Gruppen
 Häufigkeitsverteilung
 Mittelwert, Standardabweichung, Minimum, Maximum, usw. bei
 quantitativen Größen
2. Prüfen der Strukturgleichheit der Gruppen bezüglich der Ausgangsbefunde
 Vergleich von Häufigkeiten: Chi²-Test
 Vergleich von Mittelwerten: t-Test
 Mann-Whitney-Test
3. Analyse der Studienziele
 Konfirmatorischer Vergleich der Gruppen in den Zielgrößen. (statistische
 Tests; Chi², t-Test, Mann-Whitney-Test, Lifetable-Analyse)
4. Weitere Analysen
 z.B. Subgruppenanalysen, Regressionsanalysen
5. Auflistung und Bewertung der unerwünschten Arzneimittelwirkungen und
 besonderen Ereignisse

Beispiel eines Prüfplans:

1. Allgemeine Angaben:

Titel der Studie: Placebo-kontrollierte klinische Studie mit einer Erkältungssalbe bei akuten katarrhalischen Infekten
(Kurztitel: Erkältungsstudie)

Leiter der Studie: Dr. A. D., praktischer Arzt

Einrichtung: Praxis von Dr. A.D.

2. Begründung und Ziele der Studie, ethische Aspekte:

Bei dem Medikament handelt es sich um ein pflanzliches Arzneimittel mit folgender Zusammensetzung (bezogen auf 100 g): Ol.Pini sibir. 8,0 g, Ol.Pini pumil. 2,0 g, Ol.Terebinth. 1,0 g, Ol.Thymi 1,0 g, Ol.Eucalypti 8,0 g, Ol.Rosmarini 8,0 g.

Das Medikament ist als Fertigarzneimittel registriert und im Handel. Die Wirksamkeit wurde bisher aber noch nicht in kontrollierten Studien überprüft. Das Ziel der Studie besteht deshalb darin, die Wirksamkeit und Verträglichkeit des Medikaments in einem kontrollierten Vergleich mit einer Placebo-Salbe (Salbengrundlage) bei akuten katarrhalischen Erkrankungen oder bei akuten Bronchitiden zu untersuchen. Bei den bisherigen Anwendungen sind keine schwerwiegenden Nebenwirkungen bekannt geworden. Die Anwendung ist deshalb ethisch gerechtfertigt. Da die Wirksamkeit noch nicht nachgewiesen ist und für die Indikation keine wirksame Standardtherapie bekannt ist, erscheint die Verwendung eines Placebos gerechtfertigt.

3. Charakterisierung der Patienten; Ein- und Ausschlusskriterien:

Es werden alle Patienten eingeschlossen, die nach Studienbeginn mit einem akuten katarrhalischen Infekt oder einer akuten Bronchitide in die Praxis kommen, die Einschlusskriterien erfüllen und keine Ausschlusskriterien aufweisen. Es ist die Erfassung von insgesamt 100 Studienpatienten vorgesehen.

Einschlusskriterien:

Eingeschlossen werden Patienten über 12 Jahre mit den oben angegebenen Erkrankungen, die nach ausführlicher Aufklärung ihre Zustimmung zur Teilnahme geben (bei unter 18-jährigen ist auch die Zustimmung des Erziehungsberechtigten erforderlich.) Die Diagnose wird anhand der subjektiven Beschwerden (Schmerzen, allgemeine Abgeschlagenheit) und der objektiven Befunde (Husten, Auswurf, Schwellung oder Rötung im Hals, Temperatur) getroffen. Der Beginn der Erkrankung darf nicht länger als 48 Stunden zurückliegen.

Ausschlußkriterien:

Auszuschließen sind:

- schwangere und stillende Frauen,
- Patienten, die eine andere anerkannte Therapie benötigen
- Patienten mit bekannter Allergie gegen die in der Salbe enthaltenen Bestandteile
- Patienten mit Polyallergieneigung
- Patienten mit chronischer Bronchitis
- Patienten, die eine Teilnahme verweigern.

4. Beabsichtigte Maßnahmen:

Die Patienten sollen mit der ihnen zugeteilten Salbe Brust, Rücken und Hals mindestens dreimal täglich einreiben. Verum und Placebo werden in Tuben mit 45 g Inhalt verabreicht. Bei Verum handelt es sich um die Erkältungssalbe mit pflanzlichen Ölen in der oben angegebenen Zusammensetzung, bei Placebo um die Salbengrundlage, die nach den Bedingungen der IFAR hergestellt wird. Es steht den Patienten frei, zusätzlich eine physikalische Therapie (Wärmebehandlung) anzuwenden. Während der Studie sind keine Medikamente gegen den katarrhalischen Infekt oder die akute Bronchitis zulässig. Die Notwendigkeit einer Antibiotikatherapie bedingt den Abbruch der Studie bzw. ist ein Ausschlußgrund. Die Behandlung von anderen chronischen oder interkurrenten Erkrankungen mit der entsprechenden Therapie ist uneingeschränkt zulässig, muß aber im Prüfbogen vermerkt werden. Die Behandlung mit der Studienmedikation wird bei Symptommfreiheit des Patienten beendet. Sie dauert maximal 14 Tage.

5. Studientyp:

Es handelt sich um eine Placebo-kontrollierte Studie der Phase III. Die Zuteilung der Patienten zu den beiden Behandlungsgruppen (Verum, Placebo) erfolgt randomisiert nach einer Blockrandomisation mit der Blocklänge 2. Verum und Placebo sind in neutralen Tuben mit der Aufschrift A oder B und 'nur zur klinischen Prüfung bestimmt' abgepackt. Die Zuteilung der Buchstaben A oder B zu den beiden Behandlungen erfolgt zufallsgemäß und ist dem Prüfarzt nicht bekannt. Allerdings ist nach Geruch und Farbe die Medikation identifizierbar. Insofern ist die Zuteilung nicht verblindet. Der Arzt hat Listen mit den fortlaufenden Patientenummern, denen die Behandlungsgruppe A oder B zugeordnet ist. Er soll bei Studienbeginn dem Patienten eine Tube mit dem entsprechenden Buchstaben aushändigen und ihn in die Benutzung einweisen. Eine Stratifikation der Patienten ist nicht vorgesehen.

6. Messungen, Befundungen und Beobachtungen:

Zu Beginn werden folgende Befunde erhoben:

- Diagnose
- Überprüfung der Ein- und Ausschlußkriterien
- Zusatzerkrankungen
- Zusatzmedikamente
- Gripeschutzimpfung in den letzten 12 Monaten
- Beginn der Erkrankung

- bisherige Selbstmedikation
- Symptome der Erkrankung (Husten, Auswurf, Schleim, Schmerzen und/oder Kratzen im Hals, Schwellung und/oder Rötung im Hals, Schnupfen, Fieber, allgemeine Abgeschlagenheit). Die Symptome werden mit einem Score 0 (nicht vorhanden), 1 (leicht) oder 2 (stark) bewertet.

Die Befundung und Bewertung der 8 Symptome werden bei Zwischenuntersuchungen nach 2-4 Tagen, 6-8 Tagen und bei der Abschlußuntersuchung nach 10-14 Tagen wiederholt (falls der Patient nicht symptomfrei ist). Bei den Zwischenuntersuchungen und der Abschlußuntersuchung werden unerwünschte Ereignisse (UE) erfragt. Außerdem wird die Zahl der täglichen Anwendungen und der Zeitpunkt erfragt, an dem der Patient eine Besserung verspürte oder symptomfrei war.

7. Ausführliche Beschreibung des Studienablaufs:

Nach Abklärung der Diagnose und Überprüfung der Ausschlusskriterien wird ein Patient, der die Einschlusskriterien erfüllt und keine Ausschlusskriterien aufweist, über die Studie und die Behandlung aufgeklärt und um seine Zustimmung zur Teilnahme er sucht. Willigt er (bei Anwesenheit einer Zeugen) ein, an der Studie teilzunehmen, wird ihm eine fortlaufende Patientenummer gegeben und er wird in die Behandlungsgruppe A oder B eingeteilt, die in der Zuteilungsliste seiner Patientenummer zugeordnet ist. Er erhält vom Arzt die entsprechende Salbe und wird in ihre Benutzung eingewiesen. Die anamnestischen Daten sowie die Symptome und Befunde werden erhoben und die Ergebnisse in den entsprechenden Feldern des Erfassungsbogens eingetragen. Der Patient soll mit der ihm zugeteilten Salbe Brust, Rücken und Hals mindestens 3 mal täglich einreiben. Bei Schnupfen kann auch eine kleine Menge in die Nasenlöcher eingestrichen werden. Die Zahl der Anwendungen wird auf dem Prüfbo gen vermerkt. Die Untersuchung und Befundung werden nach 2-4 Tagen, 6-8 Tagen und 10-14 Tagen wiederholt. Bei den Zwischenuntersuchungen und der Endunter suchung werden Nebenwirkungen oder Komplikationen (z.B. interkurrente Erkrankun gen) erfragt und auf dem Befundbogen registriert. Die Behandlung wird abgeschlos sen, wenn entweder alle Symptome verschwunden sind oder eine Behandlungsdauer von 14 Tagen erreicht ist. Ein vorzeitiger Behandlungsabbruch wird zusammen mit dem Grund dokumentiert. Am Ende beurteilt der Arzt die Therapiewirkung.

8. Zielgrößen, Störgrößen:

Primäre Zielgröße für die Wirksamkeit ist die Dauer der Symptome, die aus den Anga ben des Patienten und den Befunden des Arztes bei den Zwischenuntersuchungen festgelegt wird. Sekundäre Zielgrößen sind:

Dauer bis zur Besserung und Änderung der Symptomausprägungen.

Als mögliche Störgrößen werden berücksichtigt: Alter der Patienten, Zusatzerkrankun gen, Dauer zwischen Erkrankungsbeginn und Behandlungsbeginn, Gripeschutzimpf ung.

9. Datenerfassung und Dokumentation:

Die erfaßten Daten werden vom Arzt in die ihm bei Studienbeginn ausgehändigten Befundbogen (Case Report Forms CRF) nach den festgelegten Richtlinien (Standard Operating Procedure SOP) eingetragen. Die eingetragenen Werte dürfen nicht ausradiert oder sonst gelöscht werden. Korrekturen sind daneben einzutragen. Sie sind deutlich zu kennzeichnen (z.B. durch andere Farbe) und vom Arzt zu bestätigen.

Die erfaßten Daten werden vom Monitor auf Vollständigkeit und Richtigkeit überprüft und gegebenenfalls dem Arzt zur Korrektur oder Ergänzung vorgelegt. Die Daten werden in ein Datenbanksystem (dBase+) eingegeben, gespeichert und gesichert (backup). Die eingegebenen Daten werden (z. B. durch Doppeleingabe) kontrolliert. Korrekturen der Datenbank sind zu protokollieren; der jeweils vor den Korrekturen bestehende Zustand der Datenbank ist zu speichern und aufzubewahren, so daß alle Korrekturschritte nachvollzogen werden können.

Die zur Datenerfassung weitergeleiteten Befundbogen dürfen nicht Namen oder Adresse des Patienten enthalten. In der Datenbank werden die Daten nur mit der Studiennummer identifiziert (anonymisierte Daten).

10. Auswertung:

Alle erfaßten Befunde werden - nach Gruppen getrennt - statistisch beschrieben (Häufigkeiten, statistische Kenngrößen der Verteilungen). Die Strukturgleichheit beider Gruppen wird durch Tests der Unterschiede in den Basisdaten (Alter, Geschlecht, Zusatzerkrankungen) und Ausgangsbefunden überprüft.

Zur confirmatorischen Auswertung der primären Zielgröße wird mit dem Logrank-Test überprüft, ob die Verteilungen der Dauern bis Symptombefreiheit in beiden Gruppen gleich sind (Nullhypothese) oder ob die Verteilung der mit Verum behandelten Gruppe zu kürzeren Dauern verschoben ist (einseitige Alternative). Die Nullhypothese wird abgelehnt, wenn die Signifikanzwahrscheinlichkeit kleiner als 0.05 ist. Die Auswertung erfolgt nach dem 'intention to treat'-Prinzip; d.h. alle Patienten, die mindestens zu einer Nachuntersuchung kamen, werden in die Auswertung einbezogen, gleichgültig ob sie protokollgemäß die Studie beendet haben oder Protokollverletzungen vorlagen. Bei vorzeitigem Abbruch wird die Dauer vom Therapiebeginn bis zum Abbruch als Zielgröße genommen und der Patient als 'nicht symptomfrei' gewertet (withdrawn, zensierte Dauer). Die Verteilung der Dauer bis zur Besserung wird zwischen beiden Gruppen ebenfalls mit dem Logrank-Test verglichen. Bei den einzelnen Symptomen werden die Unterschiede in der Verteilung der Ausprägungen zwischen beiden Gruppen für jede Untersuchungszeit mit dem χ^2 -Test überprüft. Unterschiede zwischen beiden Gruppen in den Änderungen der Ausprägung zwischen Ausgangsbefund und Endbefund (gebessert, unverändert, verschlechtert) werden mit dem χ^2 -Test geprüft. Bei Patienten, die vorzeitig abbrechen (aber mindestens zu einer Zwischenuntersuchung gekommen sind), wird der zuletzt erhobene Befund als Endbefund genommen. Der Einfluß der Störgrößen auf die Therapieergebnisse wird mit Regressionsmodellen (Cox-Analyse) untersucht.

Die unerwünschten Ereignisse werden einzeln aufgelistet. Unterschiede in der Häufigkeit unerwünschter Ereignisse zwischen beiden Gruppen werden mit dem χ^2 -Test überprüft.

Die Ergebnisse der biometrischen Auswertung werden in einem Bericht (entsprechend den EG-Richtlinien) dargestellt. Die Auswertung erfolgt mit dem System SPSS. Die Daten werden vom System unmittelbar und unverändert aus der Datenbank übernommen.

Begründung für die Patientenzahl

Die Begründung der Patientenzahl basiert auf einem Vergleich der Wahrscheinlichkeiten für Symptombefreiheit (in 14 Tagen) zwischen beiden Gruppen. Es wird angenommen, daß diese Wahrscheinlichkeit in der Placebo-Gruppe 40% beträgt. Um mit einer 'Power' von 95% bei einer Signifikanzschwelle von 5% ein 'signifikantes' Ergebnis zu erhalten, wenn diese Wahrscheinlichkeit in der Verum-Gruppe 70% beträgt (30% größer ist), ist ein Stichprobenumfang von 50 Patienten pro Gruppe (insgesamt 100 Patienten) erforderlich.

11. Patientensicherheit und Compliance:

Die Überwachung der Patientensicherheit geschieht durch laufende Kontrolle der unerwünschten Ereignisse. Die Ereignisse werden vom Arzt nach ihrer Schwere und einem möglichen Zusammenhang mit der Studientherapie bewertet (WHO-Kriterien). Nach Maßgabe des behandelnden Arztes ist bei unerwünschten Ereignissen die Studie für den Patienten abzubrechen und es sind geeignete Maßnahmen vorzunehmen. Insbesondere ist im Einzelfall die Studie abzubrechen, wenn sich der Krankheitszustand verschlimmert und zusätzliche Erkrankungen (wie z. B. Otitis, Pharyngitis) hinzukommen, die eine Antibiotikatherapie erfordern. Der Patient kann jederzeit von sich aus die Teilnahme an der Studie beenden.

Die Studie wird insgesamt abgebrochen, wenn die Rate und Schwere der unerwünschten Ereignissen nicht mehr toleriert werden können oder sich neue Erkenntnisse ergeben, die eine Fortsetzung der Studie nicht mehr sinnvoll erscheinen lassen. Die Entscheidung über den Studienabbruch trifft der klinische Leiter der Studie. Für die Patienten wird eine Versicherung entsprechend dem Arzneimittelgesetz abgeschlossen. Die Aufklärung erfolgt gemäß § 40/41 AMG mündlich in Anwesenheit eines Zeugen. Der Patient wird über seine Erkrankung, die Behandlungsmöglichkeiten (einschließlich ergänzende Behandlung z.B. durch Wärme) und über die Zusammensetzung des Prüfmedikaments sowie über die bisherigen Erfahrungen mit dieser Salbe informiert. Er wird aufgeklärt über das Studienziel (Nachweis der therapeutischen Wirksamkeit) und die Studienanlage (Placebo-kontrollierte Studie mit randomisierter Zuteilung). Die Notwendigkeit und ärztliche Unbedenklichkeit einer Placebo-Behandlung werden ihm erläutert. Er wird darüber informiert, daß ihm keinerlei Nachteile entstehen, wenn er nicht an der Studie teilnimmt oder die Studie abbricht. Er wird auch darauf hingewiesen, daß er bei einer Zustimmung zur Teilnahme die Anweisungen des Arztes sorgfältig befolgen und den Arzt über alle auffälligen Krankheitserscheinungen und über zusätzlich eingenommene Mittel informieren soll.

Die Compliance (protokollgemäße Anwendung) des Patienten wird durch die Angaben über die Zahl der Anwendungen sowie durch Inspektion der zurückgegebenen Tuben kontrolliert.

Aufklärungsinhalte

Placebo-kontrollierte klinische Studie mit ERKÄLTUNGSSALBE bei akuten katarrhalischen Infekten

Im Aufklärungsgespräch wird der Patient über folgendes informiert:

1. Erkrankung und Behandlungsmöglichkeiten

Er wird über die Art seiner Erkrankung (akuter katarrhalischer Infekt, akute Bronchitide) und ihre Behandlungsmöglichkeiten informiert. Dabei wird er auf die mögliche Wirkung durch Einreiben mit der Erkältungssalbe, aber auch auf eine unterstützende Behandlung durch Wärme hingewiesen. Auf Wunsch werden ihm die Bestandteile der Salbe erläutert. Über die bisherigen Erfahrungen mit der Erkältungssalbe wird er informiert. Dabei wird ihm auch erklärt, daß Nebenwirkungen (z.B. Allergien) zwar unwahrscheinlich sind, aber nicht ausgeschlossen werden können.

