Medizinische Biometrie und Statistik

Inhaltsverzeichnis:

1) Studienplanung

- Beispiel 1: Salk-Polio-Studie
  - Präventionsstudie im öffentlichen Gesundheitswesen der USA zur Klärung der Frage, ob mit einer speziellen Impfung die Reduktion der Inzidenz der Poliomyelitis erreicht werden kann
  - **Planungsüberlegungen („Designfragen“)**
    - Definition der Zielpopulation, Ein- und Ausschlusskriterien
    - Logistischer Ablauf der Untersuchung, Standardisierung der Durchführung
    - Konzept des Studienaufbaus zur Klärung der Wirksamkeitsfrage der Impfung
  - **Vorschlag 1: Historischer Vergleich**
    Alle Kinder der definierten Altersstufe (2. Klasse der „Primary Schools“) werden in 1954 geimpft. Polio-Inzidenz der geimpften Gruppe wird mit der in früheren Jahren beobachteten Inzidenz, an Poliomyelitis zu erkranken, verglichen.
    Problem: Hohe Variabilität der Polio-Inzidenz über die Zeit, daher Verzerrung durch Zeiteffekte möglich
  - **Vorschlag 2: Prospektive Beobachtungsstudie**
    Allen Eltern wird die Impfung angeboten. Kinder von zustimmenden Eltern werden geimpft und mit den nicht geimpften Kindern von nicht zustimmenden Eltern verglichen.
  - **Vorschlag 3: Randomisierte Placebo kontrollierte Studie**
    Kinder von zustimmenden Eltern werden zufällig in zwei Gruppen aufgeteilt. Die eine Gruppe erhält eine Impfung mit dem zu evaluierenden Impfstoff von Jonas Salk, die andere Gruppe bekommt eine Placebo-Impfung mit einer Kochsalzlösung.
- In der Realität wurden die beiden Vorschläge 2 und 3 simultan (in unterschiedlichen Regionen der USA) umgesetzt. Dies bietet uns die Möglichkeit des empirischen Vergleichs der Auswirkungen der verschiedenen Designs.
  - **Ergebnisse:**
    Beobachtungsstudie (2) n Polio-Fälle (rel.)
    - geimpfte 221998 38 17/100000
    - nicht geimpfte 123605 43 35/100000
    - sonst. Kontrollen 725173 341 47/100000

    Randomisierte Studie (3)
    - geimpfte 200745 33 16/100000
    - Placebo 201229 119 59/100000
    - nicht zugestimmt 338778 121 36/100000

- Erkenntnispyramide

  Hierarchie der Aussagekraft von Studiendesigns bei der Bewertung von Therapien mit steigender Evidenz/Belegung der Wirksamkeit

  1. Fallstudien
  2. Fälle ohne Kontrollen
  3. Fälle mit Literaturkontrollen
  4. Auswertung von Register-Datenbanken
  5. Klinische Studien mit historischen Kontrollen
  6. Prospektive Beobachtungsstudien mit Kontrollgruppe
  7. Eine randomisierte kontrollierte klinische Studie (RCT)
  8. Mehrere randomisierte kontrollierte klinische Studien

  - Studien mit historischen Kontrollen

    Vorteile:
    ♦ Patienten müssen nicht einer möglicherweise schlechteren Kontrollbehandlung ausgesetzt werden
    ♦ geringerer Aufwand und Kosten
    Nachteile:
    ♦ Selektion der Kontrollen
    ♦ Allgemein: Zeiteffekte können verzerren, alle Probleme retrospektiver Erhebungen
  - Probleme von Auswertungen von Registern
    - Retrospektives Vorgehen steht oft vor dem Problem, dass nur unzureichend geklärt werden kann, weshalb ein Patient in der dokumentierten Form behandelt wurde.
    - Beispiel: EORTC (european organisation of research and treatement of cancer)-Register für Patienten mit Schilddrüsenkarzinom

- Allgemeines Dilemma der nicht randomisierten Studien: Die zu vergleichenden Gruppen unterscheiden sich in für die Prognose relevanten weiteren Merkmalen

- **Fachbegriff:** „Confounding“ (siehe auch 39)
  Diese „Vermengung“ (deutsche Übersetzung von „Confounding“) von auf die Zielgröße der Studie einflussnehmenden Effekten kann die Evaluation der Studienfragestellung in unterschiedlicher Größenordnung verzerren.

- Beispiel 2: Randomisiert klinische Studien

  1948 wurde in Großbritannien die erste randomisierte kontrollierte klinische Studie zur Evaluation der Wirkung einer Streptomycin-Behandlung der Tuberkulose durchgeführt. Dieser Studientyp trat schnell seinen Siegeszug in der klinischen Medizin an und ist heute weltweit der akzeptierte Standard für die Beurteilung der Wirksamkeit medizinischer Maßnahmen. Unterstützt wurde diese Entwicklung von behördlichen Regelungen bei der Zulassung von Arzneimitteln (FDA, BfArM).

- Randomisierungstechniken
  - Einfache Randomisation
    - Für jeden Patienten wird – unabhängig von vorherigen Zuordnungen – zufällig eine Zuteilung zu einer Behandlungsgruppe realisiert, wobei alle Behandlungsgruppen mit gleicher Wahrscheinlichkeit ausgewählt werden können.
    - Konkret kann dies geschehen durch:
      - Münzwurf (bei zwei Gruppen)
      - Würfeln (bei bis zu sechs Gruppen)
      - Zufallszahlentabelle (beliebige Gruppenzahl)
      - Computer-Randomisation (Methode der Wahl)
  - Blockrandomisation
    - Manchmal: Randomisation mit Ausgleich
    - Nach einer vorgegebenen Zahl von Patienten sollen die Behandlungs**gruppen gleich groß** sein („balancierte Aufteilung“)
  - Geschichtete Randomisation (z.B. nach männlich/weiblich oder nach Schweregrad)

    Definition von „Schichten“ bzgl. wesentlicher prognostischer Variablen und separate Blockrandomisation in den Schichten, so dass in den Schichten eine balancierte Aufteilung in den Behandlungsgruppen erzwungen wird (zur sicheren Vermeidung eines zufälligen Ungleichgewicht hinsichtlich wichtiger prognostisch relevanter Faktoren).

- Beobachtungsmethoden in Studien
  - Offen: Zuordnung des Studienteilnehmers in eine Studiengruppe allen bekannt
  - Einfachblind: Zuordnung des Studienteilnehmers nur Teilnehmern selbst nicht bekannt
  - Zweifachblind: Zuordnung des Studienteilnehmers Teilnehmern und Untersuchern nicht bekannt
  - Dreifachblind: Zuordnung des Studienteilnehmers Teilnehmern, Untersuchern und auswertenden Statistikern nicht bekannt

- Phasenmodell klinischer Studien
  - Phase 0: Präklinische Entwicklung
    Ziele:
    1. Abklärung möglicher toxischer Effekte, wie Einfluss auf zahlreiche in Laboruntersuchungen bestimmte Größen (Klinische Chemie, Hämatologie), Fertilität, Embryotoxizität/Teratogenität, Cancerogenität
    2. Abklärung sicherheitspharmakologischer Aspekte, wie Beeinträchtigung von Herz/Kreislauf, Einfluss auf Körpergewicht
    3. Hinweise auf erwünschte pharmakologische Effekte in vitro/in vivo

- Phase I: Erstanwendung am Menschen
  Meist gesunde Freiwillige („Probanden“), gegebenenfalls besondere Patientengruppe (z. B. bei Studien mit Zytostatika)
  Ziele:
  - Verträglichkeit, Pharmakokinetik/-dynamik
  - Hinweis auf wirksame Dosis (eventuell)/ Arzneimittelinteraktionen
- Phase II: Einstieg in die therapeutische Anwendung am Patienten
  Begrenzte Zahl von Patienten der anvisierten Indikation
  Ziele:
  • Verträglichkeit und Dosisfindung
  • Wirkung (pharmakologische Effekte)/Wirksamkeit (Heilerfolg)
  • Pharmakokinetik in Spezialfällen (z. B. Leber-, Nierenerkrankung)
- Phase III: Breite Anwendung im anvisierten Indikationsgebiet, Beleg für die Einsetzbarkeit als Arzneimittel [Zulassung]
  Patienten der anvisierten Indikation in Klinik/Praxis
  Ziele:
  - Beleg der Wirksamkeit an Patienten in unterschiedlichen Populationen
  - Ausreichende Beurteilung der Verträglichkeit, besondere Patientengruppen
  - Verhalten unter Langzeitbehandlung, Vergleich mit etablierter Therapie
- Phase IV: Klinische Prüfung nach der Zulassung: Erkenntniserweiterung über die Substanz, Einsatz unter Praxisbedingungen
  Einsatz an großer Zahl von Patienten entsprechend den Vorgaben der Zulassungsbehörden (unter Praxisbedingungen)
  Ziele:
  - Quantifizierung seltener Nebenwirkungen
  - Detailuntersuchungen in bestimmten Patientengruppen
  - Einfluss auf Spätfolgen einer Erkrankung (Folgemorbidität, Letalität)
  - Tatsächlicher Einsatz des Präparates (→ „Anwendungsbeobachtungen“)
  - Hinweis auf weitere Indikationen, zu modifizierende Dosis (→ Phase II)

2) Deskriptive (= beschreibende) Statistik

- Definition: Aufgabe der deskriptiven Statistik ist es, die in den Daten einer Stichprobe enthaltene relevante Information in Tabellen, Grafiken und statistischen Maßzahlen übersichtlich und in einem der Fragestellung angepassten Format zusammenzufassen.
  Im Rahmen von Studien z.B. zur ...
  - Charakterisierung einer Studienpopulation
  - Überprüfung der Verteilung von Randomisierungsmerkmalen auf Therapiegruppen
  - Beschreibung der Verteilung von interessierenden Zielgrößen
- Beispiel 3: Lübecker Blutdruckstudie
  - Ziel: Untersuchung der Häufigkeit sowie des Bekanntheits- und Behandlungsgrades der Hypertonie in der Lübecker Bevölkerung
  - Studientyp: Querschnittsstudie
  - Studienpopulation: 3100 Lübecker Bürger deutscher Nationalität im Alter von 30-69 Jahren
  - Stichprobengewinnung: Zufallsstichprobe aus dem Einwohnermeldeamt
  - Studienbeteiligung:
    - 2833 von 3100 Probanden erreichbar
    - 2359 von 2833 Probanden nahmen teil
  - Datenerhebung:
    - Blutdruckmessung durch standardisiertes Messverfahren (3mal)
    - Frage, ob Blutdruck bekannt
    - Frage, ob Blutdruck behandelt

- Blutdruckbewertung: WHO-Klassifikation
  - normoton: systol. Blutdruck < 140 mmHg und diastol. Blutdruck < 90 mmHg
  - grenzwertig: systol. Blutdruck 140 - 159 mmHg und/oder diastol. Blutdruck 90 - 94 mmHg
  - hyperton: systol. Blutdruck ≥ 160 mmHg und/oder diastol. Blutdruck ≥ 95 mmHg
- Wichtige Begriffe:
  - Beobachtungseinheit (z.B. Proband (Obs))
  - (Merkmalsträger)
  - Merkmal (z.B. SBD)
  - Merkmalsausprägung (z.B. 160 mmHg)
  - Stichprobenumfang n: Anzahl der Beobachtungseinheiten (n = 2359)

- Merkmalstypen
  - Unterscheidung nach Skalentypen:
    - Intervall- oder metrische Skala:
      - eindeutige Reihenfolge
      - Differenzen interpretierbar
      - Merkmal gemessen oder gezählt
      - Beispiele: Blutdruck (in mmHg), Kinderzahl
    - Ordinalskala:
      - eindeutige Reihenfolge
      - Differenzen nicht interpretierbar (d.h. unbekannte oder schwankende Abstände)
      - Merkmal i.d.R. abgeleitet; z.T. „willkürliche“ Ränge
      - Beispiele: WHO-Blutdruckklassen, „Scores“
    - Nominal- oder Kategorialskala:
      - keine eindeutige Reihenfolge (z.B. BBK-Status, Geschlecht)
      - dichotomes Merkmal: nur 2 Ausprägungen

    → „Umwandlungshierarchie“ (nur eine Richtung !): metrisch → ordinal → nominal → dichotom
  - Unterscheidung qualitativ-quantitativ:
    - quantitativ: Merkmalsausprägungen sind auf Zahlen angewiesen (z.B. Blutdruck in mmHg)
    - qualitativ: Merkmalsausprägungen sind nicht auf Zahlen angewiesen (z.B. BBK-Status)
  - Unterscheidung stetig-diskret: (nur quantitativ)
    - diskret: quantitatives Merkmal mit Ausprägungen, die nur vereinzelte Zahlenwerte annehmen können, d.h. es gibt keine Zwischenwerte (z.B. Kinderzahl)
    - stetig: quantitatives Merkmal mit Ausprägungen, die prinzipiell alle Zahlenwerte in einem bestimmten Intervall annehmen können. Die exakte Übereinstimmung der Merkmalsausprägungen ist sehr unwahrscheinlich (z.B. Blutdruck in mmHg)
  - z.B. Blutdruck (in mmHg): Intervallskala – quantitativ – stetig
- Häufigkeiten
  - Zusammenfassung qualitativer Daten → Häufigkeiten der Merkmalsausprägung
    - Absolute Häufigkeit (Habs; durch Zählung):
      Habs = Anzahl Beobachtungseinheiten mit Ausprägung y
    - Relative Häufigkeit (Hrel; zum Stichprobenumfang in Beziehung gesetzt):

      $$Hrel = \frac{\text{Anzahl Beobachtungseinheiten mit Ausprägung y}}{\text{Anzahl aller Beobachtungseinheiten}} = \frac{ny}{n}$$
- Grafische Darstellung qualitativer Merkmale
  - Kreisdiagramme: der zentrale Winkel eines Kreissegmentes ist proportional zur relativen Häufigkeit der zugehörigen Merkmalsausprägung; gesamte Kreis entspricht 100%; Nicht gut für ordinale Merkmale geeignet, da Kreise keine Richtung haben
    Block- oder Stabdiagramme; Höhe eines Blocks (Stabs) entspricht der relativen Häufigkeit der zugehörigen Merkmalsausprägung; gesamte Diagramm muss nicht 100% entsprechen; auch für ordinale Merkmale geeignet

- Grafische Darstellung quantitativer Merkmale
  - Quantitative (v.a. stetige) Merkmale haben oft eine Vielzahl von Ausprägungen. Mögliches Vorgehen bei der Deskription:
    - Bildung von **Klassen** (eindeutiges Definieren von Intervallgrenzen; z.B. WHO-Blutdruckklassen)
    - Angabe von absoluten und relativen Häufigkeiten

  - **Histogramm**
    - X-Achse: Merkmalsausprägungen (darauf die Klassen als unmittelbar benachbarte Blöcke)
    - Y-Achse: Hrel (oder Habs) der Klassen
    - Bei gleichen Klassenbreiten: Höhe der Blöcke proportional zu Hrel der Klassen (bzw. Habs)
    - Bei ungleichen Klassenbreiten: Jeweils Dividieren der Klassen-Hrel durch die Klassenbreite;
    - Y-Achse: Hrel pro Merkmalsausprägung der Klasse

    → Fläche der Blöcke ist proportional zu Hrel der Klasse
  - Die Anzahl der Klassen hat großen Einfluss auf das Erscheinungsbild. Fehlen objektive Kriterien, so hilft folgende Faustformel für die **optimale Anzahl *K* der Klassen**:

$$K = \begin{cases} \sqrt{n} & \text{falls } n \leq 1000 \\ 10*\log_{10}(n) & \text{falls } n > 1000 \end{cases}$$

  - Aus Histogrammen lässt sich die Hrel einzelner Messwerte bzw. einzelner Klassen ablesen. **Kumulative (= summierte) relative Häufigkeiten** sind in der Praxis jedoch häufig interessanter:
    - Wieviel Prozent der Probanden der Blutdruckstudie sind normoton gemäß der Hypertonieskala des Joint National Committee (DBD < 90mmHg)?
    - Wieviel Prozent der Probanden sind schwach (90 ≤ DBD <105), mäßig schwer (105 ≤ DBD < 115), schwer hyperton (DBD ≥ 115)?
  - Häufigkeitstabelle mit kumulierten Habs und Hrel des DBD:
    - Sortieren der Merkmalsausprägungen nach Größe
    - Für die vorhandenen Merkmalsausprägungen des DBD jeweils Berechnung der Probandenzahl (bzw. des Anteils an Probanden) mit einem DBD-Wert, der kleiner oder gleich dem betrachteten Wert ist.
  - Grafische Darstellung kumulierter Häufigkeiten
    - Empirische Verteilungsfunktion:
      - sie weist jeder möglichen Zahl x den Prozentsatz aller beobachteten Ausprägungen des Merkmals zu, deren Zahlenwert kleiner oder gleich x ist.
      - ist eine Treppenfunktion
      - Funktion springt bei solchen x-Werten, die mindestens einmal beobachtet wurden (Treppenstufen)
      - Die Höhe der Sprünge entspricht der relativen Häufigkeit der betreffenden Ausprägung
    - Quantile
      - Zusammenfassung der Häufigkeitsverteilung
      - Ein q-Quantil ist eine Zahl Q, welche die sortierten Merkmalsausprägungen
      - dergestalt teilt, dass gilt:
        ≥ 100*q % der Merkmalsausprägungen sind ≤ Q
        ≥ 100*(1-q) % der Merkmalsausprägungen sind ≥ Q
      - Beispiel: Das 25%-Quantil ist die Zahl, bei der mindestens 25% der Merkmalsausprägungen kleiner oder gleich und mindestens 75% der Merkmalsausprägungen größer oder gleich dieser Zahl sind.
      - Ausgewählte Quantile:

q	≤ Q	≥ Q	Name des zugehörigen Quantils		
0,05	5%	95%	5. Perzentil		
0,25	25%	75%	25. Perzentil	1. Quartil	
0,50	50%	50%	50. Perzentil	2. Quartil	Median
0,75	75%	25%	75. Perzentil	3. Quartil	
0,95	95%	5%	95. Perzentil		

- Boxplot (Box and Whiskers Plot)
  - Übersichtliche graphische Darstellung von quantitativen Merkmalen, in der fünf wichtige Kenngrößen der Verteilung visualisiert werden:
    - Minimum
    - Maximum
    - 25%-, 50%- (Median) und 75%-Quantil

- Zusammenfassung Diagramme

Qualitativ	nominales Merkmal:	Kreis- oder Blockdiagramm
	ordinales Merkmal:	vornehmlich Blockdiagramm
quantitativ	ohne kumulative Information:	Histogramm
	mit kumulativer Information:	grob: Boxplot
		exakt: Verteilungsfunktion

- Diagramme/Achsen immer beschriften
- Immer den Umfang der Grundgesamtheit angeben
- Dreidimensionale Diagramme vermeiden

- Statistische Maßzahlen
  - Zusammenfassung der Datenmenge mit Hilfe weniger Maßzahlen
  - Lagemaße
    - Wichtige Maße zur Kennzeichnung der zentralen Lage einer Merkmalsverteilung sind:
      - **Arithmetisches Mittel (Mittelwert)**
        - Sind *x1*, ..., *xn* die an *n* Probanden beobachteten Ausprägungen eines Merkmals, so berechnet sich das arithmetische Mittel als:

$$\bar{x} = \frac{(x_1 + x_2 + ... + x_n)}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

        - **Beispiel**: Mittelwert des DBD von 7 Probanden Urliste in mmHg

$x_1 = 90$ $X_2=75$ $x_3=80$ $x_4=60$

$x_5=100$ $X_6=85$ $x_7=100$ → $\bar{x} = \frac{(90 + 75 + ... + 100)}{7} = \frac{575}{7} = 82{,}14$

        - Berechnung des arithmetischen Mittels nur bei **metrischen Merkmalen** erlaubt
        - empfindlich gegenüber extremen Werten, was v.a. bei kleinen Stichproben ins Gewicht fällt.
        - es gibt noch weitere Mittelwerte (geometrische, harmonische), die aber selten verwendet werden.
      - **Median**
        - Der Median x~ ist der „mittlere Wert“ der nach Größe sortierten Merkmals-ausprägungen. In der Rangreihe befinden sich oberhalb des Medians genauso viele Elemente wie unterhalb. Er entspricht dem **50%-Quantil**.
        - Bei einer geraden Anzahl von Merkmalsausprägungen wird meist der Mittelwert aus den beiden mittleren Werten als Median bezeichnet (s. Abschnitt „Quantile“).
        - Der Median kann bei metrisch und ordinal skalierten Merkmalen gebildet werden.
        - Er ist robust gegenüber Ausreißern.
        - Bei einer völlig symmetrischen Verteilung der Merkmalsausprägungen sind Mittelwert und Median identisch.
      - **Modalwert (Modus)**
        - häufigste Ausprägung eines Merkmals
        - bei allen Merkmalstypen ermittelbar
  - Streuungsmaße
    - Streuungsmaße dienen der Quantifizierung des Grades der **Variabilität** der beobachteten Ausprägungen in der Stichprobe.
      Stichprobe A: n = 10; = 5,7; „kleine Streuung“

Stichprobe B: n = 10; = 5,7; „große Streuung"
→ Mittelwert allein beschreibt die Daten nicht ausreichend.

- Varianz
  - Summe der quadrierten Abweichungen jedes Einzelwertes vom Mittelwert, dividiert durch *n-1* wenn *x1*, ...,*xn* die an *n* Probanden beobachteten Ausprägungen eines Merkmals

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

  - Empfindlich gegenüber Ausreißern
- Standardabweichung
  - Quadratwurzel der Varianz $s = \sqrt{s^2}$
  - Standardabweichung hat die gleiche Dimension/Einheit wie das Merkmal
  - Empfindlich gegenüber Ausreißern
- Spannweite
  - r = Maximum − Minimum
  - anschaulich; Extremwerte an sich wichtige Information
  - aber: Die Spannweite ist von nur zwei Werten (oft Ausreißer!) abhängig, daher schlechte Beschreibung der Daten insgesamt; als Streuungsmaß ungeeignet
- Interquartilsabstand
  - q = 25% – 75%-Quartil (→ Boxplot)
  - wie Interperzentilbereich robust gegenüber Ausreißern
  - auch bei ordinalen Merkmalen verwendbar
- Variationskoeffizient
  - $V = \frac{s}{\bar{x}}$ = Standardabweichung / Mittelwert
  - Standardabweichungen in Einheiten des Mittelwertes
  - Zum Vergleich der Variabilität verschiedener Verteilungen geeignet

3) Korrelation

- Motivation
  - Die Verfahren der Korrelations- und Regressionsanalyse dienen der **Beschreibung eines Zusammenhanges** zwischen zwei (oder mehreren) Merkmalen.
  - Die **Korrelationsanalyse** untersucht den (ungerichteten) Zusammenhang **zweier gleichberechtigter Merkmale** und quantifiziert die **Stärke** des Zusammenhangs. Sie wird primär in der deskriptiven Statistik eingesetzt, wenn man die Richtung des Zusammenhanges nicht kennt bzw. diese nicht von Interesse ist.
  - Die zu einem späteren Zeitpunkt vorgestellte **Regressionsanalyse** dient dabei zur Beschreibung der **Art** eines (gerichteten) Zusammenhanges (Je-Desto-Beziehung).
  - Beispiel:
    Im Rahmen einer Untersuchung an 149 Probanden wurde die Anzahl der Nävi (=Muttermale) an beiden Oberarmen von den Probanden selbst und von einem geschulten Untersucher erhoben.
    Fragen: Wie analysiert man den Zusammenhang zwischen den beiden Messgrößen? Wie beschreibt man quantitativ die Stärke des Zusammenhangs?
  - Grafische Darstellung
    - Boxplot

      Zur deskriptiven Aufbereitung werden die Verteilungen der Nävizahlen in separaten Boxplots dargestellt. Damit wird keine Information zum Zusammenhang gegeben.

- Scatterplot

  Zur korrekten deskriptiven Aufbereitung erfolgt die grafische Darstellung des Zusammenhangs zweier quantitativer Merkmale mit Hilfe eines Scatterplots („Punktewolke“).
  - **Je stärker** der Zusammenhang (oder die Korrelation), **desto schmaler** erstreckt sich die Punktwolke entlang einer gedachten Kurve.
  - Bei einer **starken Korrelation** liegen die meisten Punkte sehr **nahe an der Kurve** (im Falle eines linearen Zusammenhangs: an einer Gerade).
  - Eine **unstrukturierte Punktwolke** ist ein Hinweis auf einen **fehlenden Zusammenhang**.

- Korrelationskoeffizienten
  - Die Stärke eines Zusammenhanges zweier Variablen x und y soll mit Maßzahlen beschrieben werden.
  - Korrelationskoeffizient nach Pearson

$$r = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

    Voraussetzungen: linearer Zusammenhang, beide Merkmale sind quantitativ
  - Korrelationskoeffizienz nach Spearman
    - Eigenschaften
      - Maß für die **lineare Assoziation** zwischen zwei stetig-quantitativen Merkmalen
      - liegt im **Wertebereich von –1 bis +1**
      - wird **positiv**, wenn hohen Werten des einen Merkmals jeweils hohe Werte des anderen Merkmals entsprechen, d.h. die gedachte Gerade durch die Punktwolke hat eine positive Steigung, wird **negativ** bei fallender Gerade
      - nimmt Werte nahe 0 an, wenn keine lineare Beziehung vorhanden ist, und nimmt betragsmäßig Werte nahe 1 an, wenn ein starker linearer Zusammenhang besteht
      - Analog zum Korrelationskoeffizienten nach Pearson; Unterschied: ein breiter gefasstes Verständnis von Assoziation wird bei der Quantifizierung des Zusammenhanges zugrunde gelegt
    - Der Korrelationskoeffizient nach Spearman entspricht dem Korrelationskoeffizienten nach Pearson, beruht aber bei der Berechnung nicht auf den Originalmesswerten (xi, yi), sondern auf deren zugehörigen **Rangzahlen** (ri, si), die separat nach der Stellung der jeweiligen Beobachtung in der nach Größe geordneten Messreihe aller x und y Messwerte bestimmt werden.

$$r_S = \frac{\sum (r_i - \bar{r}) \cdot (s_i - \bar{s})}{\sqrt{\sum (r_i - \bar{r})^2 \cdot \sum (s_i - \bar{s})^2}}$$

    - Voraussetzungen
      - Der aus der Punktwolke vermutete Zusammenhang muss nur **monoton** sein (monoton wachsend oder fallend). Es ist kein linearer Zusammenhang notwendig.
      - Beide Merkmale sind mindestens **ordinalskaliert**.
    - Beispiel Nävi-Zählung:
      Zusammenhang zwischen Selbstzählung der Probanden und Untersucherzählung der Nävi:
      Korrelationskoeffizient nach Pearson: 0.817
      Korrelationskoeffizient nach Spearman: 0.792
  - Vergleich der Korrelationskoeffizienten

    Abb.