2. Studienziel und Studienanlage

Es wird ihm erläutert, daß zum Nachweis der therapeutischen Wirksamkeit eine kontrollierte Studie durchgeführt wird. Dies bedeutet, daß gleichzeitig mit dem Erkältungsbalsam auch eine Vergleichsbehandlung geprüft werden muß und die Behandlungsbedingungen für beide möglichst gleich sein sollen. Dies kann am besten dadurch gewährleistet werden, daß die Zuteilung der beiden Behandlungen zu den Patienten zufallsgemäß erfolgt, so daß jeder Patient die gleiche Chance hat, das Prüfpräparat oder Vergleichspräparat zu erhalten.

Als Vergleichspräparat wird die Salbengrundlage ohne pflanzliche Wirkstoffe verwendet. Da der therapeutische Wert dieser Wirkstoffe noch nicht bewiesen ist, hat nach dem jetzigen Kenntnisstand der Patient dadurch keinen Nachteil zu erwarten.

3. Freiwilligkeit der Teilnahme und Versicherungsschutz

Der Patient wird darauf hingewiesen, daß es ihm völlig freisteht, ob er an der Studie teilnehmen will oder nicht. Sowohl bei einer Zustimmung als auch bei einer Ablehnung wird ihm vom Arzt die nach dem jetzigen Erkenntnisstand bestmögliche Behandlung zuteil. Er kann seine Zustimmung zur Teilnahme jederzeit widerrufen, ohne daß ihm dadurch Nachteile entstehen. Er soll die Anweisungen des Arztes sorgfältig befolgen und den Arzt über alle auffälligen Krankheitserscheinungen und über zusätzlich eingenommene Mittel informieren. Für die Teilnehmer an der Studie wurde entsprechend den Vorschriften des Arzneimittelgesetzes eine Versicherung abgeschlossen, die sie bei möglicherweise vorkommenden Schäden entschädigt.

Beispiel eines Prüfbogens

Studie: Erkältungssalbe

Pat. Nr. _____

Beh. Gruppe: _____

Blatt 1

Patientendaten und Anamnese

Pat.Initialen _____ Geb.Dat.____.____.____. Geschl. weibl. männl.

Diagnose: akuter katarrhalischer Infekt akute Bronchitis

Zusatzangaben: _____

Zusatzerkrankungen und Medikation: keine ja

wenn ja:

Erkrankung

seit

Medikation

seit

_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____

Gripeschutzimpfung in den letzten 12 Monaten? ja nein

Beginn der akuten Erkrankung am: _____.____._____

bisherige Selbstmedikation: keine ja

wenn ja welche: _____

Ausschlusskriterien wurden überprüft und liegen nicht vor:

Der Patient hat nach Aufklärung sein Einverständnis erteilt:

Beginn der Studienbehandlung am: _____.____._____

Bemerkungen:

Verlaufsbefunde

Datum	Behandlungs- beginn	nach 2-4 Tagen	nach 6-8 Tagen	nach 10-14 Tagen
1. Husten	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>
	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>
	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>
2. Auswurf	keiner <input type="checkbox"/>	keiner <input type="checkbox"/>	keiner <input type="checkbox"/>	keiner <input type="checkbox"/>
	gering <input type="checkbox"/>	gering <input type="checkbox"/>	gering <input type="checkbox"/>	gering <input type="checkbox"/>
	erhöht <input type="checkbox"/>	erhöht <input type="checkbox"/>	erhöht <input type="checkbox"/>	erhöht <input type="checkbox"/>
3. Schleim:	feststehend <input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	zäh <input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	locker <input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Schmerzen und Kratzen in Hals	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>
	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>
	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>
5. Schwellung oder Rötung im Hals	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>
	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>
	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>
6. Schnupfen	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>
	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>
	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>
7. Fieber	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>
	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>
	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>
8. allgemeine Abgeschlagenheit	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>	nein <input type="checkbox"/>
	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>	leicht <input type="checkbox"/>
	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>	stark <input type="checkbox"/>

Zahl der Anwendungen _____

Nebenwirkungen oder Komplikationen nein ja nein ja nein ja
wenn ja, welche? _____

Der Patient verspürte eine Besserung am: ____.

Symptomfreiheit am: ____.

Abbruch der Behandlung am: ____ wegen: _____

Beurteilung der Wirkung durch den Arzt: gut mäßig schlecht

(Unterschrift)

Zufällige Zuteilung der Behandlungen zu den Patientennummern

Pat Nr.	Beh.	Pat.Init.	Datum	Pat.Nr.	Beh.	Pat.Init.	Datum
1	A			26	B		
2	B			27	A		
3	A			28	B		
4	B			29	B		
5	B			30	A		
6	A			31	A		
7	A			32	B		
8	B			33	A		
9	B			34	B		
10	A			35	B		
11	B			36	A		
12	A			37	A		
13	A			38	B		
14	B			39	B		
15	A			40	A		
16	B			41	A		
17	A			42	B		
18	B			43	A		
19	A			44	B		
20	B			45	A		
21	B			46	B		
22	A			47	A		
23	A			48	B		
24	B			49	A		
25	A			50	B		

Teil II: Grundlagen der statistischen Auswertung

Beschreibende Statistik

Der erste Schritt der Auswertung besteht darin, die Gesamtheit der erfaßten Daten statistisch zu beschreiben. Daten sind die bei den 'Beobachtungseinheiten' (Patienten) erfaßten Werte oder Ausprägungen von 'Merkmalen'. Die Beobachtungseinheiten werden durch eine fortlaufende Nummer $i=1,2,\dots$ gekennzeichnet, wobei die Gesamtzahl der Beobachtungseinheiten oft mit n bezeichnet wird. Merkmale werden durch kleine Buchstaben x, y, z, \dots symbolisiert; x_i, y_i, z_i, \dots sind dann die bei der Beobachtungseinheit i erfaßten Merkmalwerte (Daten).

Als Beispiel sollen die Daten von 20 Patienten der im Teil I besprochenen Erkältungsstudie beschrieben werden:

Nr.	Gruppe	Geschlecht	Alter (Jahre)	Husten zu Beginn	Husten am Ende
1	Verum	männlich	33	leicht	keiner
2	Placebo	weiblich	34	stark	leicht
3	Verum	weiblich	35	stark	leicht
4	Placebo	weiblich	20	leicht	leicht
5	Placebo	weiblich	31	leicht	keiner
6	Verum	männlich	43	leicht	keiner
7	Verum	weiblich	54	stark	keiner
8	Placebo	weiblich	39	stark	leicht
9	Placebo	männlich	27	leicht	stark
10	Verum	weiblich	46	stark	keiner
11	Placebo	weiblich	56	leicht	keiner
12	Verum	männlich	33	leicht	keiner
13	Verum	männlich	40	leicht	leicht
14	Placebo	männlich	47	stark	stark
15	Verum	weiblich	55	stark	leicht
16	Placebo	männlich	23	leicht	leicht
17	Verum	männlich	38	stark	keiner
18	Placebo	weiblich	46	stark	keiner
19	Verum	weiblich	31	leicht	keiner
20	Placebo	männlich	27	leicht	leicht

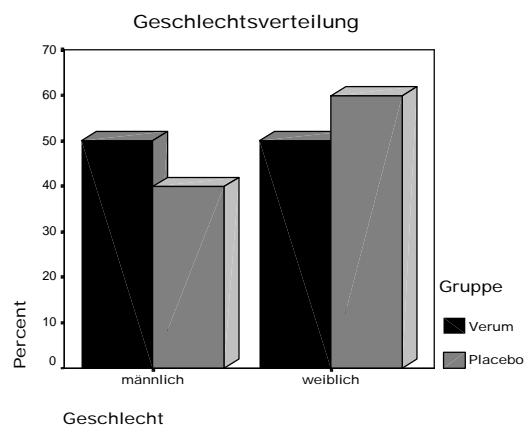
In der Tabelle sind für jeden Patienten die Werte von 5 Merkmalen aufgeführt: Behandlungsgruppe, Geschlecht, Alter, Husten zu Beginn und Husten am Ende der Studie. Diese Merkmale haben eine unterschiedliche Struktur, die bei der Datenbeschreibung und statistischen Auswertung zu berücksichtigen ist. Die Merkmale Behandlungsgruppe und Geschlecht sind **nominal** oder **kategorial**, da für die möglichen Werte nur eine Bezeichnung oder Kategorie vorgegeben ist. Bei der Beschrei-

bung der Daten kategorialer Merkmale läßt sich lediglich auszählen, wie häufig die einzelnen Kategorien in der Gesamtheit der Beobachtungen vorkommen. Dividiert man die Anzahl r der Fälle, bei denen eine bestimmte Kategorie beobachtet wurde, durch die Gesamtzahl n der beobachteten Fälle, dann erhält man die (relative) Häufigkeit h der Kategorie: $h=r/n$; oft wird die Häufigkeit – nach Multiplikation mit 100 – als Prozentzahl angegeben. Die Verteilung der Häufigkeiten über alle möglichen Kategorien des Merkmals nennt man die **Häufigkeitsverteilung** des Merkmals. Die Summe der Häufigkeiten über alle Kategorien eines Merkmals muß stets 1 (bzw. 100%) ergeben. Die folgende Tabelle (erstellt mit dem Programmsystem SPSS) zeigt die Häufigkeitsverteilung des Geschlechts für die beiden Behandlungsgruppen. Eine solche Tabelle, in der die Häufigkeitsverteilung eines Merkmals (Geschlecht) für alle Kategorien eines zweiten Merkmals (Behandlungsgruppe) dargestellt wird, nennt man eine **Kreuztabelle**.

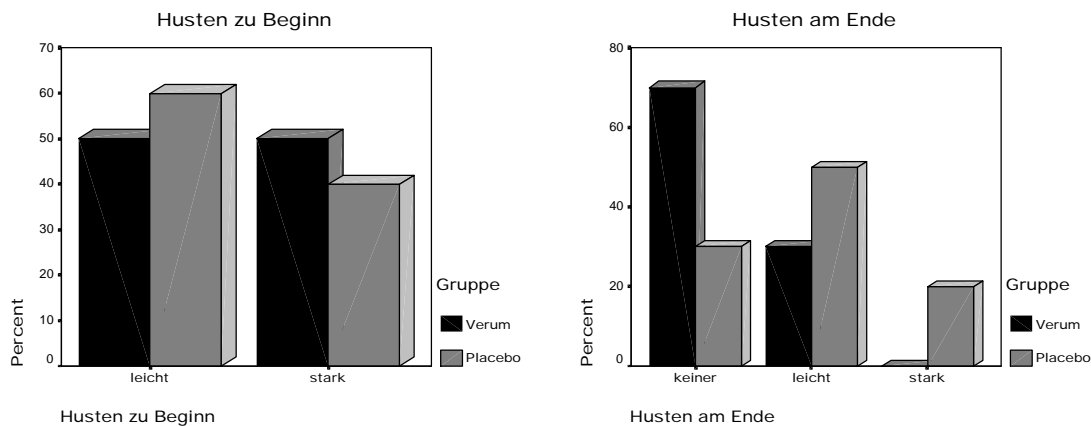
Geschlecht * Gruppe Kreuztabelle

			Gruppe		Gesamt
			Verum	Placebo	
Geschlecht	männlich	Anzahl	5	4	9
		% von Gruppe	50,0%	40,0%	45,0%
	weiblich	Anzahl	5	6	11
		% von Gruppe	50,0%	60,0%	55,0%
Gesamt		Anzahl	10	10	20
		% von Gruppe	100,0%	100,0%	100,0%

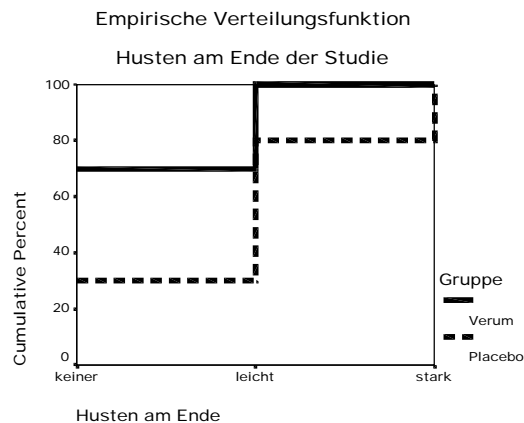
In der Verum-Gruppe sind 50% Frauen und 50% Männer, in der Placebo-Gruppe 60% Frauen und 40% Männer. Die Häufigkeitsverteilung kann in einem Säulendiagramm (Histogramm) grafisch dargestellt werden. Dabei wird über jede Kategorie eine Säule aufgetragen, deren Höhe dem Wert der Häufigkeit entspricht:



Bei den Merkmale 'Husten zu Beginn' und 'Husten am Ende' ist die Rangordnung: 'keiner', 'leicht', 'stark' vorgegeben. Man nennt diese Merkmale deshalb auch **ordinale Merkmale**. Die Daten dieser Merkmale werden primär auch mit der Häufigkeitsverteilung beschrieben. Die folgenden Abbildungen zeigen die Ausprägungen des Hustens zu Beginn und am Ende der Studie für beide Gruppen.



Zur Beschreibung der Verteilung ordinaler Daten in der Gesamtheit der erfassten Daten kann auch die **empirische Verteilungsfunktion** benutzt werden. Diese ordnet jeder Kategorie die Häufigkeit zu, mit der Kategorien, die schlechter oder höchstens gleich dieser Kategorie sind, vorkommen (Summenhäufigkeit, kumulierte Häufigkeit). Grafisch wird die empirische Verteilungsfunktion als Treppenkurve über die geordneten Merkmalkategorien dargestellt, die bei jeder Kategorie um die Häufigkeit springt, mit der diese Kategorie vorkommt. Die folgende Abbildung zeigt die empirische Verteilungsfunktionen für die Ausprägungen des Hustens am Ende der Studie in beiden Behandlungsgruppen. Die Verteilung der Verum-Gruppe liegt stets über der Verteilung der Placebo-Gruppe. Das bedeutet, daß der Husten am Ende der Studie in der Verum-Gruppe eher geringer ausgeprägt war als in der Placebo-Gruppe.



Das Alter wird durch eine Zahl (vollendete Lebensjahre) ausgedrückt und ist somit ein **quantitatives Merkmal**. Da man das Alter statt in vollendeten Lebensjahren auch in Tagen oder Stunden ausdrücken kann, ist das Alter ein 'stetiges' quantitatives Merkmal, bei dem im Prinzip jeder positive, reelle Zahlenwert vorkommen kann. Zur Erstellung der Häufigkeitsverteilung müssen die stetigen Merkmalwerte in Klassen eingeteilt werden; z.B. in Jahresklassen oder besser in 10-Jahresklassen. Bei quantitativen Merkmalen kann die Verteilung auch durch

statistische Kenngrößen beschrieben werden. Die wichtigsten Kenngrößen sind:

$$\text{Mittelwert (arithmetisches Mittel): } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{Varianz: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right)$$

$$\text{Standardabweichung } s = \sqrt{s^2}$$

Minimum x_{\min} = kleinster Wert

Maximum x_{\max} = größter Wert

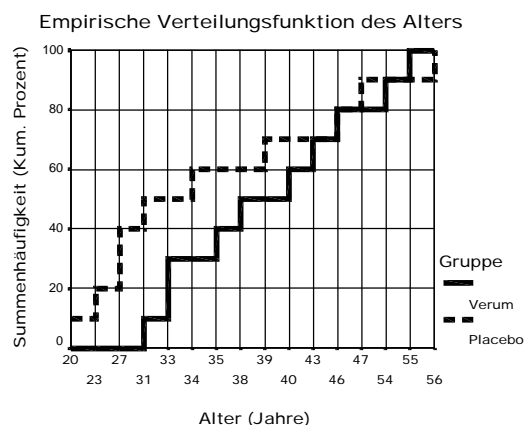
Spannweite R: = $x_{\max} - x_{\min}$

Für die Altersangaben der Erkältungsstudie erhält man folgende Kenngrößen:

Kenngröße	Gruppe: Verum	Gruppe: Placebo
Mittelwert \bar{x}	40.8 Jahre	35.0 Jahre
Varianz s^2	74.178 Jahre ²	137.333 Jahre ²
Standardabweichung s	8.61 Jahre	11.72 Jahre
Minimum x_{\min}	31 Jahre	20 Jahre
Maximum x_{\max}	55 Jahre	56 Jahre
Spannweite R	24 Jahre	36 Jahre

Mittelwert, Minimum und Maximum kennzeichnen die Lage, Varianz, Standardabweichung und Spannweite die Variabilität (Streuung) der Daten.