4) Wahrscheinlichkeitsrechnung

- Wahrscheinlichkeitsangaben in der Medizin
  - sind entweder„**empirisch-basiert**“, also nichts anderes als relative Häufigkeiten bei sehr großen Studienkollektiven oder „**modell-basiert**“, also aus rein theoretischen Überlegungen zu den Gesetzmäßigkeiten eines Modells resultierend. Es existieren auch Mischformen.
  - Modell basierte Wahrscheinlichkeiten am Beispiel der Genetik:
    - Kreuzungsexperiment: Intermediärer Erbgang; ein Gen mit den Allelen A und a. Reintypen (AA und aa) werden gekreuzt. → Mendelsche Gesetze
      - 50% (=0.5) haben Aa
      - 25% (=0.25) haben AA
      - 25% (=0.25) haben aa
  - Notationskonventionen für den Umgang mit Wahrscheinlichkeiten
    - **Versuchergebnisse („Ereignisse“)** werden mit **Großbuchstaben** (A,B,C,...) bezeichnet, wobei durchaus auch mehrere Versuchsausgänge zusammengefasst werden können.
    - **Wahrscheinlichkeiten** beziehen sich stets auf Ereignisse und werden mit **P(.)** abgekürzt.
      - **sicheres Ereignis (S)** : das Versuchsergebnis, das auf jeden Fall eintritt (also „mit Sicherheit“)
      - **Komplementärereignis** zum Ereignis A: das „Gegenteil“ des Ereignisses A, stets bezeichnet mit Aquer.
      - Fasst man A und Aquer zusammen, in Zeichen A ∪ Aquer, so entsteht immer ein sicheres Ereignis S.
  - **Rechenregeln** für Wahrscheinlichkeiten P(.)
    - $0 \leq P(A) \leq 1$
    - $P(S) = 1$, S ist sicheres Ereignis
    - $P(\text{Aquer}) = 1 - P(A)$
    - $P(A \cup B) = P(A) + P(B)$, falls $A \cap B$ leer ist
    - bzw. allgemein: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - **Unabhängigkeit** von Ereignissen
    - $P(A \cap B)$ ist auch die entscheidende Größe, um die Unabhängigkeit zweier Ereignisse zu beurteilen.
    - Inhaltlich bedeutet Unabhängigkeit: die beiden **Ereignisse haben „nichts miteinander zu tun“, „beeinflussen sich nicht gegenseitig**“ o.ä.
    - Formal bedeutet Unabhängigkeit: $P(A \cap B) = P(A) \bullet P(B)$
  - **Bedingte Wahrscheinlichkeit**
    - Falls zwei Ereignisse A und B **nicht unabhängig** sind, so stellt die Kenntnis des einen Ereignisses einen „Informationsgewinn“ für die Angabe der Auftretenswahrscheinlichkeit des anderen Ereignisses dar. Dies wird formalisiert durch den Begriff der bedingten Wahrscheinlichkeit.
    - $P(A \mid B) = P(A \cap B) / P(B)$ gelesen: A „gegeben“ B oder A „unter Bedingung“ B
    - Für unabhängige Ereignisse A und B gilt: $P(A \mid B) = P(A) \bullet P(B) / P(B) = P(A)$
  - **Wahrscheinlichkeitsverteilungen**
    - Zur theoretischen **Beschreibung des Ausgangs von zufallsbehafteten Experimenten**/Studien werden oft Wahrscheinlichkeitsverteilungen verwendet. Je nach Skalenniveau der interessierenden Zielgrößen bedeutet dies, dass modell-basierte Aussagen über die Wahrscheinlichkeit, mit der die Zielgröße einen bestimmten Wert innerhalb ihres Wertebereichs annimmt (diskreter Fall) oder in einem (Teil-)Intervall des Wertebereichs liegt (stetiger Fall), getroffen werden können.
    - In Anwendung häufig verwendete WKverteilungen
      - Im diskreten Fall:
        - Binomialverteilung

Die Zielgröße „Anzahl geheilter Patienten“ innerhalb einer Studie des Umfangs N=10 mit Wahrscheinlichkeit einer Heilung eines einzelnen Patienten von p=0.2 ist durch eine Binomialverteilung mit den Parameter N=10 und p=0.2 beschrieben.

Formel: $P(X = k) = n! / ((n-k)! \bullet k!) \bullet p^k \bullet (1-p)^{n-k}$

- ❖ Hypergeometrische Verteilung

➢ Im stetigen Fall:

- ❖ Normalverteilung
- ❖ t-Verteilung
- ❖ $\chi 2$ –Verteilung

➢ Zusätzlich zur obigen Angabe des Typs der Wahrscheinlichkeitsverteilung müssen in der konkreten Anwendung auch die jeweiligen Parameter der Verteilung spezifiziert werden.

- Beispiel 4: fiktiv
  - Cyclosporin A soll hinsichtlich seiner Sicherheit bei indikationskorrekter Anwendung überprüft werden. Dazu werden akute Nebenwirkungen bei 1000 nach Herstellerangaben korrekt behandelten Patienten dokumentiert. Es ergab sich folgendes Bild:
    - 150 Patienten klagten über Übelkeit (A = Übelkeit nach Einnahme von Cyclosporin A)
    - 200 Patienten gaben starke Kopfschmerzen an (B = starke Kopfschmerzen)
    - 50 Patienten berichteten über Erbrechen (C = Erbrechen)
  - Bisher deskriptive Statistik
    - Angabe von relativen Häufigkeiten:
      - ➢ Übelkeit: 15% (=0.15)
      - ➢ starke Kopfschmerzen: 20% (=0.2)
      - ➢ Erbrechen: 5% (=0.05)
    - Dies liefert eine empirische Beschreibung des Auftretens von Nebenwirkungen des Präparats in diesem Studienkollektiv. Der Hersteller des Präparats ist an allgemeineren Aussagen interessiert, er möchte die Wahrscheinlichkeit, des Auftretens bestimmter Nebenwirkungen bei indikationskorrekter Einnahme quantifizieren (→ „Waschzettel“).
  - Wahrscheinlichkeitsrechnung

    P(A) bezeichnet Wahrscheinlichkeit für Auftreten der Nebenwirkung Übelkeit; P(A) = 0.15 Aquer = keine Übelkeit; $P(A \cup Aquer) = 1$; P(B) = 0.2; P(C) = 0.05
    - Wie groß ist die Wahrscheinlichkeit eines Patienten, nach Einnahme von Cyclosporin A unter starken Kopfschmerzen oder Übelkeit zu leiden?
      Formal: $P(A \cup B) = ?$ → Nur falls beide Nebenwirkungen nie gemeinsam auftreten gilt: $P(A \cup B) = P(A) + P(B) = 0.2 + 0.15 = 0.35$ (sonst notwendig: Kenntnis von $P(A \cap B)$!)
    - Es sei aus der empirischen Studie zusätzlich bekannt, dass 30 der 1000 Patienten über starke Kopfschmerzen und Übelkeit sowie 50 Patienten über Übelkeit und Erbrechen geklagt haben. Somit können die „empirisch-basierten“ Wahrscheinlichkeiten wie folgt angegeben werden:
      $P(A \cap B) = 0.03 = 0.15 \bullet 0.2 = P(A) \bullet P(B)$
      $P(A \cap C) = 0.05 \neq 0.15 \bullet 0.05 = P(A) \bullet P(C)$
      $P(A \cap B) = 0.03$
    - Wie können wir in Aussagen über die Wahrscheinlichkeit des Auftretens des Ereignisses Erbrechen die Information über das Ereignis Übelkeit nutzen?
      $P(C \mid A) = P(C \cap A) / P(A) = 0.05 / 0.15 = 0.33$
      $P(C \mid –A) = P(C \cap –A) / P(–A) = 0 / 0.85 = 0$
    - Dagegen gilt:
      $P(B \mid A) = P(B \cap A)/P(A) = 0.03/0.15 = 0.2 = P(B)$

5) Diagnostische Tests

- Was ist ein diagnostischer Test?
  - In der klinischen Praxis werden vielfältige Vorgehensweisen in Form von mehr oder weniger komplexen diagnostischen Tests eingesetzt, um das Vorliegen einer Erkrankung zu beurteilen.
  - Was ist ein „guter“ diagnostischer Test?
    - Er findet möglichst alle tatsächlich Erkrankten.
    - Er identifiziert nicht erkrankte Personen als solche.
  - Zur Sicherung der Diagnose wird meist nicht nur ein sondern eine Kombination von mehreren Einzeltests durchgeführt.
- Diagnoseevaluation mittels Goldstandard
  - Der diagnostische Test kann nur im Falle des „genauen Wissens“ über das tatsächliche Vorliegen der Erkrankung auf Basis eines sogenannten „Goldstandards“ evaluiert werden.
  - Der Goldstandard kann ein etabliertes (mehr oder weniger gutes) Diagnoseverfahren sein.
  - Optimal ist das genaue Wissen über die tatsächliche Erkrankung z.B. durch die Autopsie.
- Validität
  - = Richtigkeit, Gültigkeit
  - Die gemessen Daten geben das wieder, was auch gemessen werden soll.
- Reproduzierbarkeit
  - = Wiederholbarkeit
  - Messungen führen bei Wiederholung zum selben Ergebnis
  - Synonym verwendete Begriffe: Präzision, Reliabilität, Zuverlässigkeit, Stabilität
- Vergleich von Validität und Reproduzierbarkeit
- Maßzahlen für die Güte eines diagnostischen Tests
  - Sensitivität = P(positiver Test | krank) = P(Positiver Test und krank) / P(krank)
  - Spezifität = P(negativer Test | gesund) = P(negativer Test und gesund) / P(gesund)
  - Es existiert eine Reihe von Maßzahlen, die aus Sensitivität und Spezifität gebildet werden und die Güte eines diagnostischen Tests in einem Wert zusammenfassen. Dazu gehört:
    Youden-Index = Sensitivität + Spezifität -1
    - bei „vernünftigen“ diagnostischen Tests zwischen Null und Eins.
    - Test ist „gut“ wenn der Youden-Index nahe bei Eins liegt.
    - Genau dann erreichen die Sensitivität als auch die Spezifität Werte nahe bei Eins.
- Allgemeine Darstellung (4-Felder-Tafel)

		tatsächlicher Zustand ja	tatsächlicher Zustand nein	
Test-ergebnis	ja	richtig positiv	falsch positiv	positiv
	nein	falsch negativ	richtig negativ	negativ
		krank	gesund	gesamt

  - Sensitivität = bedingte Wahrscheinlichkeit als krank identifiziert zu werden, gegeben man ist wirklich krank = Anzahl richtig positiver Personen / Anzahl aller kranken Personen
  - Spezifität = bedingte Wahrscheinlichkeit als gesund identifiziert zu werden, gegeben man ist wirklich gesund = Anzahl richtig negativer Personen / Anzahl aller gesunden Personen
- ROC-Kurven
  - Um den Zusammenhang von Sensitivität und Spezifität graphisch darzustellen, zeichnet man zuerst eine Receiver Operating Characteristic **(ROC)** Kurve.
  - Der **optimale Cutpoint** ist der Punkt, bei dem Kranke und Gesunde am besten getrennt werden. Er lässt sich in der Grafik als derjenige Punkt bestimmen, bei dem der Abstand zur Winkelhalbierenden am größten ist. Das ist gleichzeitig auch der Punkt, bei dem der Youden-Index maximal ist. In diesem Fall handelt es sich um den Cutpoint 213.
- Prädiktive Werte
  - Die beiden bislang eingeführten Maßzahlen der Validität eines diagnostischen Tests, **Sensitivität** und **Spezifität** beschreiben allein die Testeigenschaften und sind

prävalenzunabhängig (Prävalenz = Häufigkeit, mit der die untersuchte Zielerkrankung in der untersuchten Population vorkommt).

- Sie liefern deshalb auch keine Information über die klinische Verwertbarkeit des konkreten Resultats der diagnostischen Testung, da dies stets **prävalenzabhängig** gesehen werden muss.
- Hierzu ist die Kenntnis der **prädiktiven Werte** des diagnostischen Tests notwendig.
  - Der **positiv prädiktive Wert (PPW)** gibt an, wie hoch der Anteil der tatsächlich Kranken unter Patienten mit positivem Test ist.
    PPW = bedingte Wahrscheinlichkeit, tatsächlich krank zu sein, gegeben der Test ist positiv = P(krank | positiver Test) = P(krank und positiver Test) / P(positiver Test)
  - Der **negativ prädiktive Wert (NPW)** gibt an, wie hoch der Anteil der tatsächlich Gesunden unter Patienten mit negativem Test ist.
    NPW = bedingte Wahrscheinlichkeit, tatsächlich gesund zu sein, gegeben der Test ist negativ= P(gesund | negativer Test) = P(gesund und negativer Test) / P(negativer Test)

- Zusammenhang zwischen Maßzahlen
  - Der allgemeine Zusammenhang zwischen Sensitivität (SE), Spezifität (SP), Prävalenz der Erkrankung (P) und den prädiktiven Werten (PPW und NPW) lässt sich in den beiden folgenden Formeln zusammenfassen:
  - **Formel von Bayes**

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)} = \frac{P(A) \cdot P(B \mid A)}{P(A) \cdot P(B \mid A) + P(\bar{A}) \cdot P(B \mid \bar{A})}$$

  - Gegeben: A=krank und B=Test pos.
    $P=P(A)$, $SE=P(B|A)$, $1\text{-}SP=P(B \mid \bar{A})$
  - Gesucht: $PPW=P(A|B)$
  - Lösung:

$$PPW = \frac{SE \cdot P}{SE \cdot P + (1 - SP) \cdot (1 - P)}$$

$$NPW = \frac{SP \cdot (1 - P)}{SP \cdot (1 - P) + (1 - SE) \cdot P}$$

  - Grafische Darstellung:
    Mit Hilfe dieser Formeln lässt sich die Abhängigkeit der prädiktiven Werte positiver und negativer Tests von der Prävalenz bei bekannter Sensitivität und Spezifität graphisch darstellen: Abb.

- Kombination von diagnostischen Tests
  - In der Regel existieren mehrere diagnostische Tests, die zur Diagnosestellung einer Krankheit eingesetzt werden können. Wie beeinflusst die mögliche Kombination mehrerer diagnostischer Test die Güte der Diagnosestellung?
  - Es ist dahingehend zu unterscheiden, wie die kombinierte Information aus mehreren Tests im Falle widersprüchlicher Testergebnisse bewertet wird (falls alle Tests positiv (bzw. negativ) ausfallen, ist die Testentscheidung des kombinierten Vorgehens naheliegenderweise positiv (bzw. negativ)).
  - **„Believe-the-negative"-(BTN)-Regel:** Falls einer der Einzeltests negativ wird, ist die Testentscheidung des kombinierten Vorgehens ebenfalls negativ. Anders ausgedrückt: Um eine positive Testentscheidung zu treffen, müssen alle Einzeltests positiv sein.
  - **„Believe-the-positive"-(BTP)-Regel:** Falls einer der Einzeltests positiv wird, ist die Testentscheidung des kombinierten Vorgehens ebenfalls positiv. Anders ausgedrückt: Um eine negative Testentscheidung zu treffen, müssen alle Einzeltests negativ sein.
  - Die Strategie zur Auswertung mehrerer diagnostischer Einzeltests hat auch direkte Konsequenzen für eine effiziente praktische Durchführung der diagnostischen Testung.
  - **Serielles oder sequentielles Testen:** Die diagnostischen Einzeltests werden nacheinander durchgeführt. Nur positiv Getestete werden weiteren Testungen unterzogen (⇒ BTN-Regel).

- **Paralleles Testen:** Alle diagnostischen Einzeltests werden möglichst simultan durchgeführt. Für alle Getesteten liegen am Ende alle Einzeltestergebnisse vor (⇒ BTP-Regel).

- Beispiel 6: Creutzfeld-Jakob-Erkrankung
  - Bei der Creutzfeldt-Jakob-Erkrankung (CJD) handelt es sich um eine spongiforme Encephalopathie. Die Inzidenz (Anzahl der Neuerkrankungen in einem Jahr / Anzahl der Gesunden am Beginn des Jahres) beträgt 1 Fall pro 1 Million pro Jahr.
  - Das große Problem dieser Krankheit ist eine **Diagnose zu Lebzeiten**. Die definitive Diagnose von CJD findet durch den Nachweis des PRPSc (gestörte Isoform des Prion-Proteins PRP) im Gehirn bei der Autopsie statt. Nur so erhält man einen **gesicherten** CJD-Fall.
  - Seit 1993: CJD Surveillance Studie, die alle CJD-Verdachtsfälle bundesweit registriert, dokumentiert und weiterhin überwacht. (sammeln Gehirnproben)
  - Bis Dezember 1996 (Datengrundlage für das 1. Beispiel) wurden 289 Patienten Liquor (CSF-Proben) entnommen und analysiert. Davon war bei 238 Patienten eine klinische Diagnosestellung möglich.
  - Bis Dezember 1995 (Datengrundlage für das 2. Beispiel) wurden 224 Patienten Serum entnommen und analysiert. Davon war bei 182 Patienten eine klinische Diagnosestellung möglich.
  - Diagnostischer Tests für CJD:
    - Es wird ein diagnostischer Test gesucht, der genau die Situation abdeckt, in der der Test auch später bei dementen Patienten angewandt werden kann, um die CJD-Erkrankung von anderen Demenzen zu trennen. D.h. zuerst wird die klinische Diagnose gestellt und dann noch zusätzlich ein Liquor- oder Serum-Test durchgeführt.
    - **Idee:** Falls das 14-3-3-Protein im Liquor nachweisbar ist, dann handelt es sich um einen CJD-Fall
    - **Alternative:** Falls S100 im Serum in großer Höhe nachweisbar ist, dann handelt es sich um einen CJD-Fall
    - **Frage:** Welcher der beiden oben genannten Tests ist „gut“ ?
    - **Anmerkung:** Falls S100 ein „guter“ diagnostischer Test wäre, hätte sein Einsatz Vorteile, da S100 leichter bestimmbar ist.
    - 14-3-3 als diagnostischer Test

      „Herausfischen“ der Kranken 126/134 = 0.940

      „Herausfischen“ der Gesunden 97/104 = 0.933

      Häufig werden die Maßzahlen diagnostischer Tests als Prozentzahlen angeben. Im Beispiel ergibt sich: Sensitivität = 94.0%, Spezifität = 93.3%

      Diese Resultate unterstreichen, dass die Bestimmung des 14-3-3 Proteins im Liquor ein sehr guter diagnostischer Test zur CJD-Diagnosestellung im Verdachtskollektiv ist.

      Youden-Index = 0.940 + 0.933 – 1 = 0.873
    - S-100 als diagnostischer Test

      Hier sind nur zwei Cutpoints exemplarisch dargestellt. Um den optimalen zu finden, d.h. denjenigen, bei dem Sensitivität und Spezifität gleichzeitig am höchsten sind, müssen beide Maßzahlen für alle möglichen Ausprägungen von S100 ausgerechnet werden. Dieses Ergebnis kann dann graphisch dargestellt werden.

      Achtung: PPW und NPW hängen von der Prävalenz ab und gelten somit nur für die betrachtete Population und nicht allgemein für den diagnostischen Test.

6) Schätzverfahren

- Deskriptive versus induktive Statistik
  - Zielsetzung statistischer Auswertungen:
    - **Deskriptive Statistik:** Beschreibung des konkreten Datensatzes einer Studie, alle Aussagen beziehen sich auf das vorliegende empirische Datenmaterial
    - **Induktive Statistik:** allgemeinere Aussagen über den konkreten Datensatz hinausgehend und auf den zugrunde liegenden Bezugsrahmen der Studie (Fachbegriff: **Grundgesamtheit**) schließend

- Methoden der deskriptiven und induktiven Statistik unterscheiden sich aufgrund der unterschiedlichen Zielsetzung.
- Beispiel 3: Lübecker Blutdruckstudie
  - **Deskriptive Statistik:** Beschreibung des Datensatzes der Studie, z.B. Angabe der relativen Häufigkeit der Hypertonie im untersuchten Studienkollektiv
  - **Induktive Statistik:** Aussage über die Verbreitung der Hypertonie in der zugrunde liegenden Grundgesamtheit der Studie
  - **Grundgesamtheit hier:** 30-69jährige Lübecker Bevölkerung deutscher Nationalität
  - **Studienkollektiv hier:** Zufallsstichprobe aus dieser Grundgesamtheit
  - Dies ist der Idealfall! In vielen praktischen Fällen ist die tatsächlich zugrunde liegende Grundgesamtheit nicht so einfach anzugeben und das Studienkollektiv ist streng genommen keine richtige Zufallsstichprobe.

- Was sind Schätzer?
  - In der induktiven Statistik unterscheidet man zwischen den aus den Daten einer Studie errechneten Werten für interessierende Größen und den theoretisch zugrunde liegenden Größen selbst. Die theoretischen Größen nennt man oft **Parameter**, die aus den Daten berechneten Größen **Schätzer** (der Parameter).
  - **Beispiel:** Hypertonieverbreitung in der Lübecker Bevölkerung
  - **Parameter:** Wahrscheinlichkeit für Hypertonie in der Lübecker Bevölkerung (Grundgesamtheit)
  - **Schätzer** (dieses Parameters): relative Häufigkeit der Hypertonie aus den Daten des Studienkollektivs berechnet
  - Punktschätzung:
    - **Weiteres Beispiel:** durchschnittlicher diastolischer Blutdruckwert in der Lübecker Bevölkerung
    - **Parameter:** „Erwartungswert“ der diastolischen Blutdruckverteilung in der Lübecker Bevölkerung **Schätzer** (dieses Parameters): arithmetischer Mittelwert aller diastolischen Blutdruckwerte aus den Daten des Studienkollektivs berechnet
    - Die angegebenen Schätzer sind jeweils die so genannten **Punktschätzer** des interessierenden Parameters. Sie spiegeln die aus den Daten ermittelbare quantitative Information über die Größe des Parameters wider. Gleichzeitig erlaubt die Angabe eines Punktschätzers allein keine Einschätzung über die Präzision, mit der ein interessierender Parameter aus den Studiendaten ermittelt werden kann. Dieses Ziel wird mit **Intervallschätzern** verfolgt.
  - Punkt- und Intervallschätzung
    - Die Punktschätzung liefert einen Wert für den interessierenden Parameter und kann damit nicht gleichzeitig auch die Unsicherheit der Aussage vermitteln.
    - **Beispiel:** In der Lübecker Blutdruckstudie betrug der arithmetische Mittelwert aller 2359 diastolischen Blutdruckmesswerte 81,5 mmHg. Bei einer Studie an 20 Probanden in einem gleich alten Kollektiv betrug der arithmetische Mittelwert 86,5 mmHg.
    - Wie kann man die höhere Präzision, mit der eine große Studie wie die Lübecker Blutdruckstudie eine quantitative Aussage über den Erwartungswert des diastolischen Blutdrucks trifft, transparent machen bzw. die hohe Unsicherheit, die in einer auf einer kleinen Studie basierenden Punktschätzung dieses Erwartungswertes steckt, vor Augen führen?
  - Intervallschätzer
    - Intervallschätzer geben einen Bereich an, in dem der interessierende Parameter mit einer vorgegebenen „Sicherheit“ (statistisch ausgedrückt über eine Wahrscheinlichkeitsangabe) liegen wird, wenn man entsprechende Wiederholungen der Studiendurchführung realisiert.
    - Üblicherweise werden Intervallschätzer unter Hinzufügung der konkreten Wahrscheinlich-keitsangabe (nennt man auch **Konfidenzniveau**) als „(1–α)-Konfidenzintervall“ bezeichnet, wobei α Zahl zwischen 0 und 1 ist.

- Ein **95%-Konfidenzintervall** (entspricht α = 0.05) gibt also einen Bereich an, in dem der interessierende Parameter bei zwanzigfacher Wiederholung der Studie in 19 Fällen anzutreffen sein wird. Konfidenzintervalle sind in ihrer Breite vom Studienumfang und der Variabilität der interessierenden Messgröße abhängig.
- Je größer die Studie und je kleiner die Varianz der Messgröße, desto schmaler ist auch das Konfidenzintervall. Je höher das Konfidenzniveau, desto breiter das Konfidenzintervall bei gleichen Daten.

- Konfidenzintervalle – Anwendungen
  - Die Berechnung von Konfidenzintervallen bei vorgegebenem Konfidenzniveau benötigt meist eine zusätzliche Annahme über die Wahrscheinlichkeitsverteilung, aus welcher der interessierende Parameter stammt. Für die unterschiedlichen Anwendungen von Konfidenzintervallen gibt es jeweils Formeln und Software.

7) Regression

- Die Verfahren der Regressions- und Korrelationsanalyse dienen der **Beschreibung eines Zusammenhanges** zwischen zwei (oder mehreren) Merkmalen.
- Die **Regression** beschreibt dabei die **Art** eines (gerichteten) Zusammenhanges (Je-Desto-Beziehung), die **Korrelation** misst die **Stärke** eines (ungerichteten) Zusammenhanges.
- Bei der **Regressionsanalyse** unterscheidet man zwischen einer **abhängigen** und einer (oder mehreren) **unabhängigen** Variablen.
- Ziel der Analyse ist es festzustellen, wie sich Änderungen der unabhängigen Variablen auf die abhängige Variable auswirken.
- Die Regressionsanalyse beschreibt also die **Art** des Zusammenhanges und ermöglicht über die reine Beschreibung hinaus eine Voraussage (**Prädiktion**).
- Praktisches Vorgehen
  - Einteilung

    Die Einteilung der zu untersuchenden Variablen in abhängige und unabhängige Variable muss zuvor aufgrund inhaltlicher Überlegungen festgelegt werden.
  - Grafische Darstellung

    Wie bei der Korrelation erfolgt die grafische Darstellung des Zusammenhangs zweier quantitativer Merkmale mit Hilfe eines Scatterplots („Punktewolke“).
  - Formulierung des Modells
    - Die Punktwolke soll mit Hilfe einer Funktion beschrieben werden. Die grafische Darstellung lässt aufgrund der Form der Punktwolke einen **linearen Zusammenhanges** vermuten; die Punkte streuen bandförmig um eine Gerade.
    - Die gesuchte Funktion ist somit eine **Gerade**, die mit Hilfe der linearen Regressionsanalyse bestimmt werden kann.
    - Lineares Modell: $Y = \alpha + \beta \cdot X + \varepsilon$ wobei Y die abhängige Variable, α das Absolutglied („Intercept“), β den Regressionskoeffizienten und ε einen zufälligen Fehlerterm darstellt.
  - Schätzung der Regressionsfunktion

    Zu den gegebenen Wertepaaren (xi, yi) werden die entsprechenden Koeffizienten (α, β) geschätzt. Die Regressionsgerade soll die zugrundeliegenden Daten so gut wie möglich repräsentieren.
  - Methode der kleinsten Quadrate

    Die Summe der Quadrate der „Abstände“ aller Punkte von der Geraden, d.h. die Summe der quadrierten Differenzen zwischen dem beobachteten Wert und dem vorhergesagten Wert – auch **Residuum** genannt – soll minimal werden.

    $$\sum (\hat{y}_i - y_i)^2 \stackrel{!}{=} \min$$

- Beurteilung der Regression
  - Die **Güte der Regressionsgeraden** als ganze soll geprüft werden. Hierfür wird die Streuung (= Varianz) der Regression zerlegt:
    $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$
  - Gesamtstreuung = erklärte Streuung + Residualstreuung
  - Als Maßzahl zur Beurteilung der Güte der Regressionsschätzung dient das **Bestimmtheitsmaß r2**. Es stellt das **Verhältnis von erklärter Streuung zur Gesamtstreuung** dar ($0 \leq r2 \leq 1$).
- Überprüfung der Voraussetzungen

  Folgende Aspekte sind zu überprüfen:
  - Linearitätsannahme (funktionale Form des Modells)
    Scatterplott: Residuen gegen unabhängige Variable (siehe links)
  - Multikollinearität (nur bei Regressionsmodellen mit mehreren unabhängigen Variablen: keine (zu starke) Abhängigkeit der unabhängigen Variablen untereinander)
  - Spezifikation der „richtigen" unabhängigen Variablen
  - Verteilungsannahme für abhängige Variable
    Abhängige Variable sind Lungenkrebstote (siehe oben Mitte)
  - Homoskedastizitätsannahme für die Residuen (bedeutet: die Residuen des Modells sollen über den gesamten Wertebereich gleich stark um Null schwanken)
    Scatterplott: Residuen gegen vorhergesagte Werte (siehe rechts oben)

- Beispiel 7: Zigarettenkonsum

  Untersuchung über den Zusammenhang zwischen dem Zigarettenkonsum und der Anzahl an Todesfällen durch verschiedene Krebsarten in 43 US-Staaten. Daten: cig = Anzahl verkaufter Zigaretten (in 100) pro Kopf lung = Tode durch Lungenkrebs (pro 100.000 Einwohner)
  Wie sieht der Einfluss des Zigarettenkonsums auf die Anzahl der Todesfälle durch Lungenkrebs aus? Eine umgekehrte Fragestellung ist mit den vorhandenen Daten nicht sinnvoll.
  - $Y = \alpha + \beta \cdot X + \varepsilon = 6.47 + 0.53 \cdot X + \varepsilon$
  - **Interpretation:** Ohne Zigarettenkonsum bliebe eine Lungenkrebsmortalität von 6.47 Tote pro 100.000 Einwohner/Jahr. Diese nimmt mit 0.53 pro 100 verkauften Zigaretten zu.
- Zusammenhang Regression-Korrelation
  - Die Methoden der Regressions- und Korrelationsanalyse hängen formal eng zusammen.
  - Dies wird auch daran deutlich, dass das im Rahmen der Regressionsanalyse eingefügte lineare Bestimmtheitsmaß r^2 bei einer einfachen linearen Regression dem Quadrat des in der Korrelationsanalyse eingeführten (Pearsonschen) Korrelationskoeffizienten für den Zusammenhang von unabhängiger und abhängiger Variable entspricht.
  - Unterschiede zwischen Regressions- und Korrelationsanalyse ergeben sich primär aus der unterschiedlichen, inhaltlich bestimmten Ausgangsfragestellung, die bei „gerichteten Zusammenhängen" zur Anwendung von Methoden der Regressionsanalyse, bei „ungerichteten Zusammenhängen" zur Anwendung der Korrelationsanalyse führt.