Die empirischen Verteilungsfunktion $\hat{F}(x)$ gibt für jeden Merkmalwert x die Häufigkeit an, mit der Merkmalwerte $\leq x$ vorkommen. Sie ist eine Treppenkurve, die bei jedem beobachteten Wert x_i um $1/n$ springt; falls der Wert x_i k-mal vorkommt, beträgt die Sprunghöhe k/n . Die Abbildung zeigt diese Funktionen für das Alter:



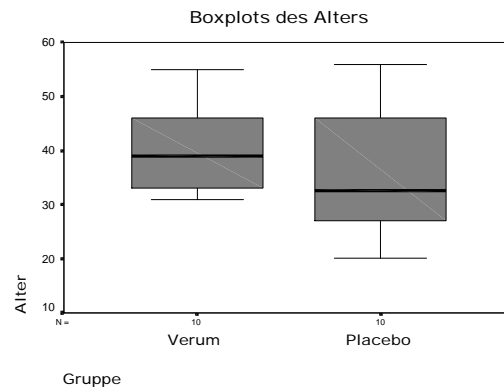
Aus der empirischen Verteilungsfunktion können als weitere Kenngrößen die **q-Quantilen** x_q hergeleitet werden. Die q-Quantile x_q ist für einen gegebenen Wert q ($0 < q < 1$) der kleinste Merkmalwert, für den gilt: $\hat{F}(x_q) \geq q$; d.h. der Anteil der Daten,

die kleiner oder gleich x_q sind, soll mindestens q betragen. Die q -Quantile ist der Merkmalwert, bei dem in der grafischen Darstellung der empirischen Verteilungsfunktion eine Parallele zur Merkmalachse (x -Achse) durch den Wert q der Ordinate (Summenhäufigkeit) die empirische Verteilungsfunktion schneidet. Falls diese Parallele genau auf eine Stufe trifft, wird der x -Wert in der Mitte der Stufe als q -Quantile genommen. Mit den der Größe nach geordneten Daten $x_{[1]}, x_{[2]}, \dots, x_{[n]}$ (wobei $x_{[1]}$ der kleinste und $x_{[n]}$ der größte Meßwert ist) kann die q -Quantile nach folgender Regel berechnet werden: Wenn $n \cdot q = k + \text{Rest}$ (d.h. das Produkt $n \cdot q$ keine ganze Zahl ist, sondern sich in der Dezimalschreibweise als eine ganze Zahl k und einem Dezimalrest hinter dem Komma darstellt), dann ist $x_q = x_{[k+1]}$. Für $n \cdot q = k$ ist $x_q = (x_{[k]} + x_{[k+1]})/2$. Die 0.5-Quantile nennt man auch den **Median**, die 0.25-Quantile die untere und die 0.75-Quantile die obere **Quartile**. Die Differenz zwischen der oberen und unteren Quartile ist der 'Interquartile Range' IQR.

Für das Alter ergeben sich in den jeweiligen Gruppen folgende Werte:

	Verum-Gruppe	Placebo-Gruppe
25%-Quartile $x_{0.25}$	33 Jahre	27 Jahre
Median $x_{0.5}$	39 Jahre	33 Jahre
75%-Quartile $x_{0.75}$	46 Jahre	46 Jahre
Interquartile Range IQR	13 Jahre	19 Jahre

Mit diesen Quantilen kann die Verteilung eines quantitativen Merkmals durch einen **Boxplot** veranschaulicht werden:



Der untere Wert der Box entspricht der unteren Quartile $x_{0.25}$, der obere der oberen Quartile $x_{0.75}$; der Strich in der Box entspricht dem Median $x_{0.5}$. Die unteren und oberen Balken kennzeichnen den kleinsten und größten Stichprobenwert (soweit diese keine 'Ausreißer' oder 'Extremwerte' sind). Werte, die um mehr als das 1,5-fache der Länge der Box die untere Grenze unterschreiten oder die obere Grenze überschreiten, werden als 'Ausreißer' gekennzeichnet; Werte, die um mehr als das 3-fache der Länge die Grenzen der Box über- oder unterschreiten als 'Extremwerte'.

Die Boxplots des Alters in beiden Behandlungsgruppen zeigen, daß das Alter der Patienten der Verum-Gruppe einen größeren Median aber eine geringere Variabilität (d.h. geringere IQR) hat als das der Patienten der Placebo-Gruppe.

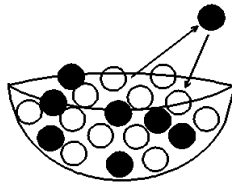
Grundgesamtheit und Wahrscheinlichkeit

Ziel der statistischen Auswertung ist es, aus den gemessenen oder beobachteten Daten Aussagen über zukünftige Ereignisse zu machen; z.B. über die Wirkungen, die bei zukünftigen Anwendungen der Arzneimittel zu erwarten sind. Hierfür wird das Konzept der **Grundgesamtheit** eingeführt: Man stellt sich vor, daß die Messung oder Beobachtung, die zu den Daten geführt hat, unendlich oft wiederholt wird. Die Gesamtheit der dabei erhaltenen Ergebnisse bildet die Grundgesamtheit. Die Grundgesamtheit umfaßt also sowohl die tatsächlich beobachteten Daten als auch die Daten, die bei zukünftigen Beobachtungen (z.B. bei zukünftigen Behandlungen mit den Arzneimitteln) zu erwarten sind. Die tatsächlich gemessenen oder beobachteten Daten nennt man eine **Stichprobe**. Es wird angenommen, daß die Daten der Stichprobe 'zufällig' und 'unabhängig' aus der Grundgesamtheit entnommen sind.

Urnschema

Alle Ergebnisse der Grundgesamtheit werden auf Kugeln geschrieben; die Kugeln werden in eine Urne geworfen und gut durchmischt.

Bei einem Ereignis wird eine Kugel zufällig entnommen und das Ergebnis notiert. Die Kugel wird dann wieder zurückgelegt (Unabhängigkeit).



In der Grundgesamtheit sind i.a. die verschiedenen möglichen Werte eines beobachteten Merkmals verschieden häufig enthalten. Die relative Häufigkeit, mit der ein Merkmalwert x oder eine Menge A von Merkmalwerten in der Grundgesamtheit vorkommen, nennt man die **Wahrscheinlichkeit** $P(x)$ bzw. $P(A)$ für diesen Wert oder diese Menge. Sie ist ein Maß für die Zuverlässigkeit, mit der der Wert x oder die Menge A bei **zukünftigen** Beobachtung zu erwarten sind. Das Ergebnis einer Beobachtung nennt man auch ein Ereignis. Ist das Ereignis bereits eingetreten (d.h. wurde ein Wert x_0 beobachtet), dann besteht keine Unsicherheit bezüglich des Ergebnisses und es ist daher nicht sinnvoll, eine Wahrscheinlichkeit zuzuordnen.

Für Wahrscheinlichkeiten gelten folgende Regeln (Axiome):

- Die Wahrscheinlichkeit für unmögliche Merkmalwerte ist 0 (da diese Werte in der Grundgesamtheit nicht vorkommen).
- Die Wahrscheinlichkeit, irgendeinen möglichen Merkmalwert zu beobachten, ist 1.
- Die Wahrscheinlichkeit, daß entweder der Merkmalwert x_1 oder ein davon verschiedener Wert x_2 beobachtet wird (bzw. daß Werte aus der Menge A_1 oder aus einer davon verschiedenen (elementfremden) Menge A_2 beobachtet werden), ist gleich der Summe aus den beiden Wahrscheinlichkeiten.

Werden zwei verschiedene Ereignisse beobachtet (z.B. das Geschlecht und die Haarfarbe einer Person oder die Augenzahlen bei zwei Würfeln mit einem Würfel), dann

interessiert die **Produktwahrscheinlichkeit**; d.h. die Wahrscheinlichkeit $P(x_1, x_2)$ für die möglichen Werte-Kombinationen (x_1, x_2) dieser beiden Ereignisse. Diese kann interpretiert werden als die relative Häufigkeit, mit der die Kombination (x_1, x_2) in der Grundgesamtheit aller möglichen Wiederholungen der Beobachtung beider Ereignisse vorkommt. Die **bedingte Wahrscheinlichkeit** $P(x_2|x_1)$ ist die Wahrscheinlichkeit, daß beim zweiten Ereignis x_2 beobachtet wird, wenn beim ersten x_1 beobachtet wurde. Dies entspricht der relativen Häufigkeit, mit der in der Teilgesamtheit der Kombinationen, bei denen im ersten Ereignis x_1 beobachtet wurde, im zweiten Ereignis x_2 beobachtet wird. Dies ergibt natürlich nur dann Sinn, wenn Werte-Kombinationen mit x_1 vorkommen. Die Wahrscheinlichkeit $P(x_1)$, im ersten Ereignis x_1 zu beobachten, gleichgültig welches Ergebnis das zweite Ereignis hat, muß also größer als 0 sein. Diese Wahrscheinlichkeit nennt man die **Randwahrscheinlichkeit**. Für die bedingte Wahrscheinlichkeit gilt dann:

$$P(x_2|x_1) = P(x_1, x_2)/P(x_1)$$

Analog wird die bedingte Wahrscheinlichkeit $P(x_1|x_2)$ als $P(x_1, x_2)/P(x_2)$ definiert (vorausgesetzt, daß $P(x_2) > 0$ ist). Die beiden Ereignisse sind (stochastisch) **unabhängig**, wenn die bedingte Wahrscheinlichkeit $P(x_2|x_1)$ nicht von x_1 abhängt und für jeden möglichen Wert von x_1 gleich der Randwahrscheinlichkeit $P(x_2)$ ist. Es folgt dann sofort aus der Definition der bedingten Wahrscheinlichkeit, daß dann auch die bedingte Wahrscheinlichkeit $P(x_1|x_2)$ für jedes x_2 gleich der Randwahrscheinlichkeit $P(x_1)$ ist und als Kriterium für die Unabhängigkeit gilt:

$$P(x_1, x_2) = P(x_1) \cdot P(x_2).$$

Die Verteilung der Wahrscheinlichkeiten über die möglichen Merkmalwerte x nennt man die **Wahrscheinlichkeitsverteilung**. Ziel der statistischen Auswertung ist es, aus den erfaßten Daten Aussagen über die Wahrscheinlichkeitsverteilung zu machen, die den Daten zugrunde liegt. Der einfachste Fall liegt vor, wenn ein Merkmal entweder vorhanden oder nicht vorhanden ist (binäres Merkmal, z.B. geheilt – nicht geheilt). Die Grundgesamtheit ist vollständig durch die Wahrscheinlichkeit π für das Vorhandensein des Merkmals gekennzeichnet, da die Wahrscheinlichkeit für das Nicht-Vorhandensein gleich $1 - \pi$ ist. Die Wahrscheinlichkeiten für alle möglichen Merkmalwerte müssen sich stets zu 1 aufsummieren. Sind nur endlich viele Merkmalwerte möglich (z.B. Geschlecht, Augenzahl bei einem Würfel), dann ist die Grundgesamtheit durch die Angabe der Wahrscheinlichkeiten für jeden Merkmalwert charakterisiert. Bei unendlich vielen möglichen Merkmalwerten, insbesondere bei stetigen quantitativen Merkmalen, die jeden reellen Zahlenwert in einem Intervall annehmen können (z.B. Alter, Größe, Gewicht), wird die Verteilung der Wahrscheinlichkeiten in der Grundgesamtheit durch die **Verteilungsfunktion $F(x)$** charakterisiert. Diese gibt zu jedem möglichen Merkmalwert x die Wahrscheinlichkeit an, mit der Werte, die kleiner oder höchstens gleich x sind, in der Grundgesamtheit vorkommen. Man bezeichnet die (quantitativen) Werte der Grundgesamtheit auch als **Zufallsvariable X** und spricht davon, daß die Verteilungsfunktion $F(x)$ die Wahrscheinlichkeit für $X \leq x$ angibt: $F(x) = P(X \leq x)$. Die Verteilungsfunktion ist für Werte, die kleiner als der kleinste mögliche Wert sind, gleich 0 (z.B. wenn nur positive Werte möglich sind, für $x \leq 0$; wenn alle reellen Werte möglich sind, für $x = -\infty$). Sie steigt mit zunehmenden x -Werten monoton auf 1 an. Bei stetigen quantitativen Merkmalen kann die Verteilung der Wahrscheinlichkeiten auch durch die **Verteilungsdichte $f(x)$** charakterisiert werden. Mathematisch ist $f(x)$ das Differential von $F(x)$: $f(x) = dF(x)/dx$. Für eine kleines Intervall $(x, x+dx)$ ist $f(x) \cdot dx$ die Wahrscheinlichkeit, einen Merkmalwert aus diesem Intervall zu beobachten. Die Verteilungsdichte hat oft

eine 'Glockenform'; d.h. sie hat bei einem bestimmten Merkmalwert (dem Modalwert) ein Maximum und fällt für größere oder kleinere Merkmalwerte auf 0 ab.

Wahrscheinlichkeiten bzw. Verteilungsfunktionen können nicht direkt ausgezählt werden, da die Grundgesamtheit unendlich viele Werte enthält. Man kann aus theoretischen Überlegungen für die Verteilungsfunktionen einen bestimmten Typ annehmen, der von unbekanntem Kenngrößen, den **Parametern**, abhängt. Parameter werden im folgenden meist mit kleinen griechischen Buchstaben bezeichnet. Wichtige Parameter für die Verteilung quantitativer Zufallsgrößen sind:

- **Mittelwert μ**

Das ist die mit den Wahrscheinlichkeiten bzw. der Verteilungsdichte gewichtete Summe bzw. das Integral der möglichen Merkmalwerte: $\mu = \sum_j x_j P(x_j) = \int x f(x) dx$.

Der Mittelwert wird auch 'Erwartungswert' der Zufallsgröße X genannt.

- **Quantilen x_q**

Das sind die Merkmalwerte, die den Anteil q ($0 < q < 1$) der Grundgesamtheit nach oben begrenzen; d.h. für die gilt: $F(\xi_q) = q$. Die 0.5-Quantile $\xi_{0.5}$, die bei stetigen Merkmalwerten die Eigenschaft hat, daß genau die Hälfte der Grundgesamtheit kleiner oder gleich $\xi_{0.5}$ ist, nennt man den **Median**. Die 0.25-Quantile $\xi_{0.25}$ heißt untere Quartile, die 0.75-Quantile $\xi_{0.75}$ ober Quartile.

- **Varianz s^2 und Standardabweichung s**

Die Varianz ist der Erwartungswert der quadratischen Abweichung der Zufallsgröße X vom Mittelwert μ : $\sigma^2 = \sum_j (x_j - \mu)^2 P(x_j) = \int (x - \mu)^2 f(x) dx$.

Die Standardabweichung σ ist die positive Wurzel aus der Varianz: $\sigma = \sqrt{\sigma^2}$.

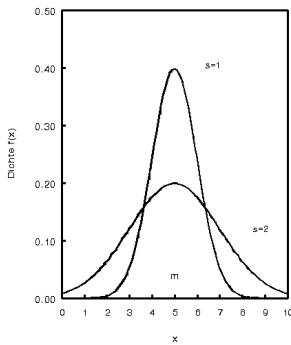
Eine große Rolle spielt in der Statistik die Klasse der **Normalverteilungen**. Sie wurde u.a. von C.F. Gauß bei der Beschreibung von Meßfehlern bei astronomischen Messungen hergeleitet. Er hat dabei angenommen, daß der von einem Stern ausgesandte Lichtstrahl beim Eintritt in die Atmosphäre durch Streuung an den Luftpartikeln abgelenkt wird, wobei sich entgegengesetzte Ablenkungen im Mittel aufheben. Der beobachtete Einfallswinkel des Lichtstrahl in das Objektiv eines Fernrohrs ist daher Realisation einer Zufallsgröße, die durch Summation vieler kleiner, unabhängiger Beträge zustande gekommen ist.

Die Normalverteilung ist durch den Mittelwert μ und die Standardabweichung σ vollständig gekennzeichnet. Die Formeln und der Verlauf von Verteilungsfunktion und Verteilungsdichte der Normalverteilung sind im folgenden gezeigt:

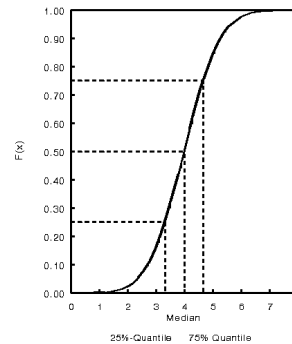
$$\text{Verteilungsfunktion } F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{s}} e^{-\frac{1}{2}t^2} dt = F\left(\frac{x-\mu}{s}\right)$$

$$\text{Verteilungsdichte } = f(x) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{s^2}} = \frac{1}{s} f\left(\frac{x-\mu}{s}\right)$$

Dichten der Normalverteilung



Verteilungsfunktion F(x)
Mittelwert=4, Standardabweichung=1



Die Verteilung ist symmetrisch um den Mittelwert μ ; d.h. für jedes positive a gilt: im Intervall $(\mu-a, \mu)$ liegt derselbe Anteil von Werten der Grundgesamtheit wie im Intervall $(\mu, \mu+a)$. Im Bereich $(\mu-\sigma, \mu+\sigma)$ liegen etwa 68% und im Bereich $(\mu-2\sigma, \mu+2\sigma)$ etwa 95% aller Werte der Grundgesamtheit.

Eine Normalverteilung mit Mittelwert $\mu=0$ und Standardabweichung $\sigma=1$ nennt man die **Standard-Normalverteilung**. Durch die Transformation:

$$z = \frac{x - \mu}{\sigma}$$

wird eine Normalverteilung der Zufallsgröße X mit Mittelwert μ und Standardabweichung σ auf die Standard-Normalverteilung der Zufallsgröße Z zurückgeführt.

Verteilungsfunktion und Verteilungsdichte der Standard-Normalverteilung werden mit $\Phi(z)$ und $\phi(z)$ bezeichnet. Die Umkehrfunktion zu $\Phi(z)$, d.h. die Funktion die jedem Wert $\Phi=q$ den entsprechenden z -Wert zuordnet, wird mit $z=\Psi(q)$ bezeichnet. Sie gibt die Quantilen z_q der Standard-Normalverteilung an. Die folgende Tabelle zeigt zu verschiedenen q -Werten die entsprechenden Quantilen z_q :

Die Quantilen z_q der Standard-Normalverteilung

q	z_q	q	z_q
0.01	-2.326	0.5	0
0.025	-1.960	0.6	0.253
0.05	-1.645	0.7	0.524
0.1	-1.282	0.8	0.842
0.2	-0.842	0.9	1.282
0.3	-0.524	0.95	1.645
0.4	-0.253	0.975	1.960
0.5	0	0.99	2.326

Wegen der Symmetrie der Normalverteilung gilt: $\Phi(z) = 1-\Phi(-z)$ und $z_q = -z_{1-q}$.

Schätzwerte (Statistiken) und Konfidenzintervalle

Die gemessenen oder beobachteten Werte x_i werden als zufällig und unabhängig aus der Grundgesamtheit herausgegriffen angesehen; sie sind **Realisationen** der Zufallsgröße X . Sie repräsentieren die Grundgesamtheit im dem Sinne, daß bei wiederholter Beobachtung die Realisationen genauso wie die Zufallsgröße variieren. Die z.B. in einer Studie beobachteten n Werte x_1, \dots, x_n bilden eine Stichprobe aus der Grundgesamtheit. Mit den Werten dieser Stichprobe können Aussagen über die Wahrscheinlichkeitsverteilung und ihre Parameter gewonnen werden. Dies geschieht dadurch, daß mit den Stichprobenwerten x_1, \dots, x_n **Schätzwerte** oder **Statistiken** für die unbekanntenen Werte der Parameter berechnet werden:

- Die Häufigkeit h , mit der in einer Stichprobe von n Beobachtungen das Ergebnis x vorkommt, ist Schätzwert der Wahrscheinlichkeit $\pi=P(x)$ für das Ergebnis x .
- Die empirische Verteilungsfunktion $\hat{F}(x)$ ist Schätzwert für die Verteilungsfunktion $F(x)$.
- Die aus der empirischen Verteilungsfunktion $\hat{F}(x)$ ermittelten Quantilen x_q sind Schätzwerte für die Quantilen ξ_q .
- Das arithmetische Mittel $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$ ist Schätzwert von μ .
- Die Stichprobenvarianz $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ ist Schätzwert der Varianz σ^2
- Die Wurzel aus s^2 , d.i. $s = \sqrt{s^2}$, ist Schätzwert für die Standardabweichung σ .

Da die Schätzwerte aus einer zufälligen Auswahl aus der Grundgesamtheit gewonnen werden, weichen sie mehr oder minder stark vom Parameterwert ab, den sie schätzen sollen. Um diese Abweichung genauer kennzeichnen zu können, führt man das Konzept der **Stichprobengesamtheit** ein. Man stellt sich vor, daß die Stichprobenentnahme aus der Grundgesamtheit unendlich oft wiederholt wird und aus jeder Stichprobe der Schätzwert (z.B. \bar{x}) berechnet wird. Man erhält so eine Gesamtheit von Schätzwerten, die zufällig variieren und somit eine neue Grundgesamtheit, die Stichprobengesamtheit, bilden. Der aus der konkret vorliegenden Stichprobe berechnete Schätzwert ist eine Realisation dieser Zufallsgröße, die 'Schätzgröße' genannt werden soll (die Bezeichnung 'Schätzfunktion' ist ebenfalls gebräuchlich). Die relativen Häufigkeiten der möglichen Schätzwerte in der Stichprobengesamtheit bilden die Wahrscheinlichkeitsverteilung der Schätzgröße (die Schätzverteilung). Diese Verteilung hängt von der Verteilung der Beobachtungswerte (und damit auch vom zu schätzenden Parameterwert) und vom Stichprobenumfang n ab. Mit ihr werden die Abweichungen der Schätzwerte von dem zu schätzenden Parameterwert charakterisiert. Als wichtige Kriterien werden der Mittelwert (Erwartungswert) und die Standardabweichung der Schätzgröße benutzt.

Bei vielen Schätzgrößen, z.B. der Häufigkeit (als Schätzgröße für die Wahrschein-

lichkeit), dem arithmetischen Mittel der Stichprobenwerte (als Schätzgröße für den Mittelwert μ) und der Stichprobenvarianz s^2 (als Schätzgröße für die Varianz σ^2), stimmt für jeden Stichprobenumfang n der Mittelwert der Schätzverteilung mit dem zu schätzenden Parameterwert überein. Man nennt solche Schätzgrößen **unverzerrt** (unbiased) oder **erwartungstreu**. Andere Schätzgrößen, wie z.B. die Standardabweichung s der Stichprobe und die Stichproben-Quantilen x_q sind zwar für kleine Stichprobenumfänge noch nicht unverzerrt; ihr Mittelwert weicht aber um so weniger vom zu schätzenden Parameterwert ab, je größer der Stichprobenumfang n ist. Man nennt solche Schätzgrößen **asymptotisch unverzerrt**.

Die Standardabweichung einer Schätzgröße nennt man den **Standardfehler**. Er charakterisiert bei unverzerrten Schätzgrößen die Variation der Schätzwerte um den unbekanntem Parameterwert. Schätzgrößen, die bei gegebenem Stichprobenumfang n von allen zulässigen Schätzverfahren den geringsten Standardfehler haben, nennt man **effizient**. Bei den gebräuchlichen Schätzgrößen ist der Standardfehler proportional zu $1/\sqrt{n}$; d.h. er nimmt mit zunehmendem Stichprobenumfang n ab. Das bedeutet, daß sich unverzerrte Schätzgrößen mit zunehmendem Stichprobenumfang immer enger um den zu schätzenden Wert des Parameters konzentrieren. Diese Eigenschaft nennt man **Konsistenz**.

Für die Standardfehler des arithmetischen Mittels \bar{x} und der Häufigkeit h gilt:

$$\begin{aligned} \text{Standardfehler des Mittelwertes } \bar{x}: \quad \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ \text{Standardfehler der Häufigkeit } h: \quad \sigma_h &= \frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}} \end{aligned}$$

Praktisch ist die Standardabweichung σ nicht bekannt. Einen Schätzwert für den Standardfehler des Mittelwertes erhält man, wenn in der Formel σ durch den Schätzwert s ersetzt wird.

Für das Alter der Patienten der Erkältungsstudie erhält man folgende Schätzwerte für den Mittelwert und seinen Standardfehler:

Verum-Gruppe:

$$n=10; \bar{x} = 40.8 \text{ Jahre}; s = 8.61 \text{ Jahre}; s_{\bar{x}} = 8.61 / \sqrt{10} = 2.72 \text{ Jahre}$$

Placebo-Gruppe:

$$n=10; \bar{x} = 35.0 \text{ Jahre}; s = 11.72 \text{ Jahre}; s_{\bar{x}} = 11.72 / \sqrt{10} = 3.71 \text{ Jahre}$$

Da der Standardfehler vom Stichprobenumfang n abhängt, läßt sich n so festlegen, daß der Standardfehler eine vorgegebene Grenze nicht überschreitet. Verlangt man z.B., daß der Standardfehler des Mittelwerts nicht größer als $\frac{1}{2}$ der Standardabweichung σ sein soll, dann ist ein Stichprobenumfang $n=4$ ausreichend. Soll der Standardfehler nicht größer als 10% der Standardabweichung sein, dann ist bereits ein Stichprobenumfang $n=100$ erforderlich, und bei der Forderung nach einem Standardfehler von 1% der Standardabweichung sind 10000 Beobachtungen erforderlich.

Bei der Einführung der Normalverteilung wurde bereits darauf hingewiesen, daß sich

diese Verteilung einstellt, wenn die beobachtete Zufallsgröße als Summe vieler unabhängiger Zufallseinflüsse dargestellt werden kann. Dieses Resultat nennt man den **Hauptgrenzwertsatz**. Viele Schätzwerte sind Summen der Stichprobenwerte oder von Funktionen der Stichprobenwerte (z.B. Abweichungsquadrate). Daher kann bei nicht zu kleinem Stichprobenumfang n nach dem Hauptgrenzwertsatz für die Verteilung dieser Schätzgrößen eine Normalverteilung angenommen werden, auch wenn die Stichprobenwerte keine Normalverteilung haben. So kann das arithmetischen Mittel \bar{x} als normal verteilt mit dem Mittelwert μ und der Standardabweichung σ/\sqrt{n} angesehen werden, auch wenn die Daten nicht normal verteilt sind. Bei einer nicht zu extrem schiefen Verteilung der Stichprobenwerte ist die Normalverteilung des arithmetischen Mittels bereits ab einem geringen Stichprobenumfang von 6-8 in recht guter Näherung gegeben.

Schätzwerte geben für den unbekanntem Wert eines Parameters eine Zahl, gewissermaßen einen 'Punkt' an. Man nennt sie deshalb auch 'Punktschätzer'. Eine andere Möglichkeit, aus den Daten Aussagen über den unbekanntem Wert des Parameters zu bekommen, bietet die Methode der **Konfidenzintervalle** zu einer vorgegebenen Konfidenzwahrscheinlichkeit (meist 95%). Ein Konfidenzintervall ist ein aus den Daten berechnetes Intervall, das den unbekanntem Wert des Parameters mit der vorgegebenen Konfidenzwahrscheinlichkeit überdeckt; d.h., wenn die Stichprobenentnahme unendlich oft wiederholt wird und jedesmal das Konfidenzintervall berechnet wird, dann werden diese Intervalle mit der durch die Konfidenzwahrscheinlichkeit vorgegebenen Häufigkeit den 'wahren' Wert des Parameters enthalten.

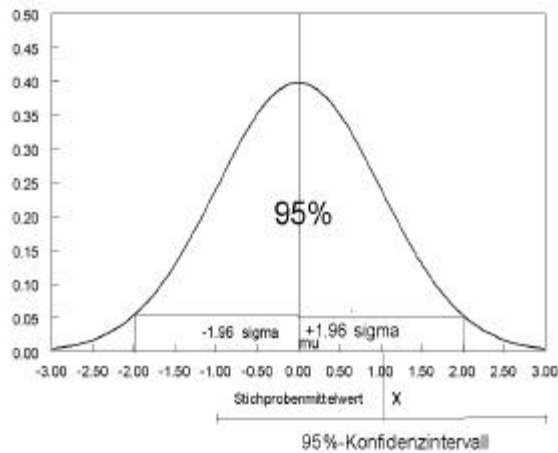
Für den 'wahren' Wert des Mittelwertes μ läßt sich mit dem Stichproben-Mittel \bar{x} ein Konfidenzintervall zur Konfidenzwahrscheinlichkeit 95% leicht konstruieren, wenn die Standardabweichung σ bekannt ist. Wie oben ausgeführt wurde, ist \bar{x} Realisation der Zufallsgröße aller bei unendlicher Wiederholung der Stichprobe erhaltenen Mittelwerte. Diese Zufallsgröße kann (bei nicht zu kleinem Stichprobenumfang n) als normal verteilt mit dem Mittelwert μ und der Standardabweichung σ/\sqrt{n} angesehen werden. Nach den oben genannten Eigenschaften der Normalverteilung sind im Intervall $(\mu - 2\sigma/\sqrt{n}, \mu + 2\sigma/\sqrt{n})$ etwa 95% aller \bar{x} -Werte der Stichprobengesamtheit zu erwarten. Daraus folgt, daß das Intervall:

$$\left(\bar{x} - 2 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \frac{\sigma}{\sqrt{n}}\right)$$

mit der Wahrscheinlichkeit von etwa 95% den 'wahren' Mittelwert μ enthält (nämlich immer dann, wenn das Stichprobenmittel \bar{x} im Intervall $\mu \pm 2\sigma/\sqrt{n}$ liegt).

Wird statt 95% eine andere Konfidenzwahrscheinlichkeit, die allgemein mit $1-\alpha$ bezeichnet werden soll (wobei α den Anteil des 'Nicht-Überdeckens' angibt) vorgegeben, dann muß die Zahl 2 durch die $(1-\alpha/2)$ -Quantile der Normalverteilung ersetzt werden. Die $(1-\alpha/2)$ -Quantile (statt der $(1-\alpha)$ -Quantile) ist zu nehmen, da mit Wahrscheinlichkeit $\alpha/2$ der 'wahre' Wert μ größer als die obere Grenze und mit Wahrscheinlichkeit $\alpha/2$ kleiner als die untere Grenze des Konfidenzintervalls sein kann.

Konstruktion eines 95%-Konfidenzintervalls



Im allgemeinen ist σ nicht bekannt und wird durch die Standardabweichung s der Stichprobe geschätzt. Damit muß aber – neben der Variabilität von \bar{X} - auch die Variabilität von s berücksichtigt werden, die vom Stichprobenumfang n abhängt. Dies geschieht dadurch, daß s/\sqrt{n} mit der $(1-\alpha/2)$ -Quantile $t_{1-\alpha/2, n-1}$ der zentralen t -Verteilung mit $n-1$ Freiheitsgraden multipliziert wird. Die zentrale t -Verteilung wurde 1908 von W.S. Gosset (der unter dem Pseudonym 'Student' schrieb) hergeleitet. Es ist die Verteilung der Zufallsgröße $t = \frac{\bar{x} - \mu}{s} \sqrt{n}$ wobei \bar{x} und s als Zufallsgrößen

aufgefaßt werden, die aus normal verteilten Daten mit Mittelwert μ und Standardabweichung σ berechnet werden. Der Zähler hat eine Normalverteilung mit dem Mittelwert 0 und der Standardabweichung σ , der Nenner ist davon unabhängig wie die Wurzel aus $\sigma^2 \chi^2 / (n-1)$ verteilt, wobei χ^2 die Summe aus $n-1$ Quadraten von unabhängigen, normal verteilten Zufallsgrößen mit Mittelwert 0 und Standardabweichung 1 ist (χ^2 -Verteilung mit $n-1$ Freiheitsgraden). Die Zahl der Freiheitsgrade (auch mit df (degrees of freedom) bezeichnet) entspricht dem Nenner bei der Berechnung von s^2 ; es ist die Zahl der 'freien' Informationen, die bei der Bildung der Summe der Abweichungsquadrate noch verbleibt, wenn das arithmetische Mittel (entspricht 1 Freiheitsgrad) berechnet ist. Das $(1-\alpha)$ -Konfidenzintervall für μ lautet somit:

$$\left(\bar{x} - t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2, n-1} \frac{s}{\sqrt{n}} \right)$$

Bei der Erkältungsstudie liegen sowohl für die Verum- als auch für die Placebo-Gruppe $n=10$ Altersdaten vor. Die 0.975-Quantile der t -Verteilung mit 9 Freiheitsgraden ist 2.262. Damit ergeben sich folgende 0.95-Konfidenzintervalle für den Mittelwert μ :

Verum-Gruppe:

$\bar{x}=40,8$; $s=8.61$ Jahre; 95%-Konfidenzintervall: (34.64 – 46.96 Jahre)

Placebo-Gruppe:

$\bar{x}=35,0$; $s=11.72$ Jahre; 95%-Konfidenzintervall: (26.62 – 43.38 Jahre)

Tabelle der Quantilen der zentralen t-Verteilung für $q = 0.95$ und $q = 0.975$

Freiheits- grade df	$q=0.95$ $\alpha=0.05$	$q=0.975$ $\alpha=0.025$	Freiheits- grade df	$q=0.95$ $\alpha=0.05$	$q=0.975$ $\alpha=0.025$
1	6.314	12.706	11	1.796	2.201
2	2.920	4.303	12	1.782	2.179
3	2.353	3.182	13	1.771	2.160
4	2.132	2.776	14	1.763	2.145
5	2.015	2.571	15	1.753	2.132
6	1.934	2.447	20	1.724	2.086
7	1.895	2.365	30	1.697	2.042
8	1.860	2.306	60	1.671	2.000
9	1.833	2.262	100	1.660	1.984
10	1.812	2.228	∞	1.645	1.960

Grundlagen des statistischen Testens

Mit statistischen Tests sollen Hypothesen über **Parameter von Grundgesamtheiten** anhand von Daten (Stichprobenergebnissen) überprüft werden.

Die Grundgesamtheit ist durch ihre **Verteilungsfunktion $F(x;Q)$** mit unbekanntem Parameter Θ gekennzeichnet; d.h durch die Wahrscheinlichkeit für Realisationen von X , die kleiner oder gleich x sind: **$F(x;Q)=Pr(X \leq x;Q)$** hängt vom Parameter Θ ab.

Parameter von Grundgesamtheiten:

Wahrscheinlichkeiten π_1, π_2, \dots

Mittelwerte μ_1, μ_2, \dots

Verteilungsfunktionen $F_1(x), F_2(x), \dots$

Hypothesen legen für die Parameter der Grundgesamtheit bestimmte Werte (oder Wertebereiche) fest. Man unterscheidet eine zweiseitige und einseitige Testung:

Zweiseitige Testung:

Nullhypothese $H_0: \Theta = \Theta_0$ Alternativhypothese $H_1: \Theta \neq \Theta_0$

Einseitige Testung:

Nullhypothese $H_0: \Theta \leq \Theta_0$ Alternativhypothese $H_1: \Theta > \Theta_0$
 oder Nullhypothese $H_0: \Theta \geq \Theta_0$ Alternativhypothese $H_1: \Theta < \Theta_0$

Ein Test vergleicht die Daten der Stichprobe mit der Nullhypothese. Hierzu wird aus den Stichprobenergebnissen x_1, x_2, \dots, x_n eine **Teststatistik $T(\mathbf{x})$** berechnet, die den Unterschied zwischen den Daten und der Nullhypothese mißt. Der aus den beobachteten Daten berechnete Wert von $T(\mathbf{x})$ wird mit t_0 bezeichnet. Bei zukünftigen Stichproben werden andere Daten \mathbf{x} und damit auch andere Werte t der Teststatistik $T(\mathbf{x})$ vorkommen. In der Grundgesamtheit aller möglichen Wiederholungen der Beobachtungen (des Experiments) variiert der Wert von $T(\mathbf{x})$ zufällig. In dieser Grundgesamtheit ist $T(\mathbf{x})$ eine Zufallsgröße T , deren Verteilung $F_T(t, \Theta)$ von der Verteilung $F(x; \Theta)$ der Beobachtungen und dem Stichprobenumfang n abhängt.