8) Überlebenszeitanalyse

- Ausgangssituation
  - Was unterscheidet die Überlebenszeitanalyse von bisherigen Fragestellungen?
  - Zielgröße ist hier stets eine Zeit**dauer**, also nicht ein (Mess-)Wert, den man zu einem fixen Zeit**punkt** erheben kann. Daher ergeben sich folgende Probleme:
    - Solange das interessierende Zielereignis (Tod, Rezidiv, Krankheitsschub etc.) nicht eingetreten ist, kann die exakte Zeitdauer nicht angegeben werden.
    - Da die Zeitdauern für die Studienpatienten individuell zu erheben sind und innerhalb einer Studie die Rekrutierung bzw. Behandlung zu unterschiedlichen Zeitpunkten erfolgt („staggered entry"), benötigt jeder Studienpatient seine eigene „Zeitachse"

- Zensierung
  - Die Besonderheit der Analyse von Überlebenszeiten liegt im Auftreten von **Zensierungen**, bzw. von **zensierten Überlebenszeiten**. Davon spricht man, wenn die Überlebenszeit nicht exakt bestimmbar ist, d.h., wenn bei Beobachtungsende das interessierende Zielereignis noch nicht eingetreten ist.
  - Es können drei Situationen eintreten:
    - Zielereignis tritt ein
    - Zielereignis tritt bis zum Studienende nicht ein
    - Teilnehmer scheidet vorzeitig aus der Studie aus
  - Eine exakte Überlebenszeit kann nur im 1. Fall beobachtet werden. In den Fällen 2. und 3. Können nur Mindestüberlebenszeiten angegeben werden.
  - Zensierungsmodell
    - Bezeichnet wird mit Ti die echte Überlebenszeit eines Probanden und mit Ci seine sog. Zensierungszeit (Zeit unter Beobachtung).
    - Beobachten kann man in einer Studie stets nur das Minimum dieser beiden Größen (Xi). Zusätzlich weiß man, ob dieses Minimum eine exakte ($Ti \leq Ci$) oder eine zensierte Zeit ($Ti > Ci$) ist. Diese Tatsache wird in einem Zensierungsindikator ($\delta i=1$ für exakte, $\delta i=0$ für zensierte Daten) notiert.
    - Somit sind die Studiendaten, die zur Überlebenszeitanalyse herangezogen werden, durch die beobachtete Zeit und den Zensierungsindikator gegeben: (X1, δ1) (X2, δ2) (X3, δ3) (X4, δ4) ... (Xn, δn)
- Überlebenszeitfunktionen
  - Wie lässt sich das im Beispiel ersichtliche Problem der Berücksichtigung unterschiedlicher Überlebenszeitinformationen („exakt“ versus „zensiert“) lösen?
  - Man kann nur bestimmte Informationen aus den Daten ableiten, z.B. macht es keinen Sinn Mittelwerte, Standardabweichungen etc. auszurechnen. Sinnvoll analysierbar ist allerdings die Überlebenszeitfunktion S.
  - Die Überlebenszeitfunktion $S(t):=P(T>t)$ gibt die Wahrscheinlichkeit an, den Zeitpunkt t zu überleben. Sie hängt direkt mit der Verteilungsfunktion F der Überlebenszeitdaten zusammen. $S(t)=1-F(t)$. Diese Überlebenszeitfunktion S kann auch aus einer Mischung von exakten und zensierten Überlebenszeiten ermittelt werden.
  - Schätzung der Überlebenszeitfunktion
    - Falls keine zensierten Überlebenszeiten beobachtet werden, lässt sich die Überlebenszeitfunktion S analog zur empirischen Verteilungsfunktion ermitteln:

      $$S(t) = \frac{\text{Anzahl der Teilnehmer mit Überlebenszeit} \geq t}{\text{gesamte Fallzahl}}$$

    - Bei Vorliegen zensierter Überlebenszeiten zerlegt man die Zeitachse gemäß den unzensierten Überlebenszeitpunkten und ermittelt iterativ die Funktion S abschnittsweise durch die **Kaplan-Meier- Methode**.
      - Notation
        - (t0,) t1, t2, t3,...: angeordnete unzensierte Zeiten, nj: Anzahl der lebenden Patienten unmittelbar vor dem Zeitpunkt tj, dj: Anzahl der Zielereignisse zum Zeitpunkt tj
        - Mittels der Kaplan-Meier-Methode wird nun iterativ die Überlebenszeitfunktion S durch ein Produkt bedingter Wahrscheinlichkeiten bestimmt:
          Startpunkt t0: $S(t0) = P(T>t0) =1$
        - Kleinste (exakte) Überlebenszeit t1:
          $S(t1)=P(T> t1)=S(t0)\cdot P(T>t1|T>t0)$ ... $S(tj)=P(T> tj)=S(tj-1)\cdot P(T>tj|T>tj-1)$
      - Durch die Vorgabe $S(t0)=1$ lässt sich für jeden Zeitpunkt tj, an dem eine exakte Überlebenszeit beobachtet wurde, der zugehörige Funktionswert der Überlebens-zeitfunktion S nach schrittweisem Auflösen der Rekursionsformel angeben.

- Zwischen zwei aufeinander folgenden Zeitpunkten tj-1 und tj bleibt die Funktion S konstant und springt nur an den tj („Treppenfunktion“).
- Zu klären ist somit nur noch, wie Term P(T>tj|T>tj-1) aus den Daten berechnet wird.
- Die bedingte Wahrscheinlichkeit, den Zeitpunkt tj zu überleben, unter der Bedingung unmittelbar vor diesem Zeitpunkt noch gelebt zu haben, lässt sich aus den Daten ermitteln als (nj – dj)/nj = 1- dj/nj.
- Die mittels Kaplan-Meier-Methode ermittelten Überlebenszeitfunktionen werden häufig grafisch in Form von so genannten Kaplan-Meier-Kurven dargestellt.

- Beispiel 8:
  - Es gibt eine „neue“ Therapie (Bestrahlung und Chemotherapie) des Mundbodenkarzinoms. Es soll ein Unterschied zur „alten“ Therapie (nur Bestrahlung) festgestellt werden, das Zielereignis ist der Tod in Folge des Karzinoms.
  - Die Studie dauert insgesamt drei Jahre, die Rekrutierungsphase geht über die ersten zwei Jahre. Es wurden 15 Patienten pro Therapiegruppe eingeschlossen, und die Beobachtungszeiten in der Studie wurden in Monaten erfasst.
  - Das im Beispiel ersichtliche Problem der unterschiedlichen Eintrittszeitpunkte in die Studie („staggered entry“) ist einfach zu lösen:
    Aus der kalendarischen Zeitachse muss eine Zeitachse gemacht werden, welche die individuellen Beobachtungszeiten nach Aufnahme in die Studie beinhaltet. Somit werden die individuellen Zeitachsen synchronisiert und man startet mit einem gemeinsamen Nullpunkt der Beobachtungszeit.
  - Berechnung der S(tj) in der Therapiegruppe der kombinierten Bestrahlungs- und Chemotherapie (Grafik rechts)
  - Im Beispiel sieht dies wie folgt aus: (siehe rechts)

9) Statistisches Testen

Was ist ein statistischer Test?

- ein **Entscheidungsverfahren**, das über eine a priori formulierte „statistische Vermutung“ durch das Aufstellen von **Nullhypothese** (H0) und zugehöriger **Alternative** (H1) formalisiert. nach der Sammlung von empirischer Information (Daten) eine konkrete Entscheidung trifft.
- „Entscheidung“ meint hier, dass die Vermutung entweder abgelehnt wird oder nicht.
- Nullhypothese enthält dabei stets die zu widerlegende Aussage, die Alternative beinhaltet die zu beweisende Aussage.
  H0: ⟨was zu widerlegen ist⟩ **H1: ⟨was zu beweisen ist⟩**
- Die getroffene Entscheidung kann richtig oder falsch sein.
- Ein statistischer Test erlaubt es, die Häufigkeit einer Fehlentscheidung im Sinne der ungerechtfertigten Ablehnung der Vermutung zu „kontrollieren“.

Allgemeines Vorgehen

- Daten werden im Rahmen einer Studie, eines Experiments o.ä. erhoben.
- Anhand des Vorwissens wird eine „Vermutung“ formuliert. Aufgrund der Daten wird nach Anwendung einer statistischen Regel eine Entscheidung über die Vermutung getroffen.
- Diese Entscheidung kann falsch sein.
- Die Wahrscheinlichkeit für eine Fehlentscheidung wird durch Vorgabe eines „Signifikanzniveaus“ nach oben begrenzt.

Ein- und zweiseitiges statistisches Testen

- Es gibt verschiedene Formulierungen der Alternative H1 beim statistischen Testen.
  - Ist eine Abweichung von der Nullhypothese (H0) nur in eine Richtung („größer“/ „kleiner“) von Interesse, so spricht man von einem **einseitigen statistischen Test**.
  - Abweichung von der Nullhypothese (H0) in beide Richtungen („unterschiedlich“ vom in H0 spezifizierten Zustand), so spricht man von einem **zweiseitigen statistischen Test**.
- Das **Standardvorgehen** des statistischen Testens in der Medizin ist der **zweiseitige Test**. Lässt sich jedoch **a priori** stringent begründen, dass eine **Abweichung von H0 nur in eine Richtung** möglich ist, so ist in diesen seltenen Ausnahmefällen ein einseitiger Test indiziert.

- Fehlentscheidungen beim statistischen Testen
  - Es gibt zwei unterschiedliche Typen von Fehlentscheidungen:
    - **Fehler 1. Art:** Nullhypothese wird abgelehnt, obwohl sie zutrifft
    - **Fehler 2. Art:** Nullhypothese wird nicht abgelehnt, obwohl sie nicht zutrifft
  - Die Wahrscheinlichkeit für einen Fehler 1. Art wird i.d.R. mit α abgekürzt, die Wahrscheinlichkeit für einen Fehler 2. Art mit β.
    - Beim statistischen Testen wird α („Signifikanzniveau“) vorgegeben und durch die Konstruktion des Entscheidungsverfahrens „kontrolliert“.
    - Dies gilt nicht für β.
  - Illustration der Fehlertypen

		Testentscheidung	
		H_0 nicht ablehnen	H_0 ablehnen
Unbekannte Realität	H_0 richtig	Richtige Entscheidung	**Fehler 1. Art** α
	H_0 falsch	**Fehler 2. Art** β	Richtige Entscheidung

  - Illustration der Fehlertypen im Beispielkontext

		Testentscheidung	
		H_0 nicht ablehnen	H_0 ablehnen
Unbekannte Realität	kein Unterschied zwischen beiden Therapien	kein Unterschied (richtig)	**signifikanter Unterschied (falsch)**
	eine Therapie ist besser als die andere	**kein Unterschied (falsch)**	signifikanter Unterschied (richtig)

- Wie trifft man auf Basis der Daten eine statistische Entscheidung?
  - **Allgemein:** Man berechnet aus den Daten eine „Prüfgröße“ (Teststatistik), die einem Auskunft über die Vermutung gibt.
  - Die Form der Teststatistik wird so gestaltet, dass man ihre Wahrscheinlichkeits-verteilung unter der Nullhypothese angeben kann. Dann weiß man, ob ein aus den Daten berechneter Wert für Teststatistik unter Gültigkeit der Vermutung „wahrscheinlich“ oder „unwahrscheinlich“ ist.
    - Ist er so unwahrscheinlich, dass er kleiner als das zuvor festgesetzte Signifikanzniveau α ist, so entscheidet man sich, die Nullhypothese abzulehnen.
    - Ist die Wahrscheinlichkeit allerdings größer als α, so lehnt man die Nullhypothese nicht ab.
- Wahl des **Signifikanzniveaus**
  - **Standard:** α = 0.05 (5%) Jede Abweichung muss gut begründet sein.
  - **Alternativen:**
    - α = 0.01 (1%), falls ein Fehler 1. Art gravierende Konsequenzen hat, ein Fehler 2. Art zu tolerieren ist
    - α = 0.10 (10%), falls ein Fehler 2. Art gravierende Konsequenzen hat, ein Fehler 1. Art zu tolerieren ist
- Konsequenzen der Entscheidung zwischen ein- und zweiseitigen statistischen Tests
  - **Einseitiges Testen:** „Kontrolle“ des Fehlers 1. Art (α) nur in eine Richtung notwendig
  - **Zweiseitiges Testen:** „Kontrolle“ des Fehlers 1. Art (α) in beide Richtungen notwendig, so dass für jede Richtung nur α/2 zur Verfügung stehen
  - Somit können bei denselben Daten unterschiedliche Testentscheidungen entstehen.
  - Bei einseitigem Testen kommt man eher zu einer Ablehnung der Nullhypothese als bei zweiseitigem Testen. Dies birgt ein manipulatives Potential in sich, so dass man strenge Anforderungen an die Begründung für ein Abweichen vom Standardvorgehen des zweiseitigen Testens stellt.
- Was sind **p-Werte**?
  - Ein p-Wert ist die Wahrscheinlichkeit dafür, dass die Prüfgröße unter H0 Werte annimmt, die größer oder gleich dem aus den Daten berechneten Wert der Prüfgröße sind.

- Statistiker nennen p-Werte daher „Überschreitungswahrscheinlichkeiten“.
- Kennt man p-Wert, so kann man auf die Ermittlung kritischer Werte aus Tabellen verzichten.
- **Begründung:** Vorgegeben sei ein Signifikanzniveau (Größenordnung des tolerierten Fehlers 1. Art). Die beiden Entscheidungsregeln
  - H0 ablehnen, falls p-Wert ≤ α
  - H0 ablehnen, falls Prüfgröße ≥ cα

  sind äquivalent. Popularität von p-Werten resultiert aus Verwendung von Statistiksoftware.
- Die Software gibt stets p-Werte als Resultate von statistischen Tests aus.

- Interpretation von Testentscheidungen
  - **Allgemeine Beschreibung:**

1. **Fall:** Prüfgröße ≥ cα oder p-Wert ≤ α
   - **Entscheidung:** H0 ablehnen
     - P(Fehler 1. Art) = α
     - P(Fehler 2. Art) = 0
   - **Interpretation:** Die in den Daten zu beobachtende Abweichung von der Nullhypothese ist auf dem (100•α)%-Niveau signifikant.
2. **Fall:** Prüfgröße < cα oder p-Wert > α
   - **Entscheidung:** H0 nicht ablehnen
     - P(Fehler 1. Art) = 0
     - P(Fehler 2. Art) = unbekannt
   - **Interpretation:** Die in den Daten zu beobachtende Abweichung von der Nullhypothese ist auf dem (100•α)%-Niveau ist nicht signifikant.
   - **Vorsicht:** H0 nicht abzulehnen, bedeutet **nicht** die Gültigkeit von H0 (mit irgendeiner Wahrscheinlichkeitsquantifizierung, z.B. 1 - α) statistisch nachgewiesen zu haben.