Signifikanzwahrscheinlichkeit P

ist die Wahrscheinlichkeit, mit der bei zukünftigen Stichproben der beobachtete Wert t_0 oder ein größerer zu erwarten ist, wenn die Nullhypothese H_0 gilt.

$$P = \Pr(T > t_0 | H_0) = 1 - F_t(t_0; \Theta_0)$$

$\Pr(\dots)$ bedeutet "Wahrscheinlichkeit für...". Je größer der Unterschied zu H_0 ist, desto kleiner ist P. Bei einseitiger Testung wird die maximale Wahrscheinlichkeit unter H_0 (für $\Theta = \Theta_0$) genommen.

Powerfunktion $P_t(Q)$ zu gegebenem t_0 :

ist die Wahrscheinlichkeit, mit der bei zukünftigen Stichproben ein Wert $t > t_0$ zu erwarten ist, wenn der Parameterwert Θ ist.

$$P_t(\Theta) = \Pr(T > t_0 | \Theta) = 1 - F_t(t_0; \Theta)$$

Test als Entscheidung:

Mit einem Test soll entschieden werden, ob die Nullhypothese oder die Alternativhypothese zutrifft. Es wird eine Schwelle α (meist 0.05) vorgegeben und H_0 abgelehnt (H_1 angenommen), wenn $P \leq \alpha$ (0.05). Die Schwelle α (0.05) ist die Wahrscheinlichkeit, H_0 irrtümlich abzulehnen (**Fehler 1. Art**). α ist die entsprechende Irrtumswahrscheinlichkeit 1. Art. Dem Wert α entspricht eine **Entscheidungsschwelle** t_s , für die gilt: $P = \Pr(T > t_s | H_0) \leq \alpha$; d.h. H_0 wird abgelehnt, wenn der Wert t der Teststatistik T größer als t_s ist. Wenn H_0 abgelehnt wird, nennt man den Unterschied zwischen Daten und Nullhypothese **signifikant**.

Die Annahme der Nullhypothese ($P > \alpha$) bedeutet **nicht**, daß H_0 mit großer Zuverlässigkeit gilt. Der Fehler, H_0 anzunehmen (H_1 abzulehnen), obwohl H_1 gilt, heißt **Fehler 2. Art**. Die Wahrscheinlichkeit hierfür bezeichnet man mit β (Irrtumswahrscheinlichkeit 2. Art). Sie hängt (bei gegebenem α) vom Wert des Parameters Θ und dem Stichprobenumfang n ab.

Die Wahrscheinlichkeit $P_\alpha(\Theta)$, bei gegebenem α H_0 abzulehnen und H_1 anzunehmen, nennt man die **Powerfunktion** (Teststärke) zu gegebenem α . Es gilt:

$$P_\alpha(\Theta) = \Pr(T > t_s | \Theta) = 1 - F_t(t_s; \Theta) = 1 - \beta$$

Tabelle der möglichen Testentscheidungen

Es gilt	Entscheidung für:	
	H_0	H_1
H_0	richtig	Fehler 1. Art
H_1	Fehler 2. Art	richtig

Festlegung des Stichprobenumfangs n

- Vorgabe von α (meist 0.05)
- Vorgabe eines von Θ_0 abweichenden Wertes Θ_1 (Referenzwert)
- Vorgabe von β_1 bzw. der Power $P_\alpha(\Theta_1) = 1 - \beta_1$ (meist $\beta_1 = 0.2$ bzw. $P_\alpha(\Theta_1) = 0.8$).

Der Stichprobenumfang n (bzw. Gesamtumfang N bei mehreren Stichproben) soll so groß sein, daß bei Gültigkeit des Referenzwertes Θ_1 und bei dem vorgegebenen α die Fehlerwahrscheinlichkeit 2. Art (β) höchstens den Wert β_1 besitzt bzw. die Power $P_\alpha(\Theta_1)$ den vorgegebenen Wert $1-\beta_1$ (z.B. 0.8) mindestens erreicht.

Der verbundene (paarweise) t-Test

Eine Meßgröße x wird zu Beginn und am Ende einer Periode bestimmt. Es wird danach gefragt, ob der Mittelwert der Meßgröße am Ende der Periode gegenüber dem Beginn nicht erhöht ist ($H_0: \mu_{\text{Differenz}} \leq 0$) oder erhöht ist ($H_1: \mu_{\text{Differenz}} > 0$). Es ist eine einseitige Testung durchzuführen.

Beispiel: Versuchsergebnisse bei $n=8$ Wiederholungen:

Beginn	x_{1i} :	5	7	3	8	6	10	3	4
Ende	x_{2i} :	8	12	9	6	5	11	9	8
Differenz	d_i :	3	5	6	-2	-1	1	6	4

Mittelwert und Standardabweichung der Differenzen d_i sind:

$$\bar{d} = 2.75 \quad s^2_d = 9.64 \quad s_d = 3.10$$

Zu testen ist die Nullhypothese $H_0: \mu_d \leq 0$ gegen die Alternativhypothese $H_1: \mu_d > 0$.

Als Teststatistik wird die t-Statistik genommen:

$$t = \frac{\bar{d}\sqrt{n}}{s_d}$$

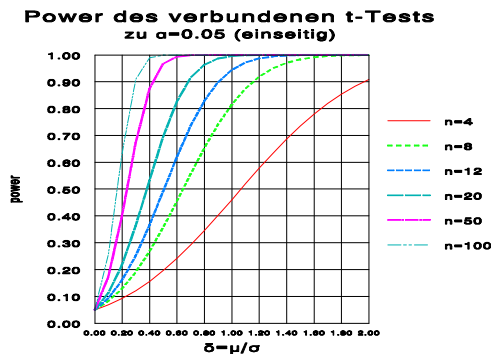
Die Daten des Experiments ergeben den Wert $t_0 = 2,5$ der Teststatistik.

Zur Berechnung der Signifikanzwahrscheinlichkeit P benötigt man die Kenntnis der Verteilung der Teststatistik, wenn H_0 gilt. Unter H_0 hat t (bei $n > 6$) eine zentrale t-Verteilung $F_{t,n-1}(t)$ mit $n-1$ Freiheitsgraden, die bereits im Abschnitt 'Schätzwerte und Konfidenzintervalle' erklärt wurde. Die Signifikanzschwelle t_s bei einem vorgegebenen α ist die $(1-\alpha)$ -Quantile $t_{1-\alpha,n-1}$ der zentralen t-Verteilung mit $n-1$ Freiheitsgraden. Die 0.95- und 0.975-Quantilen der zentralen t-Verteilung sind am Ende des Abschnitts 'Schätzwerte und Konfidenzintervalle' angegeben. Bei $\alpha=0.05$ und $n-1=7$ Freiheitsgraden ist $t_{1-0.05,7}=1.895$. Der aus den Daten des Experiments berechnete t-Wert t_0 ist 2.5 und somit größer als $t_{1-\alpha,n-1}$. Die Nullhypothese wird verworfen. Die mittlere Änderung ist **signifikant** größer als 0. Die Signifikanzwahrscheinlichkeit ist: $P = \Pr(t > t_0 | \mu_d = 0) = 1 - F_{t,n-1}(2.5) = 0.03$

Die Powerfunktion des Tests:

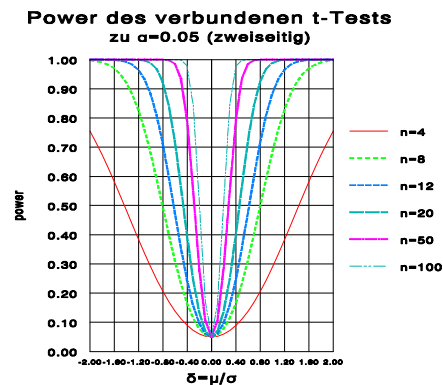
Für μ_d größer als Null (Alternativhypothese) gibt die Power die Wahrscheinlichkeit an, mit der bei dem gegebenen Stichprobenumfang n ein signifikantes Ergebnis zu erwarten ist. Diese Power hängt aber nicht nur von μ_d (sowie α und n) ab, sondern auch noch von der Standardabweichung σ_d der Differenzen in der Grundgesamtheit. Die t-Statistik hat unter der Alternative eine nichtzentrale t-Verteilung mit dem Nichtzentralitätsparameter $nc = (\mu_d / \sigma_d) \sqrt{n}$. Zur Berechnung der Power ist als Referenzwert nicht μ_d , sondern der

Quotient $\delta = \mu_d / \sigma_d$ vorzugeben. In der folgenden Abbildung sind für verschiedene n-Werte die Powerfunktionen des t-Tests bei einseitiger Testung mit $\alpha = 0.05$ gezeigt. Man erkennt, daß mit zunehmendem n die Kurven immer steiler verlaufen. Für die Stichprobengröße $n = 8$ des Beispiels ist bei $\delta = 1.0$ mit einer Power von etwa 0.8 ein signifikantes Ergebnis zu erwarten.



Zweiseitige Alternative ($H_0: \mu_d = 0; H_1: \mu_d \neq 0$):

Bei der Testung der Nullhypothese $H_0: \mu_d = 0$ gegen die zweiseitige Alternative $\mu_d \neq 0$ ist als Teststatistik der Betrag von t ($|t|$) zu nehmen. H_0 wird abgelehnt, wenn dieser Betrag größer als die Signifikanzschwelle t_s ist; d.h. entweder $t > t_s$ oder $t < -t_s$ ist. Die Wahrscheinlichkeit, daß unter H_0 entweder $t > t_s$ oder $t < -t_s$ eintritt, soll insgesamt α sein. Dies ist gegeben, wenn t_s so gewählt wird, daß unter H_0 die Wahrscheinlichkeit für $t > t_s$ gleich $\alpha/2$ und die Wahrscheinlichkeit für $t < -t_s$ ebenfalls gleich $\alpha/2$ ist. Wegen der Symmetrie der zentralen t-Verteilung um 0 ist als Signifikanzschwelle die $(1 - \alpha/2)$ -Quantile $t_{1-\alpha/2, n-1}$ der zentralen t-Verteilung mit $n-1$ Freiheitsgraden zu nehmen (für $\alpha = 0.05$ Quantile für $q = 0.975$). Darin besteht der wesentliche Unterschied zwischen einseitiger und zweiseitiger Testung. Da $t_{1-\alpha/2, n-1}$ stets größer als $t_{1-\alpha, n-1}$ ist, hat für eine gegebene Alternative $\delta = \mu_d / \sigma_d$ der zweiseitige Test eine geringere Power als der einseitige Test. In der folgenden Abbildung sind die Powerfunktionen des zweiseitigen Tests für $\alpha = 0.05$ und verschiedene n-Werte gezeigt. Für $n = 8$ und $\delta = 1$ ist die Power bei zweiseitiger Testung etwa 0.7. Für den Referenzwert $\delta = 1$ wird die Power 0.8 mit dem Stichprobenumfang $n = 10$ erreicht.



Bestimmung des Stichprobenumfangs

Der Stichprobenumfang n soll so groß sein, daß bei einem Referenzwert $\delta_1 = \mu_d / \sigma_d$ und der vorgegebenen Irrtumswahrscheinlichkeit $1 - \alpha$ die Nullhypothese H_0 mit der vorgegebenen Power $1 - \beta$ abgelehnt wird. Die Nullhypothese wird abgelehnt, wenn (bei einseitiger Testung) $t > t_{1-\alpha, n-1}$ ist. Bei gegebenem δ_1 ist die Wahrscheinlichkeit dafür: $P_{\alpha, n}(\delta_1) = 1 - F_{t, n-1}(t_{1-\alpha}, \delta_1 \sqrt{n})$, wobei $F_{t, n-1}(\cdot, \cdot)$ die Verteilung der nichtzentralen t-Verteilung mit $n-1$ Freiheitsgraden und dem Nichtzentralitätsparameter $nc = \delta_1 \sqrt{n}$ ist. Setzt man diesen Ausdruck gleich $1 - \beta$, dann kann daraus der erforderliche Stichprobenumfang berechnet werden. Die Berechnung muß iterativ erfolgen, da ja das gesuchte n auch in den Freiheitsgraden vorkommt. In der unten stehenden Tabelle sind für verschiedene α und β die erforderlichen Stichprobenumfänge n angegeben. Bei einem zweiseitigen Test mit der (gesamten) Irrtumswahrscheinlichkeit $\alpha = 0.05$ sind die Werte n in der Spalte mit $\alpha = 0.025$ zu nehmen.

Erforderlicher Stichprobenumfang n beim verbundenen t-Test

δ_1	$\alpha=0.025$ $\beta=0.2$	$\alpha=0.05$ $\beta=0.2$	$\alpha=0.025$ $\beta=0.1$	$\alpha=0.05$ $\beta=0.1$	$\alpha=0.025$ $\beta=0.05$	$\alpha=0.05$ $\beta=0.05$
0.1	787	620	1053	858	1302	1084
0.2	199	156	256	216	327	272
0.3	90	71	119	97	147	122
0.4	52	41	68	55	84	70
0.5	34	27	44	36	54	45
0.6	24	19	32	26	39	32
0.7	19	15	24	19	29	24
0.8	15	12	19	15	23	19
0.9	12	10	16	13	19	15
1.0	10	8	13	11	16	13

Eine einfache Formel für den erforderlichen Stichprobenumfang n erhält man, wenn auch für den Test die Standardabweichung σ_d als bekannt angesehen wird und daher die Teststatistik $z = (\bar{d} / \sigma_d) \sqrt{n}$ (bzw. bei zweiseitigem Test $|z|$) genommen wird. Diese ist für $\delta_1 = \mu_d / \sigma_d$ normalverteilt mit dem Mittelwert $\delta_1 \sqrt{n}$ und der Standardabweichung 1. Die Nullhypothese wird abgelehnt, wenn z die Signifikanzschwelle z_s überschreitet, die bei einseitigem Test gleich der $(1-\alpha)$ -Quantile $z_{1-\alpha}$ und beim zweiseitigen Test gleich der $(1-\alpha/2)$ -Quantile $z_{1-\alpha/2}$ der Standardnormalverteilung ist. Die Power bei gegebenem δ_1 ist: $P_{\alpha, n}(\delta_1) = 1 - \Phi(z_s - \delta_1 \sqrt{n})$. Der Stichprobenumfang n soll so groß sein, daß $P_{\alpha, n}(\delta_1) = 1 - \beta$ ist.

Die Lösung dieser Gleichung ist:
$$n = \frac{(z_s + z_{1-\beta})^2}{\delta_1^2}$$

Für $\alpha = 0.05$ und $\beta = 0.2$ (Power = 0.8) ist folgender Stichprobenumfang erforderlich:

$$\text{einseitige Testung: } n = \frac{7}{\delta_1^2}; \quad \text{zweiseitige Testung: } n = \frac{8}{\delta_1^2}$$

Der unverbundene t-Test zum Vergleich von zwei Mittelwerten

Es soll ein neuer Blutdrucksenker A gegen ein Vergleichspräparat B geprüft werden. Zielgröße ist die Blutdrucksenkung x (mmHg) nach einer 6-wöchigen Therapie. Folgende Blutdruckänderungen wurden beobachtet:

Gruppe 1 mit Behandlung A ($n_1=10$ Patienten): x_{1j} : 10, 5, -2, 3, 8, 2, 0, 4, 2, 6.
Gruppe 2 mit Behandlung B ($n_2= 6$ Patienten): x_{2j} : 3, -4, 2, 0, 5, -2.

Daraus berechnet man folgende Mittelwerte und Standardabweichungen:

in Gruppe 1: $\bar{x}_1 = 3.8$ $s_1=3.6$
in Gruppe 2: $\bar{x}_2 = 0.7$ $s_2=3.3$

Nullhypothese:

einseitig: $\mu_1 \leq \mu_2$ bzw. $\mu_1 \geq \mu_2$
zweiseitig: $\mu_1 = \mu_2$

Alternativhypothese:

einseitig: $\mu_1 > \mu_2$ bzw. $\mu_1 < \mu_2$
zweiseitig: $\mu_1 \neq \mu_2$

(μ_1 und μ_2 sind die Mittelwerte in den Grundgesamtheiten)

Es wird angenommen, daß in der Grundgesamtheit die Standardabweichung σ für beide Gruppen gleich ist. Ein Schätzwert ist die "gepoolte" Standardabweichung:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Teststatistik ist:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad (\text{einseitig}) \quad t = \frac{|\bar{x}_1 - \bar{x}_2|}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad (\text{zweiseitig})$$

Im Beispiel ist: $s=3.5$; $\bar{x}_1 - \bar{x}_2 = 3.8 - 0.7 = 3.1$; $s=3.5$; $n_1=10$; $n_2=6$; **$t=1.712$** .

Die Signifikanzwahrscheinlichkeit hängt von der "Zahl der Freiheitsgrade" $df = n_1 + n_2 - 2$ ab und ist durch die zentrale t-Verteilung $F_{t,df}(t)$ gegeben. Die Signifikanzschwelle t_s ist bei einseitiger Testung die $(1-\alpha)$ -Quantile und bei zweiseitiger Testung die $(1-\alpha/2)$ -Quantile der zentralen t-Verteilung mit $df=n_1+n_2-2$ Freiheitsgraden (Tabelle).

Im Beispiel ist $t=1.712$, $df=14$, $t_s=1.763$ (einseitig) bzw. $t_s=2.145$ (zweiseitig), **$P_{\text{eins}}=0.054$** ; **$P_{\text{zweis}}=0.108$** . Der Unterschied ist "nicht signifikant".

Für $n_1+n_2 > 15$ kann als Entscheidungsschwelle t_s bei $\alpha=0.05$, zweiseitig ($t_s=t_{0.975,df}$), der Wert 2 genommen werden; d.h. die Nullhypothese wird abgelehnt (der Unterschied ist "signifikant") für $t > 2$.

Erforderliche Stichprobenumfänge:

Werden in jeder Gruppe gleich viele Werte erhoben ($n_1=n_2=n$), dann sind die Stichprobenumfänge n pro Gruppe für die zweiseitige Testung ($\alpha=0.025$) und einseitige Testung ($\alpha=0.05$) in der folgenden Tabelle angegeben, die erforderlich sind, um mit der Power $1-\beta$ ($\beta=0.2, 0.1, 0.05$) ein signifikantes Ergebnis erwarten zu können, wenn die auf die Standardabweichung σ bezogenen Differenz der Erwartungswerten μ_1 und μ_2 den Wert δ_1 hat: $\delta_1=(\mu_1-\mu_2)/\sigma$.

Erforderlicher Stichprobenumfang n pro Gruppe beim unverbundenen t-Test

δ_1	$\alpha=0.025$ $\beta=0.2$	$\alpha=0.05$ $\beta=0.2$	$\alpha=0.025$ $\beta=0.1$	$\alpha=0.05$ $\beta=0.1$	$\alpha=0.025$ $\beta=0.05$	$\alpha=0.05$ $\beta=0.05$
0.1	1571	1238	2103	1714	2600	2166
0.2	394	310	527	429	651	542
0.3	176	139	235	191	290	242
0.4	100	78	133	108	164	136
0.5	64	51	86	70	105	88
0.6	45	36	60	49	74	61
0.7	34	26	44	36	55	45
0.8	26	21	34	28	42	35
0.9	21	16	27	27	34	28
1.0	17	14	23	18	27	23

Als Approximation erhält man – analog zu den beim verbundenen t-Test gebrachten Überlegungen – die Formel:

$$n = \frac{2 \cdot (z_s + z_{1-\beta_1})^2}{\delta_1^2}$$

Speziell gilt:

$$n \approx \frac{13}{\delta_1^2} \text{ für } \alpha=0.05 \text{ (einseitig) und } \beta=0.02; \quad n \approx \frac{16}{d_1^2} \text{ für } \alpha=0.05 \text{ (zweiseitig) und } \beta_1=0.2.$$

Vergleich von Wahrscheinlichkeiten (Chi²-Test)

Ein binäres Merkmal (z.B. geheilt – nicht geheilt) wird bei zwei Gruppen an n_1 bzw. n_2 Einheiten beobachtet. Z.B. wird beobachtet, daß mit dem Arzneimittel A von 20 Patienten 12 geheilt wurden (Häufigkeit $h_1 = 12/20 = 0.6$) und mit dem Mittel B von 16 Patienten 6 (Häufigkeit $h_2 = 6/16 = 0.375$). Sind die Heilungswahrscheinlichkeiten π_1 und π_2 gleich (Nullhypothese) oder verschieden (Alternative)?

Die Ergebnisse lassen sich in einer 4-Felder-Tafel darstellen:

	geheilt	nicht geheilt	Gesamt
Substanz A	a = 12	b = 8	$n_1 = 20$
Substanz B	c = 6	d = 10	$n_2 = 16$
Gesamt	$m_1 = 18$	$m_2 = 18$	$N = 36$

Teststatistik ist:

$$X^2 = \frac{(a \cdot d - b \cdot c)^2 N}{n_1 \cdot n_2 \cdot m_1 \cdot m_2}$$

Die Nullhypothese, daß die Heilungswahrscheinlichkeiten π_1 und π_2 gleich sind, wird mit Irrtumswahrscheinlichkeit $\alpha=0.05$ verworfen, wenn $X^2 > 3.84$ ist.

Bei einseitiger Alternative $\pi_1 > \pi_2$ ist die Nullhypothese ($\pi_1 \leq \pi_2$) zu verwerfen, wenn $X > 1.64$ und $ad > bc$, bei Alternative $\pi_1 < \pi_2$ (Nullhypothese $\pi_1 \geq \pi_2$), wenn $X > 1.64$ und $ad < bc$ sind. Im Beispiel ist: $X^2 = (12 \cdot 10 - 8 \cdot 6)^2 \cdot 36 / (20 \cdot 16 \cdot 18 \cdot 18) = 1.8$; $X^2 < 3.84$: kein signifikanter Unterschied! $X = 1.34 < 1.64$. Auch die einseitige Hypothese kann nicht verworfen werden.

Die erforderlichen Stichprobenumfänge n pro Gruppe, um bei einem zweiseitigen Test zum Niveau $\alpha=0.05$ mit der Power $1-\beta=0.8$ ein signifikantes Ergebnis erwarten zu können, wenn in der Gruppe 1 (Behandlung A) die Wahrscheinlichkeit π_1 und in der Gruppe 2 (Behandlung B) die Wahrscheinlichkeit π_2 beträgt, sind in der folgenden Tabelle angegeben:

Tabelle für Stichprobenumfang n pro Gruppe ($n_1=n_2=n$) für $\alpha=0.05$ (zweiseitig) und Power $1-\beta=0.8$

π_2	π_1							
	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	197	59	29	17	11	7	5	3
0.2		291	79	36	20	12	7	5
0.3			354	91	40	21	12	7
0.4				385	95	40	20	11
0.5					385	91	36	17
0.6						354	79	29
0.7							291	59
0.8								197

Nach einer Näherungsformel: $n = 4 / (\pi_1 - \pi_2)^2$ erhält man:

$\pi_1 - \pi_2$:	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$n =$	400	100	45	25	16	12	9	2

Mehr als 2 Kategorien und/oder Gruppen:

Hat das Merkmal mehr als 2 Kategorien (allgemein q Kategorien) und/oder sollen die Verteilungen des Merkmals zwischen mehr als 2 Gruppen (allgemein p Gruppen) verglichen werden, dann können die Ergebnisse in einer $p \times q$ -Feldertafel dargestellt werden.

	Kategorie 1	Kategorie 2	... Kategorie j ...	Kategorie q	Summe
Gruppe 1	n_{11}	n_{12}	... n_{1j} ...	n_{1q}	$n_{1.}$
Gruppe 2	n_{21}	n_{22}	... n_{2j} ...	n_{2q}	$n_{2.}$
...
Gruppe i	n_{i1}	n_{i2}	... n_{ij} ...	n_{iq}	$n_{i.}$
...
Gruppe p	n_{p1}	n_{p2}	... n_{pj} ...	n_{pq}	$n_{p.}$
Summe	$n_{.1}$	$n_{.2}$... $n_{.j}$...	$n_{.q}$	N

Die Wahrscheinlichkeit, in der Gruppe i die Kategorie j zu beobachten, wird mit π_{ij} bezeichnet ($\sum_j \pi_{ij} = 1$ für alle i). Zu testen ist die Nullhypothese, daß die π_{ij} bei allen Gruppen gleich sind: $\pi_{1j} = \dots = \pi_{qj} = \pi_j$ für alle j . Ist dies der Fall, dann ist $h_{.j} = n_{.j} / N$ ein Schätzwert für die Wahrscheinlichkeit π_j . Ein Schätzwert für den Erwartungswert der beobachteten Anzahl n_{ij} ist: $E_{ij} = n_{i.} \cdot n_{.j} / N$. Teststatistik ist:

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - n_{i.} \cdot n_{.j} / N)^2}{n_{i.} \cdot n_{.j} / N}$$

Unter H_0 hat diese Teststatistik (bei nicht zu kleinen Erwartungswerten E_{ij} , die zumindest größer als 0.5 sein sollten) eine zentrale χ^2 -Verteilung mit $df = (p-1)(q-1)$ Freiheitsgraden; d.h. sie ist die Summe aus df Quadraten standard-normalverteilter, unabhängiger Zufallsgrößen verteilt. Die Signifikanzschwellen $\chi^2_{0.95, df}$ zu $\alpha = 0.05$, zweiseitig, sind für $df = 1$ bis 10 Freiheitsgraden unten angegeben:

df	1	2	3	4	5	6	7	8	9	10
$\chi^2_{0.95, df}$	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51	16.92	18.31

Beispiel:

Zu Beginn dieses Abschnitts wurde die Ausprägung des Hustens am Ende der Erkältungsstudie für je 10 Patienten, die mit Verum bzw. mit Placebo behandelt wurden, angegeben. Die Ausprägung des Hustens hatte die ordinalen Kategorien: keiner, leicht, stark. Die Anzahlen sind in der folgenden 2x3-Feldertafel zusammengestellt:

Gruppe	Husten am Ende der Studie			Summe
	keiner	leicht	stark	
Verum	7	3	0	10
Placebo	3	5	2	10
Summe	10	8	2	20

Die Erwartungswerte (unter H_0) sind: $E_{11} = E_{21} = 5$; $E_{12} = E_{22} = 4$; $E_{13} = E_{23} = 1$. Damit ist: $X^2 = 4,1$ mit $df = 2$ Freiheitsgraden. Die 5%-Signifikanzschwelle bei 2 Freiheitsgraden ist: $\chi^2_{0.95, 2} = 5.99$. Da $X^2 < 5.99$, wird die Nullhypothese angenommen.

Rangtest nach Wilcoxon-Mann-Whitney

Der Test wird eingesetzt zum Prüfen der Unterschiede in den Verteilungsfunktionen zweier Gesamtheiten. Er wird häufig statt des t-Tests verwendet, der Unterschiede in den Mittelwerten zweier Gesamtheiten testet (Lokationstest). Im Unterschied zum t-Test prüft der Rangtest Unterschiede in der gesamten Verteilung. Da er nicht auf einen bestimmten Parameter der Verteilung orientiert ist, nennt man ihn auch einen **parameterfreien (nichtparametrischen)** Test.

Der Test verwendet eine Teststatistik, die aus den **Rangzahlen** der Meßwerte berechnet wird. Alle Meßwerte beider Stichproben werden der Größe nach geordnet. Die Nummer eines Meßwertes x_i in der Reihenfolge dieser geordneten Meßwerte ist die Rangzahl r_i dieses Meßwertes. Haben mehrere Stichprobenwerte den gleichen Wert, so wird ihnen der Mittelwert ihrer Rangzahlen zugewiesen (Mittelrang).

Beispiel (des t-Tests):

Folgende Blutdrucksenkungen wurden bei Gruppe A und B festgestellt:

Gruppe A ($n_1=10$): x_{1i} : 10, 5, -2, 3, 8, 2, 0, 4, 2, 6. Gruppe B ($n_2=6$): x_{2i} : 3, -4, 2, 0, 5, -2.

Nullhypothese: $F_A(x)=F_B(x)$;

Alternative: einseitig: $F_A(x)<F_B(x)$ bzw. $F_A(x)>F_B(x)$; zweiseitig: $F_A(x) \neq F_B(x)$ für alle x .

Die Rangzahlen der 16 Meßwerte und die zugehörige Gruppe zeigt folgende Tabelle:

Stichprobe	B	A	B	A	B	A	A	B
Wert	-4	-2	-2	0	0	2	2	2
Rang	1	2.5	2.5	4.5	4.5	7	7	7
Stichprobe	A	B	A	A	B	A	A	A
Wert	3	3	4	5	5	6	8	10
Rang	9.5	9.5	11	12.5	12.5	14	15	16

Der Test basiert auf der Summe W_A der Rangzahlen der Gruppe A. Bei mehr als 15 Meßwerten ist W_A normal verteilt und hat bei Gültigkeit der Nullhypothese den Mittelwert $\mu_{WA} = \frac{1}{2}n_1(n_1+n_2+1)$ und die Varianz $\sigma_{WA}^2 = (n_1n_2(n_1+n_2+1))/12$.

$$\text{Teststatistik ist: } z = \frac{W_A - \frac{1}{2}n_1(n_1 + n_2 + 1)}{\sqrt{n_1n_2(n_1 + n_2 + 1)/12}}$$

Bei Gültigkeit der einseitigen Alternative $F_A(x)<F_B(x)$ (d.h. die Meßwerte von A sind wahrscheinlich größer als die von B: $\Pr(X_A>X_B)>1/2$) ist + zu erwarten, daß W_A größer als der Mittelwert μ_{WA} ist. Die Nullhypothese wird bei gegebener Schwelle α (z. B. 0.05) abgelehnt, wenn z größer als $z_{1-\alpha}$ ist, wobei $z_{1-\alpha}$ die $1-\alpha$ -Quantile der Standard-Normalverteilung ($N(0,1)$) ist. Für $\alpha=0.05$ ist $z_{1-\alpha}=1.64$.

Bei Gültigkeit der einseitigen Alternative $F_A(x)>F_B(x)$ (d.h. die Meßwerte von A sind wahrscheinlich kleiner als die von B: $\Pr(X_A>X_B)<1/2$) ist zu erwarten, daß W_A kleiner als der Mittelwert μ_{WA} ist. Die Nullhypothese wird bei gegebenem α abgelehnt, wenn z kleiner als $-z_{1-\alpha}$ ist.

Zum Testen gegen die zweiseitige Alternative $F_A(x)\neq F_B(x)$ wird als Teststatistik der Betrag $|z|$ benutzt und die Nullhypothese abgelehnt, wenn dieser Betrag größer als $z_{1-\alpha/2}$ ist (bei $\alpha=0.05$ größer als 1.96). Alternativ kann die Teststatistik z^2 mit der Schwelle $z_{1-\alpha/2}^2$ (3.84 für $\alpha=0.05$) genommen werden.

Im Beispiel ist:

$$W_A = 99 \quad W_B = 37 \quad (W_A+W_B = 136 = (10+6)(10+6+1)/2)$$

$$\mu_{WA} = 85 \quad \sigma_{WA}^2 = (10*6*17)/12 = 85 \quad \sigma_{WA} = 9.2$$

Damit ist:

$$z = (99 - 85)/9.2 = 1.52 < 1.64: \quad \text{Annahme von } H_0$$

Dieser Test kann auch zum Vergleich der Wahrscheinlichkeitsverteilung eines ordinalen Merkmals benutzt werden, dessen Ausprägungen in zwei unabhängigen Gruppen (Stichproben) beobachtet wurde. Als Beispiel sei die Ausprägung des Hustens am Ende der Erkältungsstudie genannt, die 'keiner', 'leicht' und 'stark' sein kann. Die Ausprägungen wurden bei je 10 Patienten, die mit Verum bzw. Placebo behandelt wurden, festgestellt. In der Verum-Gruppe hatten 7 Patienten keinen Husten, 3 einen leichten und keiner einer starken Husten. In der Placebo-Gruppe hatten 3 Patienten keinen Husten, 5 einen leichten und 2 einen starken Husten. Es hatten also insgesamt 10 Patienten keinen Husten, 8 einen leichten Husten und 2 einen starken Husten. Der Mittelrang der Kategorie 'kein Husten' ist $(10+1)/2=5.5$, der der Kategorie 'leichter Husten': $10+(8+1)/2=14.5$ und der der Kategorie 'starker Husten': $18+(2+1)/2=19.5$. Der Wert der Teststatistik, die prüft, ob in der Verum-Gruppe eher geringere Ausprägungen vorkommen als in der Placebo-Gruppe, ist:

$$W_A = 7 \cdot 5.5 + 3 \cdot 14.5 + 0 \cdot 19.5 = 82$$

Unter der Nullhypothese ist der Mittelwert $\mu_{WA} = n_1(n_1+n_2+1)/2 = 10 \cdot 21/2 = 105$. Bei der Standardabweichung müssen die Bindungen (gleiche Kategorien) berücksichtigt werden, weshalb die oben angegebene Formel nicht gilt. Der korrekte Wert ist: $\sigma_{WA} = 11.92$. Die Teststatistik z hat somit den Wert: $z = (82-105)/11.92 = -1.93$. Die 5%-Quantile der Standard-Normalverteilung ist: $z_{0.05} = -1.64$. Da z kleiner als $z_{0.05}$ ist, kann die einseitige Nullhypothese, daß die Ausprägungen des Hustens am Ende der Studie unter der Verum-Behandlung wahrscheinlich nicht geringer sind als die unter der Placebo-Behandlung, abgelehnt werden.

Vergleich von Zeiten (Überlebenszeit, Krankheitsdauer)

Beispiel:

Ein neues Medikament A zur Behandlung von Erkältungskrankheiten soll mit der Standardbehandlung B verglichen werden. Zielgröße ist die Dauer der Krankheit.

Problem:

Zensierte Daten:

Ein Teil der Patienten scheidet noch krank aus. Für diese Patienten ist die Krankheitsdauer nicht bekannt; man kennt nur den Zeitpunkt der letzten Untersuchung und weiß, daß der Patient dann noch krank war.

Auswertung:

Für jeden Patienten ist die Zeit t_i bekannt, zu der er entweder gesund wurde oder aus der Studie ausschied. Diese Zeiten werden der Größe nach geordnet ($t_i < t_{i+1}$). Insgesamt wurden k verschiedene Zeiten beobachtet.