- **Statistische Power**
  - Definition: Als „statistische Power“ eines Tests bezeichnet man die Wahrscheinlichkeit, mit der ein statistischer Test eine spezifische „richtige“ Alternative unter den Rahmenbedingungen seines Einsatzes (Fallzahl, Signifikanzniveau) auch als solche entdeckt (d.h. die „falsche“ Nullhypothese ablehnt).
  - Formal gilt: **Statistische Power = 1 - Fehler 2. Art = 1 – β**
  - Die statistische Power eines Tests hängt von folgenden Größen ab:
    - **Fallzahl der Studie:** je größer die Fallzahl, desto größer die statistische Power
    - **Signifikanzniveau des statistischen Tests:** je kleiner das Signifikanzniveau, desto kleiner die statistische Power
    - **der konkreten Alternative, die vorliegt:** je „weiter weg“ von der Nullhypothese die spezifische Alternative, desto größer die statistische Power
  - und Fallzahlplanung
    - Zusammenhang zwischen statistischer Power (Pow), Fallzahl (n), Signifikanzniveau (α) und spezifischer Alternative (HP1) ist evident.
    - Bislang haben wir Pow als Funktion von n, α und HP1 aufgefasst.
    - Wir können jedoch diese Beziehung auch nach n „auflösen“ und somit die Fallzahl n als Funktion von Pow, α und HP1 beschreiben. Dies ist dann die Basis für jede Fallzahlplanung.
    - Das Einsetzen konkreter Werte für die statistische Power Pow, das Signifikanzniveau α und die spezifische Alternative H1P in diese Funktion liefert dann die Fallzahl für die vorgegebene Situation.

- Konkrete statistische Tests
  - Generell: Es existieren sehr viele verschiedene statistische Tests, von denen nur wenige Standardtests im Rahmen der Vorlesung konkret vorgestellt werden.
  - **Kriterien** für die Unterscheidung zwischen statistischen Tests:
    - **Skalentyp** der auszuwertenden Daten (teilweise zusätzlich: Verteilungsannahmen)
    - **Strukturelle Designmerkmale** der Studie (z.B. Parallel-Gruppen-Design versus Cross-Over-Design in klinischen Prüfungen)

- **unverbunden/verbunden** (unabhängig, unpaarig, z.B. Fall-Kontroll-Studie/abhängig, paarig, Individuen haben in beiden Gruppen etwas miteinander zu tun)

Anzahl und Art der Stichproben	Quantitativ		Qualitativ bzw. dichotom	Überlebens-zeiten
	normalverteilt	Verteilung unbekannt		
Eine Stichprobe	Ein-Stichproben t-Test	(nicht behandelt)	Binomial-test	
Zwei verbundene Stichproben	t-Test für verbundene St.	Wilcoxon-Vorzeichen-Test	McNemar-Test	
Zwei unverbundene Stichproben	t-Test für unverbundene St.	U-Test von Mann, Whitney und Wilcoxon	χ^2-Test, Exakter Test von Fisher	Logrank-Test

- im Einstichproben-Design
  - „Einstichproben-Design“ meint, dass **nur eine Gruppe von unabhängigen Merkmalsträgern** hinsichtlich der **Verteilung der beobachteten Merkmalsausprägungen** analysiert wird.
  - Statistische Tests in dieser Situation:
    - **Einstichproben-t-Test**
      - **Voraussetzung:** X1,..., Xn unabhängige Messgrößen, deren Verteilung aus einer Normalverteilung mit Erwartungswert μ und Varianz σ2 stammt.
      - **Testproblem:** H0: μ = μ0 versus H1: μ ≠ μ0
      - **Teststatistik:** $T = \frac{|\bar{X} - \mu_0|}{s_X} \cdot \sqrt{n}$
        Die Teststatistik T folgt unter H0 einer t-Verteilung mit n-1 Freiheits-graden. Aus dieser Tatsache kann man kritische Werte bei vorgege-benem Signifikanzniveau α zur Testdurchführung ermitteln - bzw. direkt den p-Wert ausrechnen – und dann die Testentscheidung treffen.
    - **Binomialtest**
      - **Voraussetzung:** X1,..., Xn unabhängige dichotome Messgrößen („Erfolg“ / „Misserfolg“), deren Verteilung somit aus einer Binomialverteilung mit Parametern n und p (Wahrscheinlichkeit für Erfolg) stammt. Die Anzahl der „Erfolge“ wird mit K bezeichnet.
      - **Testproblem:** H0: p = p0 versus H1: p ≠ p0
      - **Teststatistik:** $B = \frac{|K - n \cdot p_0|}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$
        Die Teststatistik B folgt unter H0 (bei nicht zu kleinem Stichproben-umfang, Daumenregel: $n \cdot p0 \cdot (1-p0) > 9$) einer Standardnormalverteilung (Erwartungswert: 0, Varianz:1).
      - Beispiel:
        Erinnerung: Im Teil Wahrscheinlichkeitsrechnung hatten wir ein Beispiel zum Erfolg einer neuen Therapie des metastasierten malignen Melanoms, der angeblich 20% betragen sollte. In einer Studie an 100 Melanompatienten wurden 14 Therapieerfolge beobachtet. Ist damit die Aussage der ‚Erfinder’ des neuen Therapieverfahrens widerlegt?
        Testproblem: H0: p = 0.2 versus H1: p ≠ 0.2
        Teststatistik: $B = \frac{|14 - 100 \cdot 0.2|}{\sqrt{100 \cdot 0.2 \cdot 0.8}}$
        Die Teststatistik B ergibt den Wert 1.5. Dies liegt unterhalb des kritischen Wertes von 1.96 (der p-Wert ist 0.13). Somit ist H0 zu α = 5% nicht abzulehnen.
- **Zweistichproben-Tests**
  - Beim Vergleich zweier Stichproben ist zunächst zu klären, ob durch das Design der Studie Abhängigkeiten der beiden Stichproben zu berücksichtigen sind. Es ist zwischen folgenden Situationen zu unterscheiden:

- **Zwei unverbundene Stichproben** („Parallel-Gruppen-Design“): zwei Gruppen von unabhängigen Merkmalsträgern sollen hinsichtlich möglicher (Gruppen-) Unterschiede in der Verteilung der Merkmalsausprägungen analysiert werden.
- **Zwei verbundene Stichproben** (z.B. „Cross-over-Design“, „Vorher-Nachher-Vergleich“): in einer Gruppe von Merkmalsträgern soll die Verteilung von zwei an den Merkmalsträgern beobachteten Merkmalsausprägungen analysiert werden.

- T-Test für eine und zwei Stichproben
  - **Achtung:** Der zu Beginn eingeführte Einstichproben-t- Test und der t-Test für zwei verbundene Stichproben sind formal betrachtet derselbe Test. Der einzige formale Unterschied besteht darin, dass in der Situation **zweier verbundener Stichproben** in der Formulierung des Testproblems stets die **Nullhypothese μ = 0 überprüft** wird, während im **Einstichprobenfall beliebige Nullhypothesen** möglich sind.
  - Im Fall zweier verbundener Stichproben wird durch die Differenzbildung der beiden Beobachtungen am selben Merkmalsträger für die Auswertung **eine** Stichprobe von Werten gebildet. Dadurch wird dieser Zweistichprobenfall in den Einstichprobenfall überführt.
- **Kontingenztafeltests (Unabhängigkeitstests)**
  - **Voraussetzung:** An n unabhängigen Merkmalsträgern werden simultan zwei dichotome Merkmale A (Ausprägungen: u und v) und B (Ausprägungen: s und t) erfasst. Die gemeinsame Verteilung wird in einer Kontingenztafel dargestellt.
  - **Kontingenztafel:**

A \ B	s	t	Gesamt
U	a	b	a + b
V	c	d	c + d
Gesamt	a + c	b + d	n

  - **Typische Frage:** Treten A und B unabhängig auf?
  - **Testproblem:** H0: P(A ∩ B) = P(A) * P(B) versus H1: P(A ∩ B) ≠ P(A) * P(B)
  - **Teststatistik:** $\chi^2 = \frac{(a \cdot d - b \cdot c)^2 \cdot n}{(a+c) \cdot (b+d) \cdot (a+b) \cdot (c+d)}$

    Die Teststatistik χ2 dieses Unabhängigkeitstests ist identisch zur o.a. Teststatistik beim Vergleich der Erfolgswahrscheinlichkeiten zweier unverbundener Stichproben. Somit kann formal derselbe Test in zwei unterschiedlichen Situationen für verschiedene Testprobleme eingesetzt werden.
- Tests im **Parallel-Gruppen-Design**
  - „Parallel-Gruppen-Design“ meint, dass zwei (oder mehrere) Gruppen von **unabhängigen** Merkmalsträgern hinsichtlich **möglicher (Gruppen-)Unterschiede in der Verteilung der beobachteten Merkmalsausprägungen** analysiert werden.
  - Statistiker sprechen oft von dieser Situation als „Vergleich zweier (oder mehrerer) unverbundener Stichproben“.
  - **Statistische Tests in dieser Situation:**
    - **t-Test für zwei unverbundene Stichproben** (bekanntester)
      - **Voraussetzung:** X1,..., XnX unabhängige Messgrößen, deren Verteilung aus einer Normalverteilung mit Erwartungswert μX und Varianz σ2 stammt. Analog Y1,..., YnY unabhängige Messgrößen aus einer Normalverteilung mit Erwartungswert μY und Varianz σ2.
      - **Testproblem:** H0: μX = μY versus H1: μX ≠ μY
      - **Teststatistik:** $T = \frac{|\bar{X} - \bar{Y}|}{s_{XY}} \cdot \sqrt{\frac{n_X \cdot n_Y}{n_X + n_Y}}$

        Die Teststatistik T folgt unter H0 einer t-Verteilung mit $n_X + n_Y - 2$Freiheitsgraden. Die empirische Standardabweichung Sxy bezieht sich auf die „gepoolten“

Messreihen und kann auch als entsprechend gewichtetes Mittel aus Sx und Sy gewonnen werden.

➢ **U-Test von Mann-Whitney-Wilcoxon**

- **Voraussetzung:** X1,..., XnX sowie Y1,..., YnY sind unabhängige Messgrößen mit beliebiger Verteilung FX bzw. FY. In der Gesamtstichprobe werden Rangzahlen R1 bis RnX+nY für die der Größe nach geordneten
- Messgrößen vergeben und innerhalb der beiden Messreihen gemittelt (XR + YR)
- **Testproblem:** H0: FX = FY versus H1: FX ≠ FY
- **Teststatistik:** $$U = \frac{\left|\bar{R}_X - \bar{R}_Y\right|}{\sqrt{\frac{n_X + n_Y}{n_X \cdot n_Y \cdot (n_X + n_Y - 1)} \cdot \left(\sum R_i^2 - \frac{(n_X + n_Y + 1)^2 \cdot (n_X + n_Y)}{4} \right)}}$$
  U folgt unter H0 (bei nicht zu kleinem Probenumfang, sonst Vertafelung kritischer Werte verwenden) einer Standardnormalverteilung

➢ **Logrank-Test**

- **Standardtest zum Vgl von Überlebenskurven in 2 unverbundenen Stichproben**.
- nichtparametrisch (keine Voraussetzungen an Verteilung der Überlebenszeiten)
- **Nullhypothese H0** : Überlebenszeitfunktionen der zu vergleichenden Gruppen stimmen überein, d.h. es gilt für alle Zeitpunkte t:
  $$H_0 : S_{\text{Gruppe\_I}}(t) = S_{\text{Gruppe\_II}}(t)$$
  → in beiden Gruppen zu jedem Zeitpunkt tj gleiche Sterbewahrscheinlichkeit
- in jeder Gruppe (i=1,2) tatsächlich beobachtete Anzahl von Zielereignissen (Oi) mit unter H0 erwarteter Anzahl von Zielereignissen (Ei) verglichen. Ei durch Summation über alle Ereigniszeitpunkte tj unter der Annahme gleicher Ereigniswahrscheinlichkeiten in den zu vergleichenden Gruppen.
- **Teststatistik:** $T = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$ T folgt unter H0 (bei nicht zu kleinen Gruppengrößen) einer χ2-Verteilung mit 1 Freiheitsgrad

➢ **Chi-Quadrat-Test = Kontingenztafeltest unverbundener Stichproben**

- **Voraussetzung:** X1,..., Xnx sowie Y1,..., Yny sind unabhängige Beobachtungen einer dichotomen Zielgröße, deren Verteilungen in einer „Kontingenztafel“ dargestellt wird. (stets bei Tafeln anwenden!)
- **Kontingenztafel:**

Gruppe	„Erfolg“	„Misserfolg“	Gesamt
X	a	b	n_X
Y	c	d	n_Y
Gesamt	a + c	b + d	$n_X + n_Y$

  Die Wahrscheinlichkeiten für einen Erfolg in Gruppe X bzw. Y werden mit Px bzw. Py bezeichnet und durch die relativen Häufigkeiten a/nx bzw. c/ny geschätzt.
- **Testproblem:** H0: Px = Py versus H1: Px ≠ Py
- **Teststatistik:** $$\chi^2 = \frac{(a \cdot d - b \cdot c)^2}{(a + c) \cdot (b + d) \cdot n_X \cdot n_Y} \cdot (n_X + n_Y)$$
  Die Teststatistik χ2 folgt unter H0 (bei nicht zu kleinen Besetzungszahlen in der Kontingenztafel) einer Chi-Quadrat-Verteilung mit einem Freiheitsgrad.

➢ **Exakter Test von Fisher**

Falls die Besetzungszahlen in der Tafel „zu niedrig“ (Daumenregel: ≤ 5) für die Anwendung des Chi-Quadrat-Tests sind, findet an seiner Stelle der **exakte Test von Fisher** Anwendung (einer der wenigen statistischen Tests, bei dem **keine Teststatistik** ausgerechnet wird, sondern **direkt ein p-Wert bestimmt** wird).