Es bedeuten:

n_{1i} (n_{2i}) die Zahl der unmittelbar vor t_i noch kranken Patienten aus A (B),

d_{1i} (d_{2i}) die Zahl der Patienten aus A (B), die zur Zeit t_i gesund wurden,

c_{1i} (c_{2i}) die Zahl der zur Zeit t_i ausgeschiedenen Patienten aus A (B).

$q_{1i} = d_{1i}/n_{1i}$ ist die Intensität zur Zeit t_i in Gruppe A und $q_{2i} = d_{2i}/n_{2i}$ die in B.

Beispiel: Krankheitsdauer in zwei Behandlungsgruppen

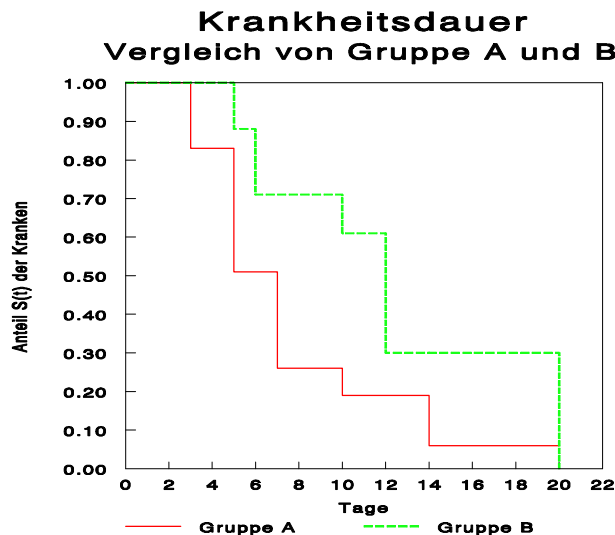
t_i	n_{1i}	d_{1i}	Gruppe A			n_{2i}	d_{2i}	Gruppe B		
			c_{1i}	q_{1i}	S_{1i}			c_{2i}	q_{2i}	S_{2i}
0	20	-	-	-	1.00	20	-	-	-	1.00
2	20	-	2	-	1.00	20	-	-	-	1.00
3	18	3	2	0.17	0.83	20	-	3	-	1.00
5	13	5	-	0.38	0.51	17	2	-	0.12	0.88
6	8	-	-	-	0.51	15	3	2	0.20	0.71
7	8	4	-	0.50	0.26	10	-	3	-	0.71
10	4	1	-	0.25	0.19	7	1	2	0.14	0.61
12	3	-	-	-	0.19	4	2	-	0.50	0.30
14	3	2	-	0.67	0.06	2	-	1	-	0.30
20	1	-	1	-	0.06	1	1	-	1.00	0.00

Die Krankheitsdauerfunktion (Überlebensfunktion) S_{1i} (S_{2i}) ist die Wahrscheinlichkeit, in der Gruppe A (bzw. B) zur Zeit t_i noch krank zu sein (zu leben). Es ist:

$$S_{1i} = S_{10}(1-q_{11})(1-q_{12})\dots(1-q_{1i})$$

$$S_{2i} = S_{20}(1-q_{21})(1-q_{22})\dots(1-q_{2i}) \quad (\text{Kaplan-Meier Schätzer; E. Halley})$$

Die mediane Krankheitsdauer (Überlebenszeit) t_{Med} ist definiert durch $S(t_{Med})=0.5$. Sie ist für die Gruppe A: $t_{Med}=5$ Tage und für die Gruppe B: $t_{Med}=11$ Tage.



Der Log-Rank-Test

Nullhypothese: $H_0: S_1(t) = S_2(t)$
 Alternative: $H_1: S_1(t) \neq S_2(t)$ ($S_1(t) = S_2(t)^Q$ $Q > 0$)

$S_1(t)$ und $S_2(t)$ sind die Krankheitsdauerfunktionen (Überlebensfunktionen) für die den Gruppen A und B zugrundeliegenden Grundgesamtheiten.

Unter H_0 ist für jede Gruppe dieselbe 'Gesundung' (Sterblichkeit) zu erwarten. Es ist daher die (unter H_0) zu erwartende Zahl von Ereignissen zur Zeit t_i :

$$e_{1i} = d_i \frac{n_{1i}}{n_i} \text{ für Gruppe A und } e_{2i} = d_i \frac{n_{2i}}{n_i} \text{ für Gruppe B}$$

mit: $d_i = d_{1i} + d_{2i}$ und $n_i = n_{1i} + n_{2i}$.

Teststatistik (bei zweiseitiger Alternative) ist:

$$\chi^2 = \frac{(e_{1.} - d_{1.})^2}{e_{1.}} + \frac{(e_{2.} - d_{2.})^2}{e_{2.}}$$

mit: $e_{1.} = \sum_{i=1}^k e_{1i}$, $e_{2.} = \sum_{i=1}^k e_{2i}$, $d_{1.} = \sum_{i=1}^k d_{1i}$, $d_{2.} = \sum_{i=1}^k d_{2i}$.

H_0 wird (bei $\alpha=0.05$, zweiseitig) abgelehnt, wenn $\chi^2 > 3.8$.

Beispiel für Logrank-Test:

t_i	Gruppe A			Gruppe B			Gesamt	
	n_{1i}	d_{1i}	e_{1i}	n_{2i}	d_{2i}	e_{2i}	d_i	n_i
0	20	0	0	20	0	0	0	40
3	18	3	1.42	20	0	1.58	3	38
5	13	5	3.03	17	2	3.97	7	30
6	8	0	1.04	15	3	1.96	3	23
7	8	4	1.78	10	0	2.22	4	18
10	4	1	0.73	7	1	1.27	2	11
12	3	0	0.86	4	2	1.14	2	7
14	3	2	1.20	2	0	0.80	2	5
20	1	0	0.50	1	1	0.50	1	2
Summe		15	10.56		9	13.44		

$$\begin{aligned} \chi^2 &= (10.56-15)^2/10.56 + (13.44-9)^2/13.44 \\ &= 1.87 + 1.47 \\ &= 3.34 \end{aligned}$$

($\chi^2 < 3.8$; nicht signifikant !)

Teil III: Statistische Methoden in der Epidemiologie

Phase IV: Therapeutische Prüfung und epidemiologische Studien

Ziel epidemiologischer Studien:

systematische Untersuchung von **Risikofaktoren**

Risiko:

Wahrscheinlichkeit für ein negatives Ereignis

Risikofaktor:

spezielle Bedingungen, die die Wahrscheinlichkeit für ein negatives Ereignis erhöhen

Personen, die den speziellen Bedingungen ausgesetzt sind, nennt man **exponiert**.

Schema zur Risikoermittlung:

Exposition	Ereignis		Gesamt
	ja	nein	
ja	a	b	n ₁
nein	c	d	n ₂
Gesamt	m ₁	m ₂	N

Risiko(exponiert):

$$R_{\text{exp}} = a/n_1$$

Risiko(nicht exponiert):

$$R_{\text{nicht exp.}} = c/n_2$$

Quote (odds) (exponiert):

$$O_{\text{exp.}} = a/b$$

Quote (odds) (nicht exponiert):

$$O_{\text{nicht exp.}} = c/d$$

Relatives Risiko

$$RR = R_{\text{exp}}/R_{\text{nicht exp.}} = (a/n_1)/(c/n_2)$$

Exzess-Risiko

$$AR = R_{\text{exp}} - R_{\text{nicht exp.}} = a/n_1 - c/n_2$$

Relative Quote (odds ratio) RO

$$RO = O_{\text{exp.}}/O_{\text{nicht exp.}} = (ad)/(bc)$$

Bevölkerungsstudien zur Feststellung von Risikofaktoren:

Kohortenstudien

Aus der (relevanten) Bevölkerung wird eine "Kohorte" von N Personen ausgewählt und es werden

- die Exponierten mit Ereignis (a)
- die Exponierten ohne Ereignis (b)
- die Nicht-Exponierten mit Ereignis (c)
- die Nicht-Exponierten ohne Ereignis (d)

ermittelt.

Fall-Kontroll-Studie

Aus der Bevölkerung werden m_1 "Fälle" (Personen mit Ereignis) und m_2 "Kontrollen" (Personen ohne Ereignis) ausgewählt und die Expositionsquoten (a/c bzw. b/d) zwischen beiden Gruppen verglichen (Relative Quote RO)

Kohortenstudien

Beispiel: Risiko für Brustkrebs nach Einnahme von Reserpin (Kewitz et al. 1977)

Reserpin	Brustkrebs		Gesamt
	ja	nein	
ja	32	57	89
nein	149	351	500
Gesamt	181	408	589

Risiko(Reserpin) = $32/89$ = 0.360

Risiko(kein Reserpin) = $149/500$ = 0.298

Relatives Risiko RR = 1.207

Exzess-Risiko AR = $0.360 - 0.298 = 0.062$

Statistische Variabilität:

Das relative Risiko wurde mit einer Auswahl aus der Gesamtheit von Frauen (Grundgesamtheit) ermittelt (Schätzwert). Der Schätzwert variiert zufällig um das "wahre" relative Risiko RR_{wahr} in der Grundgesamtheit.

Das Ausmaß der Zufallsvariation wird durch das **95%-Konfidenzintervall** ausgedrückt. Das ist ein Intervall, das aus den beobachteten Zahlen (a,b,c,d) berechnet wird und das "wahre" relative Risiko RR_{wahr} mit Wahrscheinlichkeit 95% überdeckt.

$$95\text{-CI} \approx RR(1 \pm 2 \sqrt{\frac{b}{an_1} + \frac{d}{cn_2}})$$

95%-Konfidenzintervall für RR_{wahr} bei Reserpin-Studie: 95%-CI: 0.827 - 1.587

Für $RR_{\text{wahr}} = 1$ liegt kein erhöhtes Risiko bei Exposition vor. Die Hypothese: $RR_{\text{wahr}} = 1$ wird mit dem Chi²-Test getestet:

$$\chi^2 = \frac{(ad - bc)^2 N}{m_1 m_2 n_1 n_2}$$

Die Hypothese wird (mit Irrtumswahrscheinlichkeit $\alpha=0.05$) verworfen, wenn $X^2 > 3.8$ ist.

Im Reserpin-Beispiel ist: $X^2 = 1.345$ ($P > 0.05$): Kein erhöhtes Risiko.

Systematische Fehler bei Risikoermittlung:

Berichtsfehler (bias of reporting)

falsche Angaben bei Ereignis oder Exposition

Vermengungsfehler (bias of confounding)

Inhomogenitäten bezüglich eines zusätzlichen Faktors (z.B. Alter)

Selektionsfehler (bias of selection)

Fehler bei der Auswahl

Beispiel für Berichtsfehler

In der Studie von Kewitz et al. geben 7 Frauen mit Brustkrebs eine Reserpin-Einnahme an, obwohl sie kein Reserpin genommen haben.

Es ergibt sich dann folgende Tabelle:

Reserpin	Brustkrebs		Gesamt
	ja	nein	
ja	39	57	96
nein	142	351	493
Gesamt	181	408	589

Risiko(Reserpin) = 39/96 = 0.406

Risiko(kein Reserpin) = 142/493 = 0.288

Relatives Risiko RR = 1.410

Beispiel für Vermengung (confounding) durch Alter:

Die Daten der Studie Kewitz et al. werden nach Altersgruppen aufgliedert:

Alter unter 50 Jahren:

Reserpin	Brustkrebs		Gesamt
	ja	nein	
ja	2	14	16
nein	42	221	263
Gesamt	44	235	279

$$RR = 0.125/0.160 = 0.781$$

Alter über 50 Jahre

Reserpin	Brustkrebs		Gesamt
	ja	nein	
ja	30	43	73
nein	107	130	237
Gesamt	137	173	310

$$RR = 0.411/0.451 = 0.911$$

In jeder Altersgruppe ist das berechnete relative Risiko RR kleiner als 1, obwohl es in der gesamten Tabelle größer als 1 ist (**Simpson's Paradoxon**). Dies ist darauf zurückzuführen, daß das Alter mit der Einnahme von Reserpin und dem Auftreten von Brustkrebs 'vermengt' (confounded) ist: Ältere Frauen nehmen häufiger Reserpin und bekommen häufiger Brustkrebs als jüngere Frauen.

Der Einfluß einer solchen Vermengung kann durch Stratifizierung (Bilden von Untergruppen) ausgeglichen werden. In jeder Untergruppe werden die relativen Risiken berechnet. Es wird angenommen, daß die 'wahren' relativen Risiken in allen Untergruppen gleich sind. Die Risiken für Exponierte und Nicht-Exponierte können aber bei verschiedenen Untergruppen durchaus verschieden sein. Die gewichtete Summe der relativen Risiken der Untergruppen ist ein Schätzwert für das (gemeinsame) 'wahre' relative Risiko.

Formeln zur Berechnung des relativen Risikos bei Stratifikation:

Es wurden k Strata mit jeweils n_j Beobachtungen gebildet. Es bedeuten:

- n_{1j} = Zahl der Exponierten in Stratum j
- n_{2j} = Zahl der Nicht-Exponierten in Stratum j
- a_j = Zahl der Ereignisse bei Exponierten in Stratum j
- c_j = Zahl der Ereignisse bei Nicht-Exponierten in Stratum j
- $n_j = n_{1j} + n_{2j}$ = Zahl der Personen in Stratum j
- $R_{\text{exp}}(j) = a_j/n_{1j}$ = Risiko für Exponierte in Stratum j
- $R_{\text{nicht exp.}}(j) = c_j/n_{2j}$ = Risiko für Nicht-Exponierte in Stratum j.

Für die Strata j werden Gewichte g_j vorgegeben. Das standardisierte relative Risiko ist:

$$\text{SRR} = \frac{\sum_{j=1}^k g_j R_{\text{exp}}(j)}{\sum_{j=1}^k g_j R_{\text{nicht exp.}}(j)} = \sum_{j=1}^k G_j \text{RR}_j \quad \text{mit} \quad G_j = \frac{g_j \frac{c_j}{n_{2j}}}{\sum_{j=1}^k g_j \frac{c_j}{n_{2j}}} \quad \text{und} \quad \text{RR}_j = \frac{a_j n_{2j}}{c_j n_{1j}}$$

Spezielle Gewichte:

a) Bezug auf Exponierte:

$$g_j = n_{1j} \quad (\text{Zahl der Exponierten})$$

$$\text{SRR}(\text{Exp}) = \frac{A}{\sum_{j=1}^k n_{1j} \frac{c_j}{n_{2j}}}$$

mit $A = \sum_j a_j$ = Gesamtzahl der Ereignisse bei Exponierten.

b) Mantel-Haenszel Verfahren:

$$g_j = (n_{1j} n_{2j}) / n_j \quad (\text{harmonisches Mittel})$$

$$\text{SRR}(\text{M-H}) = \frac{\sum_{j=1}^k \frac{a_j n_{2j}}{n_j}}{\sum_{j=1}^k \frac{c_j n_{1j}}{n_j}}$$

Beispiel: (Studie von Kewitz et al.)

$$\text{SRR}(\text{Exp}) = 32 / (16 * (42/263) + 73 * (107/237)) = 0.90$$

$$\text{SRR}(\text{M-H}) = ((2 * 263) / 279 + (30 * 237) / 310) / ((42 * 16) / 279 + (107 * 73) / 310) = 0.90$$

Beispiel für Selektionsfehler (Berkson's Fallacy)

In einer **Bevölkerung** von 1000 Personen gilt:

	CA	kein CA	Gesamt
exponiert	50	150	200
nicht expon.	200	600	800
Gesamt	250	750	1000

$$RR = 1.0$$

Alle Personen mit CA werden in die Klinik aufgenommen. Bei Personen ohne CA beträgt die Aufnahmequote 10% für Exponierte und 40% für Nicht-Exponierte. Es ergibt sich für die **Klinik**:

	CA	kein CA	Gesamt
exponiert	50	15	65
nicht expon.	200	240	440
Gesamt	250	255	505

$$RR = 1.7$$

Durch die unterschiedliche Aufnahmequote für andere Erkrankungen (kein CA) von Exponierten und Nicht-Exponierten wird ein erhöhtes relatives Risiko vorgetäuscht.

Deshalb: **Vorsicht bei Risikoermittlung aus Krankenhausdaten !**

Fall-Kontrollstudien

Anwendung: zur Risikoermittlung bei seltenen Ereignissen

Prinzip:

Es werden alle "Fälle" eines Gebietes (z.B. alle Frauen mit Lungen CA) erfaßt. Zusätzlich werden vergleichbare "Kontrollen" (z.B. Frauen ohne Lungen CA) des Gebiets erfaßt. Bei allen wird festgestellt (z. B. erfragt), ob eine Exposition vorliegt oder nicht. Zwischen den Fällen und Kontrollen wird die Expositionshäufigkeit verglichen.

Maß für relatives Risiko ist die **Relative Quote**, d.i. das Verhältnis:

$$\frac{(\text{Exponiert/Nicht Expon.})_{\text{Fälle}}}{(\text{Exponiert/Nicht Expon.})_{\text{Kontr}}}$$

Analog zur Kohortenstudie werden die Ergebnisse der Fall-Kontrollstudie in einer Vierfeldertafel dargestellt:

	Fälle	Kontrollen	Gesamt
Exponierte	a	b	n ₁
Nicht-Exponierte	c	d	n ₂
Gesamt	m ₁	m ₂	N

a = exponierte Fälle b = exponierte Kontrollen,
c = nicht-exponierte Fälle d = nicht exponierte Kontrollen
m₁ = Fälle, m₂ = Kontrollen, n₁ = Exponierte, n₂ = Nicht-Exponierte; N = Gesamtzahl.