- Tests für **verbundene** Stichproben

- „Verbundene Stichproben“ meint, dass **in einer Gruppe von Merkmalsträgern** die **Verteilung von zwei (oder mehreren) an den Merkmalsträgern beobachteten** Merkmalsausprägungen analysiert wird (Synonym: „abhängige Stichproben“).
- Typische Beispiele für dieses Design sind z.B. „Vorher-Nachher-Vergleiche“ oder klinische Prüfungen mit Therapiewechsel im „Cross-over- Design“.

- **Statistische Tests in dieser Situation:**

➢ **(„paired“)t-Test für zwei verbundene Stichproben = Einstichpr.-t-Test für Differenz**

❖ **Voraussetzung:** (X1 , Y1),..., (Xn , Yn) unabhängige Wertepaare von normalverteilten Messgrößen. Bezeichne Di die Differenz Xi - Yi , die dann ebenfalls aus einer Normalverteilung mit Erwartungswert µD undVarianz σ2 stammt.

❖ **Testproblem:** H0: µD = 0 versus H1: µD ≠ 0

❖ **Teststatistik:** $T = \frac{|\bar{D}|}{s_D} \cdot \sqrt{n}$

Die Teststatistik T folgt unter H0 einer t-Verteilung mit n -1 Freiheitsgraden. Das ganze Vorgehen entspricht dem Einstichproben-t-Test für die Differenzen.

➢ **Wilcoxon-Vorzeichen-Rangsummen-Test**

❖ **Voraussetzung:** (X1 , Y1),..., (Xn , Yn) unabhängige Wertepaare von Messgrößen mit Verteilung FX und FY, deren Differenz Xi - Yi (als Di abgekürzt) eine beliebige symmetrische Verteilung haben. Für die Absolutbeträge der Di ≠ 0 werden Rangzahlen von R1 bis Rn* der Größe nach vergeben. Die Rangzahlen der positiven Differenzen (Di > 0) werden zu R+ aufaddiert.

❖ **Testproblem:** H0: : FX = FY versus H1: FX ≠ FY

❖ **Teststatistik:** $W = \frac{\left|2 \cdot R^{+} - \frac{n^{*} \cdot (n^{*}+1)}{2}\right|}{\sum_{i=1}^{n^{*}} R_i^2}$

Die Teststatistik W folgt unter H0 (bei nicht zu kleinem Stichprobenumfang, sonst Vertafelung kritischer Werte anwenden) einer Standardnormalverteilung.

➢ **McNemar-Test = Kontingenztafeltest verbundener Stichproben**

❖ **Voraussetzung:** (X1 , Y1),..., (Xn , Yn) unabhängige Wertepaare einer dichotomen Zielgröße, deren Verteilung in einer Kontingenztafel dargestellt wird.

❖ **Kontingenztafel:**

X \ Y	„Erfolg“	„Misserfolg“	Gesamt
„Erfolg“	a	b	a + b
„Misserfolg“	c	d	c + d
Gesamt	a + c	b + d	n

Die Wahrscheinlichkeiten für einen Erfolg unter X bzw. Y werden mit pX bzw. pY bezeichnet und diesmal durch die relativen Häufigkeiten (a + b)/n bzw. (a + c)/n geschätzt.

❖ **Testproblem:** H0: pX = pY versus H1: pX ≠ pY

❖ **Teststatistik:** $\chi^2 = \frac{(b-c)^2}{(b+c)}$ χ2 folgt unter H0 (bei nicht zu kleinen Besetzungszahlen in der Kontingenztafel) einer Chi-Quadrat-Verteilung mit einem Freiheitsgrad. Falls die Besetzungszahlen in der Tafel „zu niedrig“ (Daumenregel: b + c ≤ 20) für die Anwendung dieses McNemar-Tests sind, so testet man mittels **Binomialtest**, ob Wahrscheinlichkeit einer der beiden Nebendiagonalzellen 0.5 entspricht.

- Zusammenfassung:

Stichprobentyp	*Grundgesamtheit normalverteilt*	*Grundgesamtheit mit unbekannter Verteilung*
Eine Stichprobe	1-Stichproben-t-Test	1-Stichproben-Wilcoxon-Test
2 Stichproben	t-Tests für verbundene Stichproben	Wilcoxon-Test für Paardifferenzen
	t-Test für unverbundene Stichproben	U-Test von Mann/Whithyney/ Wilcoxon
Mehr als 2 Stichproben	Varianzanalyse	Kurskal-Wallis-Test

- Beispiele

1. Randomisierte kontrollierte Therapiestudie zur Behandlung des schweren Atemnotsyndroms von Frühgeborenen mit zwei Dosierungsschemata eines natürlichen Surfactantpräparats

Ergebnisse:

Therapiearm-Umfang (Mortalität): einfach 176 37 (= 21%), mehrfach 167 21 (= 13%)

Frage: Mit welchem Dosierungsschema soll in der Zukunft ein schweres Atemnotsyndrom bei Frühgeborenen behandelt werden?

Nullhypothese (H0): „Die Mortalitätswahrscheinlichkeiten der beiden Therapiearme unterscheiden sich nicht."

Alternative (H1): „Die Mortalitätswahrscheinlichkeiten der beiden Therapiearme unterscheiden sich." oder „Die Mortalitätswahrscheinlichkeit der mehrfach-behandelten Gruppe ist größer/kleiner als die der einfach behandelten Gruppe."

Illustration der statistischen Power:

1) Abhängigkeit von der Fallzahl

Statistische Power zur Aufdeckung des Unterschiedes zwischen den beiden Therapiearmen (Mortalität: 21% vs. 13%) bei einem Signifikanzniveau von 5%

Fallzahl	Power
50	0.11
100	0.18
200	0.32
300	0.45
500	0.66
1000	0.92

2) Abhängigkeit vom Signifikanzniveau

Statistische Power zur Aufdeckung des Unterschiedes zwischen den beiden Therapiearmen (Mortalität: 21% vs. 13%) beieiner Fallzahl von 500

Signifikanzniveau	Power
0.001	0.18
0.01	0.43
0.05	0.66
0.1	0.77

3) Abhängigkeit von der konkreten Alternative

Statistische Power zur Aufdeckung des Unterschiedes zwischen den beiden Therapiearmen (Mortalität im einfachbehandelten Therapiearm: 21%) bei einer Fallzahl von 500 und einem Signifikanzniveau von 5%

Mortalität im anderen Arm	Power
17%	0.21
15%	0.42
13%	0.66
11%	0.86
9%	0.97

2. Laut Herstellerangabe enthält der oral-wirksame ACE-Hemmer Xanef zur Blutdrucksenkung 10mg Enalapril pro Tablette. Zur Qualitätskontrolle des Herstellers überprüft ein pharma-kologisches Institut den Wirkstoffgehalt von 100 zufällig ausgewählten Xanef-Tabletten.
Ergebnis:
Der Mittelwert der 100 Wirkstoffgehaltsangaben lag bei 9.81mg, die Standardabweichung der Messungen bei 0.76mg.
Frage: Hat der Hersteller ein Produktionsproblem bei der Fabrikation von Xanef-Tabletten?
(Erinnerung: N = 100, –X = 9.81, σ = 0.76)
Nullhypothese (H0): „Der Wirkstoffgehalt einer Xanef-Tablette ist 10 mg."
H0: Wirkstoffgehalt ist 10 mg
Alternative (H1): „Der Wirkstoffgehalt einer Xanef-Tablette ist von 10 mg verschieden." Oder „Der Wirkstoffgehalt einer Xanef-Tablette ist größer/kleiner als 10 mg."
H1: Wirkstoffgehalt unterscheidet sich von 10 mg
Prüfgröße T = (10 – 9.81) / 0.76 * √100 = 2.5
T ist unter H0 t-verteilt mit 99 Freiheitsgraden.
Signifikanzniveau α sei 0.05 ⇒ unter H0 gilt : P(|T| ≥ cα) = 0.05, wobei cα „kritischer Wert" heißt.
In unserem Fall (siehe Vertafelungen) ist cα = 1.99.
T = 2.5, T ist unter H0 t-verteilt mit 99 Freiheitsgraden ⇒ PHo (|T| ≥ 2.5) = 0.007 ist der zugehörige (zweiseitige) p-Wert des statistischen Tests
Einstichproben-Design:

- An einer Gruppe von unabhängigen Merkmalsträgern wird eine Zielgröße beobachtet. Abhängig vom Merkmalstyp der beobachteten Merkmalsaus-prägungen kommen verschiedene Testverfahren zum Einsatz.
- **Erinnerung an das Qualitätskontrollbeispiel:**
  Laut Herstellerangabe enthält der ACE-Hemmer Xanef zur Blutdrucksenkung 10mg Enalapril pro Tablette. Zur Qualitätskontrolle des Herstellers überprüft ein pharmako-logisches Institut den Wirkstoffgehalt von 100 zufällig ausgewählten Xanef-Tabletten.
- **Ergebnis:**
  Der Mittelwert der 100 Wirkstoffgehaltsangaben lag bei 9.81mg, die Standardabweichung der Messungen bei 0.76mg.

10) Epidemiologie

- geschichtliche Wurzeln
  - **16. Jahrhundert**: "De Contagiosum" - "mal aria"
  - **1833 – 1848**: Semmelweis' Interventionsstudie bei Wöchnerinnen ("Kindbettfieber")
  - **1850**: British Epidemiological Society, John Snow, Londoner Cholera-Epidemie
  - **80er Jahre**: HIV-Infektion
  - **Heutige Fragen** (Auswahl!): Verursacht Telefonieren mit Handys Hirntumoren?; Erzeugt Strahlung in der Umgebung von AKW Leukämien bei Kindern?; Beeinflusst häufiges „Sonnenbaden"das Melanom-Risiko?; Stellt Passivrauchen eine Gefahr für Nichtraucher dar?
- Definition:
  WHO: „Die Epidemiologie befasst sich mit der Untersuchung der Verteilung von Krankheiten, physiologischen Variablen und sozialen Krankheitsfolgen in menschlichen Bevölkerungsgruppen sowie mit Faktoren, die diese Verteilung beeinflussen.
- Epidemiologie versus
  - Individualmedizin: Gruppen statt Individuen, Beobachtung statt Einflussnahme
  - „klassische" Biometrie
    - Bevölkerungsstichproben –keine randomisierten Vergleichsgruppen
    - Untersuchung der „Bedingungen von Krankheit" -kein formalisierter Wirksamkeitsnachweis

- Stufen der epidemiologischen Vorgehensweise
  Abb.

- weitere Definitionen
  - Mortalität: Anzahl von Personen pro Zeiteinheit, die an einer bestimmten Ursache versterben.
  - Letalität (Fallsterblichkeit): Anteil von Personen, die eine bestimmte Erkrankung haben und daran versterben.
  - Morbidität: ungenau und summarisch für „Erkrankungszustand“
    - Prävalenz $= \frac{\text{Zahl bestehender Fälle}}{\text{Gesamtpopulation}}$
      - Krankenstand, Bestandsmaß (Punktprävalenz)
      - erweiterte Konzepte: Perioden-P., Lebenszeit-P.

      Zusammenhang Inzidenz-Prävalenz: Inzidenz → Prävalenz / Abheilung

    - Inzidenz (Veränderungsmaß) $= \frac{\text{Zahl der Neuerkrankungen pro Zeit}}{\text{Population pro Zeit}}$
      Neuerkrankungsrate:
      Bisher *nicht* Erkrankte werden über einen *definierten Zeitraum*beobachtet und die Häufigkeit neu auftretender Krankheitsfälle gezählt (z.B. 12 Melanomfälle / 100.000 / Jahr) geeignetes Maß um z.B. Zeittrends zu bestimmen (Bsp. Krebs: Erkrankte leben länger, Neuerkrankungsrate stabil, Prävalenz steigt)
      - Kumulative Inzidenz $= \frac{\text{Zahl der Neuerkrankungen}}{\text{Population}}$
        - häufig bezogen auf 100.000 Personen pro Jahr, gleichzeitig „Risiko“
        - Nenner: Zahl der Personen zu Beginn (=“population at risk”)
      - Inzidenzdichte $= \frac{\text{Zahl der Neuerkrankungen}}{\text{"Personenzeit"}}$
        - Nenner: Summe aller individuellen Beobachtungszeiten, die alle Teilnehmer der Kohorte zur Studie beitrugen, solange sie gesund und „unter Risiko“ waren
      - Personenzeit
        - von: Eintritt (Geburt, ..., Expositionsbeginn).
        - bis: Erkrankung oder Ende Beobachtungszeit
      - Unter der Annahme einer Population im Gleichgewicht (“steady state”):
        $ID \cdot T = \frac{P}{1 - P}$ wobei *T* die durchschnittliche Krankheitsdauerbezeichnet.

- **direkte Standardisierung**
  - **Direkte Altersstandardisierung**
    h*1,... , h*k: relative Anteile von Personen in **k Altersklassen** einer „Standardbevölkerung“
    I1,... , Ik: **beobachtete Inzidenzraten** in der untersuchten Region pro Altersgruppe
    → **direkt altersstandardisierte Rate:** $I_{dir} = \sum_{i=1}^{k} h_i^* \cdot I_i$
  - **Indirekte Alterstandardisierung**
    F: Anzahl **Fälle** in der **untersuchten** Bevölkerung
    h1,.., **hk**: relative Anteile von Personen in k Altersklassen der Studienpopulation
    I*1,... , **I*k**: **Inzidenzraten aus einer Standardbevölkerung** (*k* Altersklassen; Angabe „pro 100.000“)
    → **indirekt altersstandardisierte Rate** (der Standardbevölkerung), d.h. die „erwartete Rate“ :
    $I_{ind} = \sum_{i=1}^{k} h_i \cdot I_i^*$

und die indirekt standardisierte Mortalitäts- **/ Morbiditäts-Raten-Ratio:**

SMR= $\frac{\text{„gefundene Rate"}}{\text{„erwartete Rate"}}$

- Risikofaktoren
  - Faktoren („Expositionen"), die mit einer Erkrankung assoziiert sind
  - Häufungen der Erkrankung in Subgruppen (Alter, Geschlecht, Beruf, Region ...) geben oft erste Hinweise
  - ! Assoziation bedeutet nicht immer auch Kausalität; zusätzliche Kriterien müssen dafür erfüllt sein
- analytische Epidemiologie
  - Ziel: Untersuchung von Hypothesen zur Beurteilung möglicher kausaler Zusammenhänge
  - Voraussetzung: Studienplan und systematische Erhebung (≠ zufällige Beobachtungen!)
  - Wichtigste Studientypen: Querschnittsstudien, Längsschnitt- (Kohorten-) studien, Fall-Kontroll-Studien; zahlreiche Spezial- und Mischformen
  - Maßzahlen:

    Analytische Studien untersuchen die Stärke eines Zusammenhanges zwischen Expositionen („Risikofaktoren") und einer Krankheit; sie quantifizieren das *individuelle* Risiko

    Interessierende Größe ist **Relatives Risiko**:
    - das (theoretische) Verhältnis zwischen der Inzidenz in einer Population, die einer Exposition „ausgesetzt" ist, und der (theoretischen) Inzidenz, die in der selben Population aufträte, wenn keine Exposition vorgelegen hätte

$$RR = \frac{\frac{I_E}{N}}{\frac{I_0}{N}}$$

I_E: Inzidente Fälle bei Exposition
I_0: Inzidente Fälle bei Nicht-Expos.
N: Gesamte Population

⇒ nie messbar!