Beispiel: Fall-Kontroll-Studie von Garfinkel (1985)

Fälle: Nichtraucher Frauen mit Lungenkrebs aus 4 Krankenhäusern (USA)

Kontrollen: Nichtraucher Frauen mit Colon CA oder Rektum CA

Exposition: Passiv-Rauchen (rauchender Ehemann):

	Fälle	Kontrollen	Gesamt
Exponierte	92	266	358
Nicht-Exponierte	42	136	178
Gesamt	134	402	536

$$\text{Relative Quote RO} = (92/42)/(266/136) = (92 \cdot 136)/(266 \cdot 42) = 1.12$$

Konfidenzintervall für RO_{wahr} bei Fall-Kontrollstudien:

Die beobachtete relative Quote RO=(a.d)/(b.c) ist ein Schätzwert für die "wahre" relative Quote RO_{wahr}. Die Zuverlässigkeit des Schätzwerts wird durch das Konfidenzintervall zu einer vorgegebenen Konfidenzwahrscheinlichkeit 1-α ausgedrückt. Dies lässt sich einfacher für den (natürlichen) Logarithmus ln RO berechnen, der approximativ normal verteilt ist mit dem Mittelwert RO_{wahr} und dem Standardfehler:

$$S_{\ln RO} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Daraus berechnet sich der Konfidenzfaktor: F = exp(z_{1-α/2} · S_{lnRO}), wobei z_{1-α/2} die (1-α/2)-Quantile der Standard-Normalverteilung ist (bei α=0.05 ist z_{1-α/2} ≈ 2).

Das Konfidenzintervall ist:

$$CI(1-\alpha) = (RO/F \text{ bis } RO \cdot F).$$

Beispiel: a = 92 b = 266 c = 42 d = 136

$$RO = (92 \cdot 136) / (266 \cdot 42) = 1.12$$

$$s_{\ln RO} = \sqrt{0.011 + 0.004 + 0.024 + 0.007} = 0.2145$$

$$F = \exp(1.96 \cdot 0.2145) = 1.52$$

95%-Konfidenzintervall: (0.74 bis 1.70)

Anmerkung: f = F-1 = 0.52 = entspricht der relativen Variabilität von RO.

Test der Nullhypothese: RO_{wahr} = 1 (kein erhöhtes Risiko):

$$\chi^2 = \frac{(a \cdot d - b \cdot c)^2 \cdot N}{m_1 \cdot m_2 \cdot n_1 \cdot n_2}$$

Ablehnen der Nullhypothese (Schwelle $\alpha = 0.05$), wenn: $\chi^2 > 3.8$

Beispiel: $\chi^2 = 0.28$; Annahme der Nullhypothese !

Berücksichtigung von Subgruppen (Stratifikation)

Wenn z.B. die Altersverteilung bei Fällen und Kontrollen verschieden ist, dann ist die relative Quote RO verzerrt. Durch Stratifikation nach dem Verzerrungsfaktor (Alter), der Berechnung von RO_j innerhalb der homogenen Strata j und geeignetes Zusammenfassen der RO_j zu einem gemeinsamen Schätzwert (Mantel-Haenszel), kann die Verzerrung ausgeglichen werden.

Beispiel: Berücksichtigung von Altersgruppen

Alter	Fälle			Kontrollen			N _j	RO _j
	m _{1j}	a _j	c _j	m _{2j}	b _j	d _j		
<60	33	23	10	103	68	35	136	1.18
60-79	72	50	22	209	138	71	281	1.16
≥80	29	19	10	90	60	30	119	0.95
Gesamt	134	92	42	402	266	136	536	1.12

Es bedeuten für die Gruppe j:

m_{1j} = Fälle; a_j = exponierte Fälle; c_j = nicht exponierte Fälle;
m_{2j} = Kontrollen; b_j = exponierte Kontrollen; d_j = nicht exponierte Kontrollen;
N_j = Gesamtzahl der Patienten (Fälle+Kontrollen) in Gruppe j;
RO_j = Relative Quote in Gruppe j.

Berechnen eines "unverzerrten Schätzwertes" für die 'wahre' relative Quote RO_{wahr} (Mantel-Haenszel-Schätzwert):

Annahme:

Die Risiken sind in den Altersgruppen verschieden, die "wahren" relativen Quoten RO_{wahr} aber gleich. Ein Schätzwert für RO_{wahr} ist:

$$RO_{MH} = \frac{\sum_j \frac{a_j d_j}{N_j}}{\sum_j \frac{b_j c_j}{N_j}} \quad (N_j = m_{1j} + m_{2j})$$

Beispiel:

$$RO_{MH} = (23 \cdot 35 / 136 + 50 \cdot 71 / 281 + 19 \cdot 30 / 119) / (68 \cdot 10 / 136 + 138 \cdot 22 / 281 + 60 \cdot 10 / 119) = (5.92 + 12.63 + 4.79) / (5.00 + 10.80 + 5.04) = 23.24 / 20.84 = 1.115$$

Test der Nullhypothese: RO_{wahr} = 1 (kein erhöhtes Risiko):

Testgröße:

$$X^2_{MH} = \frac{(\sum_j \frac{a_j d_j - b_j c_j}{N_j})^2}{\sum_j \frac{m_{1j} m_{2j} n_{1j} n_{2j}}{N_j^2 (N_j - 1)}}$$

a_j = exponierte Fälle, b_j = exponierte Kontrollen in Gruppe j,
c_j = nicht-exponierte Fälle d_j = nicht-exponierte Kontrollen in Gruppe j;
n_{1j} = Exponierte, n_{2j} = Nicht-Exponierte, m_{1j} = Fälle, m_{2j} = Kontrollen in Gruppe j.

Ablehnen der Nullhypothese (Schwelle α = 0.05), wenn: X² > 3.8

Beispiel: Berechnung von X²_{MH} für Studie von Garfinkel.

Zähler:

$$((23 \cdot 35 - 68 \cdot 10) / 136 + (50 \cdot 71 - 138 \cdot 22) / 281 + (19 \cdot 30 - 60 \cdot 10) / 119)^2 = (0.919 + 1.829 + 0.252)^2 = 6.230$$

Nenner:

$$33 \cdot 103 \cdot 91 \cdot 45 / (136^2 \cdot 135) + 72 \cdot 209 \cdot 188 \cdot 93 / (281^2 \cdot 280) + 29 \cdot 90 \cdot 79 \cdot 40 / (119^2 \cdot 118) = 5.574 + 11.900 + 4.936 = 22.410$$

$$X^2_{MH} = 6.23 / 22.41 = 0.278$$

Ergebnis: Annahme der Nullhypothese !

95%-Konfidenzintervall für RO_{wahr}

Berechnen des Konfidenzfaktors F mit X_{MH} = √X²_{MH} :

$$\ln(F) = \ln(RO_{MH}) \cdot 2 / X_{MH} \quad F = (RO_{MH})^{\frac{2}{X_{MH}}}$$

95%-Konfidenzintervall: 95%-CI = (von RO_{MH}/F bis RO_{MH}·F)

Im Beispiel:

$$RO_{MH} = 1.115; X_{MH} = 0.527; F = 1.512; 95\%KI: (0.738 - 1.685)$$

Matched Pairs

Um die Vergleichbarkeit zwischen Fällen und Kontrollen zu verbessern, wird zu jedem Fall eine Kontrolle ausgewählt, die ihm bezüglich bestimmter Kriterien (Alter, Wohngebiet u.ä.) möglichst gut entspricht: "matched pair".

4 Möglichkeiten der Exposition (1=exponiert, 0=nicht exponiert):

Konkordante Paare:

a.) n_{11} Paare	b.) n_{00} Paare
Fall exponiert Kontrolle exponiert	Fall nicht exponiert Kontrolle nicht exponiert

Diskordante Paare:

c.) n_{10} Paare	d.) n_{01} Paare
Fall exponiert Kontrolle nicht exponiert	Fall nicht exponiert Kontrolle exponiert

Relative Quote bei "matched pairs"

Nur die diskordanten Paare tragen zur Risikoschätzung bei.

n_{10} = Anzahl der Paare: Fall exponiert, Kontrolle nicht exponiert

n_{01} = Anzahl der Paare: Fall nicht exponiert, Kontrolle exponiert

$$RO = n_{10}/n_{01}$$

Test der Nullhypothese: $RO_{\text{wahr}} = 1$.

Testmaß: $X^2 = (n_{10} - n_{01})^2 / (n_{10} + n_{01})$

Ablehnen der Nullhypothese, wenn $X^2 > 3.8$

Beispiel:

beide exponiert: $n_{11} = 27$	beide nicht exponiert: $n_{00} = 4$
nur Fall exponiert: $n_{10} = 29$	nur Kontrolle exponiert: $n_{01} = 3$

$$RO = 29/3 = 9.67$$

$$X^2 = (29-3)^2 / (29+3) = 676/32 = 21.125 \text{ Ergebnis: Nullhypothese ablehnen !}$$

Teil IV: Unterstützung von Diagnostik und Prognostik

Diagnostik und Klassifikation

Bei der Diagnostik soll auf Grund von Symptomen und Befunden eines Patienten eine Diagnose gestellt werden. Die Diagnosebezeichnungen sind seit mehr als 100 Jahren international standardisiert: International Classification of Diseases ICD. Bei der Diagnostik handelt es sich um ein Zuordnungs- oder Klassifikationsproblem, das aus Erfahrung gelöst werden soll:

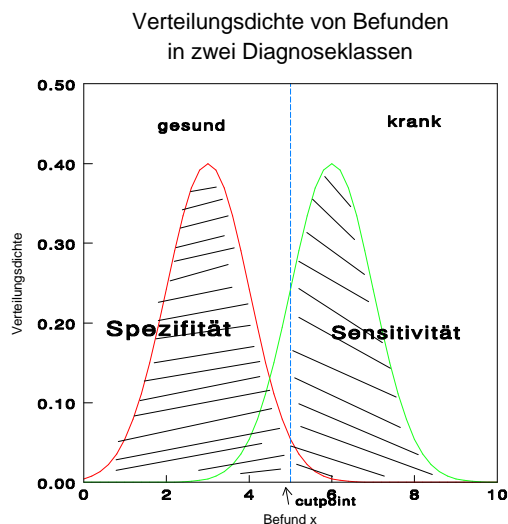
Erfahrung \rightarrow (Befund $x \rightarrow$ Diagnose d)

"Aus der Erfahrung folgt: Wenn Befund x vorliegt, dann gilt Diagnose d ".

Statistische Klassifikation:

Annahme: Die Befunde sind Zufallsgrößen, deren Verteilung von der Diagnoseklasse abhängt.

Formal kann die Klassifikation als eine Unterteilung des "Befundraums" in so viele Teilbereiche, wie Diagnoseklassen vorhanden sind, angesehen werden. Jedem Teilbereich wird eine Diagnoseklasse zugeordnet. Die Krankheit eines Patienten wird in die Diagnoseklasse eingestuft, in deren Teilbereich die Befunde liegen.



Forderung: Die Wahrscheinlichkeit der Fehlklassifikation soll minimal sein.

Sensitivität und Spezifität

Im einfachsten Fall kann ein Befund 'positiv' (d.h. er deutet auf eine Krankheit) oder 'negativ' (d.h. er deutet nicht auf eine Krankheit) sein. In beiden Fällen kann die Krankheit tatsächlich vorliegen oder fehlen.

Sensitivität ist die Wahrscheinlichkeit für einen positiven Befund, wenn die Krankheit vorliegt: $= \text{Pr}(\text{Befund positiv} \mid \text{Krankheit liegt vor})$.

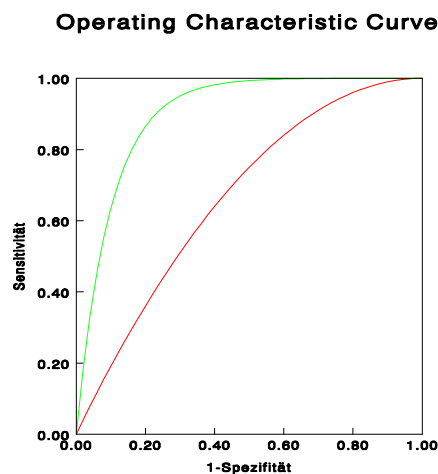
Spezifität ist die Wahrscheinlichkeit für einen negativen Befund, wenn die Krankheit fehlt: $= \text{Pr}(\text{Befund ist negativ} \mid \text{Krankheit fehlt})$.

Komplementär dazu sind die Anteile 'falsch positiver' und 'falsch negativer' Befunde:

Anteil falsch positiv ist die Wahrscheinlichkeit für einen positiven Befund, wenn die Krankheit fehlt: $\text{FP} = 1 - \text{Spezifität}$.

Anteil falsch negativ ist die Wahrscheinlichkeit für einen negativen Befund, wenn die Krankheit vorhanden ist: $\text{FN} = 1 - \text{Sensitivität}$.

Bei quantitativen Befunden x (z.B. Laborwerte) geschieht die Einteilung in 'positiv' oder 'negativ' durch Vorgabe eines 'cutpoints' x_0 . Für Befunde $x \leq x_0$ wird z.B. der Befund als 'negativ' bewertet, für Befunde $x > x_0$ als 'positiv'. Sensitivität und Spezifität hängen gegenläufig vom cutpoint ab: Bei Verkleinerung des cutpoints verringert sich die Spezifität und erhöht sich die Sensitivität, bei Vergrößerung des cutpoints erhöht sich die Spezifität und verringert sich die Sensitivität. Dieser Zusammenhang wird durch die OC-Kurve (Operating Characteristic Curve) dargestellt, in der auf der Ordinate die Sensitivität und auf der Abszisse der Anteil falsch positiver Befunde $\text{FP} = 1 - \text{Spezifität}$ aufgetragen werden:



Prädiktive Werte

Sensitivität und Spezifität sind bedingte Wahrscheinlichkeiten; d.h. sie geben jeweils die Wahrscheinlichkeit für eine richtige Klassifikation auf Grund des Befundes an, wenn die Krankheit tatsächlich vorliegt bzw. fehlt. Der praktische Wert eines diagnostischen Klassifikationsverfahrens wird durch zwei andere bedingte Wahrscheinlichkeiten ausgedrückt; nämlich durch den:

positiv prädiktiven Wert PW_+ d.i. die Wahrscheinlichkeit, daß die Krankheit vorliegt, wenn der Befund positiv ist

und den

negativ prädiktiven Wert PW_- d.i. die Wahrscheinlichkeit, daß die Krankheit fehlt, wenn der Befund negativ ist.

Diese prädiktiven Werte hängen von der **Prävalenz p** der Krankheit ab; d.i. die Wahrscheinlichkeit, mit der die Krankheit in der zu untersuchenden Population vorhanden ist. Bei bekannter Prävalenz p können die prädiktiven Werte aus der Sensitivität und Spezifität nach dem **Satz von Bayes** berechnet werden:

$$PW_+ = \Pr(\text{vorh} | \text{pos}) = \frac{\text{Sensitivität} \cdot \text{Prävalenz}}{\text{Sensitivität} \cdot \text{Prävalenz} + (1 - \text{Spezifität}) \cdot (1 - \text{Prävalenz})}$$

$$PW_- = \Pr(\text{fehlt} | \text{neg}) = \frac{\text{Spezifität} \cdot (1 - \text{Prävalenz})}{\text{Spezifität} \cdot (1 - \text{Prävalenz}) + (1 - \text{Sensitivität}) \cdot \text{Prävalenz}}$$

Bei gleicher Sensitivität und Spezifität ist der positive prädiktive Wert um so geringer und der negative prädiktive Wert um so größer, je geringer die Prävalenz für die Krankheit ist. Um z.B. bei einer Vorsorgeuntersuchung mit einem diagnostischen Verfahren (z.B. Mammografie) einen großen positiven prädiktiven Wert zu erhalten, sollte durch eine Vorauswahl der Patienten (z.B. nach bekannten Risikofaktoren) die Prävalenz für die zu diagnostizierende Erkrankung erhöht werden.

Ist von einer Anzahl von Personen einer Bevölkerung (einer Kohorte) bekannt, ob sie die Krankheit haben oder nicht und ob ihre Befunde positiv oder negativ sind, dann können Sensitivität, Spezifität und die prädiktiven Werte aus diesen Daten geschätzt werden. Die Häufigkeit der Erkrankten in dieser Kohorte ist ein Schätzwert für die Prävalenz.

Beispiele

Krankheit	Befund		Gesamt
	positiv	negativ	
vorhanden	60	20	80
fehlt	4	16	20
Gesamt	64	36	100

Sensitivität $60/80 = 0.75$ falsch negativ = 0.25
 Spezifität $16/20 = 0.80$ falsch positiv = 0.20

Prävalenz $80/100 = 0.80$

Prädiktive Werte
 positiver PW₊ $60/64 = 0.94$
 negativer PW₋ $16/36 = 0.44$

Krankheit	Befund		Gesamt
	positiv	negativ	
vorhanden	15	5	20
fehlt	16	64	80
Gesamt	31	69	100

Sensitivität $15/20 = 0.75$ falsch negativ = 0.25
 Spezifität $64/80 = 0.80$ falsch positiv = 0.20

Prävalenz $20/100 = 0.20$

Prädiktive Werte
 positiver PW₊ $15/31 = 0.48$
 negativer PW₋ $64/69 = 0.93$

Krankheit	Befund		Gesamt
	positiv	negativ	
vorhanden	9	1	10
fehlt	91	899	990
Gesamt	100	900	1000

Sensitivität $9/10 = 0.90$ falsch negativ = 0.10
Spezifität $899/990 = 0.91$ falsch positiv = 0.09

Prävalenz $10/1000 = 0.01$

Prädiktive Werte positiver PW₊ $9/100 = 0.09$
negativer PW₋ $899/900 = 0.99$