    - Schätzung des relativen Risikos

➢ Relative kumulative Inzidenzen (RR)

Beispiel: einfachster Fall einer dichotomen Erkrankung K bzw. Exposition E

	K=1	K=0	Gesamt
E=1	a	b	n_1
E=0	c	d	n_0
Gesamt	m_1	m_0	t

$$RR = \frac{P(K=1 \mid E=1)}{P(K=1 \mid E=0)}$$

Schätzer für die relativen kumulativen Inzidenzen:

$$\hat{RR} = \frac{\hat{CI}_E}{\hat{CI}_0} = \frac{a/n_1}{c/n_0} \qquad \hat{SE}(\ln \hat{RR}) = \sqrt{\frac{c}{(a \cdot n_1)} + \frac{d}{(b \cdot n_0)}} \qquad KI_{95\%}(\widehat{RR}) = e^{(\ln \widehat{RR} \pm 1{,}96 \cdot SE(\ln \widehat{RR}))}$$

Test auf Signifikanz: χ2-Test bzw. exakter Fisher-Test

➢ Inzidenzraten-Ratio (IRR)

	Fälle	Personenzeit
E=1	a	PT_e
E=0	b	PT_0
Gesamt	m_1	t

Schätzer für die relative Inzidenzraten:

$$\widehat{IRR}_E = \frac{\hat{IR}_E}{\hat{IR}_0} = \frac{a/PT_E}{b/PT_0} \qquad \widehat{SE}(\ln \widehat{IRR}) = \sqrt{\frac{1}{a} + \frac{1}{b}} \qquad KI_{95\%}(\widehat{IRR}) = e^{(\ln \widehat{IRR} \pm 1{,}96 \cdot SE(\ln \widehat{IRR}))}$$

➢ Odds Ratio (OR)

Von besonderer Bedeutung für die Epidemiologie, da es als Grundlage für den Risikoschätzer *Odds Ratio* (OR) dient, ist das **Odds** („Chancenverhältnis").

$$Odds = \frac{\text{Wahrscheinlichkeit}}{(1 - \text{Wahrscheinlichkeit})} = \frac{P}{(1-P)}$$

bei kleinen Wahrscheinlichkeiten nähert sich das *Odds* der Wahrscheinlichkeit stark an

	K=1	K=0	Gesamt
E=1	a	b	n_1
E=0	c	d	n_0
Gesamt	m_1	m_0	t

$$\hat{OR} = \frac{a/c}{b/d} = \frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c}$$

Willkommene Invarianz der OR: OR krank zu sein unter der Exposition = OR exponiert zu sein unter der Erkrankung!

$$\widehat{SE}(\ln \widehat{OR}) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \qquad KI_{95\%}(\widehat{OR}) = e^{(\ln \widehat{OR} \pm 1{,}96 \cdot SE(\ln \widehat{OR}))}$$

Test auf Signifikanz: χ2-Test bzw. exakter Fisher-Test

- **attributales Risiko**
  - Relevant für gesundheitspolitische Entscheidungen: das Risiko auf *Bevölkerungsebene*-Was hat Priorität: seltene Risikofaktoren mit hohen RR? häufige Risikofaktoren mit kleinen RR?
  - Das AR ist ein Maß für die Auswirkungen eines Risikofaktors auf Bevölkerungsebene, das berücksichtigt, wie häufig die die Exposition in der Bevölkerung vorkommt

$$AR := \frac{P(K=1) - P(K=1|E=0)}{P(K=1)}$$

$$= \frac{P(E=1) \cdot (RR - 1)}{P(E=1) \cdot (RR - 1) + 1}$$

P(E=1): Prävalenz der Exposition

  - Allgemeine Interpretation: „Vermeidbare Fälle"; d.h. der Anteil Krankheitsfälle (an allen Fällen) in einer Bevölkerung, die zusätzlich entstanden sind, weil die Exposition in der Bevölkerung vorkommt

- deskriptive Epidemiologie
  - Inhalt: systematische Untersuchung (räumlich, zeitlich) der Morbidität und Mortalität
    - Survey: Querschnitt / Querschnitts-Vergleiche
    - Register: Längsschnitt-Vergleiche
  - Nutzen
    - Identifikation eventueller räumlicher oder zeitlicher Unterschiede; Hypothesengenerierung (Ursache[n]?)
    - „Gesundheitsberichterstattung"
  - Voraussetzung
    - Standardisierte (räumlich und zeitlich) und zuverlässige Erhebungsmethodik
    - Aufbereitung und Veröffentlichung der Daten in interpretierbarer Form
  - Studientypen der deskriptiven Epidemiologie
    - Ziel: Beschreibung und Vergleiche
      - „Beobachtungsstudien": Fallberichte, Fallserien
      - **Fall-Kontroll-Studien (FKS)**
        - Ermittlung von Risikofaktoren
        - Prinzip: Vergleich von neu erkrankten Personen ("inzidenten Fällen") mit nicht erkrankten Personen bzgl. Verteilung von Risikofaktoren
        - Odds Ratio: (seltene Erkrankung: $\widehat{OR} \approx \widehat{RR}$)
        - keine Information über absolute Inzidenz bzw. Inzidenzdifferenz
        - Vorgehen
          - Krankenhausbezogene FKS:
            Definition der Fallserie Wahl einer geeigneten Kontrollgruppe (z.B. andere Pat. derselben Klinik); Problem: Übereinstimmung der Bezugsbevölkerung?
          - Populationsbezogene FKS:
            Definition der Bezugsbevölkerung Rekrutierung aller inzidenten Fälle in der Bezugsbevölkerung; Rekrutierung von Kontrollpersonen aus der Bezugsbevölkerung (z.B. Zufallsstichprobe)
          - Bei Fall- und Kontroll-Gruppe: identische (!) Erfassung von Expositionen

- ❖ Kontrollen
  - sollten bezüglich der Expositionsprävalenz repräsentativ sein für die Bezugsbevölkerung der Fälle (keine Patienten mit expositionsbezogenen Erkrankungen)
  - sollten, wären sie erkrankt, Teil der Fallgruppe geworden sein
  - sollten zeitgleich mit den Fällen rekrutiert werden. Sie bleiben Kontrollen, auch wenn sie später erkranken
  - sollten den gleichen Erhebungsprozeduren unterzogen werden wie Fälle
- ❖ Hauptvorteile: besonders effizient für seltene Erkrankungen; Untersuchung mehrerer Risikofaktoren möglich; weniger zeit- und kostenaufwändig als Kohortenstudien
- ❖ Hauptnachteile: retrospektive, nicht blinde Erhebung von Exposition und Kovariablen (⇒ *Recall-Bias;* Selektion der Kontrollgruppe oft problematisch (Zweifache Stichprobe ⇒ *Selection-Bias* ?*;* uneffizient für seltene Expositionen; zeitliche Sequenz nicht immer klar; keine Ermittlung absoluter Inzidenzraten bzw. Risiken möglich

➢ Registerstudien

➢ **Längsschnittstudien**

- ❖ Synonym: Kohortenstudien
- ❖ Kohorte in der Epidemiologie: eine Personengruppen, die ein bestimmtes Merkmal gemeinsam hat (z.B.: Geburtsjahr (⇒Geburtskohorte); Arbeitsplatz in einer best. Firma; Wohnort...)
- ❖ **geschlossene („fixe“) Kohorte**:
  Beobachtung Aller ab einem gemeinsamen Startpunkt über einen festen Zeitraum (*Follow-up*) →kumulative Inzidenz
- ❖ **offene („dynamische“) Kohorte**:
  individuelle Start- und Endpunkte für einzelne Personen → Personenzeit / Inzidenzrate
- ❖ Vergleich von mindestens zwei Kohorten (mit/ohne interessierender Exposition) in der Kohorten*studie*
- ❖ Auswertungen von Kohortenstudien: Inzidenz
  Betrachtung von neu über einen **Zeitraum** auftretenden Ereignissen
  Auswertung: Inzidenz-Vergleiche zwischen Kohorten mit und ohne Exposition, Schätzung des relativen Risikos bzw. der Odds Ratio (logistische Regression; Cox-Regression)
- ❖ Vorteile von Kohortenstudien
  zeitliche Folge (E→K) klar, d.h. validerer Schluss „von Assoziation auf Kausalität“
  seltene Expositionen können durch gezielte Auswahl untersucht werden
  Biasfreie Expositionsermittlung
  mehrere Krankheiten können untersucht werden, oder mehrere Definitionen eines Erkrankungs-Komplexes als Zielereignis definiert werden
  Direkte Inzidenz-Berechnung
- ❖ Nachteile von Kohortenstudien
  aufwändig (teuer!)
  seltene Krankheiten erfordern die Beobachtung *großer* Kohorten
  Krankheiten mit *langer Induktionszeit* oder *langer Latenz* erfordern *lange* Beobachtungszeit
  nur relativ wenige Expositionen können untersucht werden
  In jedem Fall relativ lange Zeit, bis Ergebnisse vorliegen
- ❖ Beispiel Kohortenstudie
  Zusammenhang zwischen Passivrauchen und KHK bei Nichtrauchern, prospektive Erhebung („*NursesHealth Study Cohort*“); 32.046 Frauen; 10 Jahre Follow-up ⇒

300.325 Personenjahre; 152 inzidente KHK [Kawachi et al. (1997) Circulation 95: 2374-9]
Auswertung: Berechnung der Inzidenzraten, multiple logistische Regression → $\widehat{OR} \approx \widehat{RR}$

- **Querschnittsstudien** (Prävalenz in verschiedenen Subkollektiven)
  - Untersuchung einer Stichprobe aus interessierender Grundgesamtheit an einem „Stichtag“
  - Bestimmung von Krankheitsstatus, auch bezüglich mehrerer Erkrankungen (sowie aktueller oder früherer Exposition[en])
  - Vorteil: schnelle Orientierung, „preiswert“ (z.B. via Fragebogen)
  - Einschränkung: nur bei chronischen Krankheiten sinnvoll, die nicht selten (in der Subgruppe) sind; keine Inzidenz ermittelbar
  - Bsp.: Leipziger Allergiestudie

Sensibilisierung (Pricktest)	'91/'92	'95/'96
Milbe (Dpt.)	4,6	8,1
Hund	2,8	2,6
Graspollen	9,1	11,5
Birkenpollen	8,4	14,2

Innenraum-Bedingungen (Kohle vs. Gas, Fensterdichtigkeit, Passivrauchen, Teppichböden, Haustiere), Ernährung, Kinderzahl, Verkehr
  - Untersuchung einer Stichprobe aus interessierender Grundgesamtheit an einem „Stichtag“
  - Bestimmung von Krankheitsstatus, sowie aktueller oder früherer Expositionen
  - Vorteil: „preiswert“, Hypothesengenerierung, mehrere Expositionen erfassbar
  - Nachteil: zeitliche Abfolge Exposition vs. Krankheit unklar („Reversed Causality“?), anfällig für Selektionsfehler
    → relativ schwache Evidenz für Kausalität
  - Auswertung

	K=1	K=0	Gesamt
E=1	a	c	n_e
E=0	b	d	n_0
Gesamt	m_1	m_0	t

Prävalenz: Exponierte: $\frac{a}{n_e}$ Nicht Exponierte: $\frac{b}{n_0}$
P.-Ratio: $\frac{\frac{a}{n_e}}{\frac{b}{n_0}}$ Odds Ratio: $\frac{\frac{a}{c}}{\frac{b}{d}}$

In multifaktorieller Situation: z.B. logistische Regressionsanalyse
Bei stetigen Daten: z.B. lineare Regression

- **Ökologische Studien** (Beobachtungen auf aggregierter Ebene)
  - Synonym.: Ökologische Relationen, Korrelationsstudien
  - Zusammenhang Einflussfaktoren - Erkrankungsraten auf aggregierter Ebene, z.B.Lungenkrebsrate auf Betriebsebene Rotweinkonsum und KHK-Mortalität auf Länderebene
  - Hauptvorteil: schnell, billig, effektiv bei großer Variation von Morbidität und Einflussfaktoren
  - Hauptnachteil: Gefahr des "ökologischen Trugschlusses“: Zusammenhang besteht nur auf aggregierter, nicht aber auf individueller Ebene
  - Bsp.: ökologische Studien
    Frage: Gibt es einen Zusammenhang zwischen Ernährung und koronarer Herzerkrankung (KHK)?
    Daten: Aggregierte (länderspezifische) Daten zum Verzehr (Fett-, Weinkonsum) sowie zur KHK-Mortalität.
    Berechnung der Korrelation zwischen Fettkonsum und KHK-Mortalität (mit „Korrektur“ für Weinkonsum)
  - Bsp.: ISAAC-Studie
    13- bis 14jährige in 56 Ländern (N=463.801)

Prävalenz von Asthma, Rhinokonjunktivitis und atopischem Ekzem

	Asthma	Ekzem
Großbrit.	30%	16%
Deutschl.	14%	8%(?)
Griechenl.	4%	4%

Fettverzehr und Atopie

Hoher Anteil von *trans*-Fettsäuren signifikant assoziiert mit
– Asthma bronchiale
– Atopischem Ekzem
– Allergische Rhinokonjunktivitis

- Fehlertypen in epidemiologischen Studien
  - Zur Erinnerung: Eine Assoziation, die man in einer Studie zwischen einer Exposition und dem Vorliegen der Zielgröße (Krankheit) gefunden hat, kann auf drei Gründen beruhen:
    - „echte“ Assoziation ☺
    - zufälliger Fehler (*error*)
    - systematischer Fehler (*bias*)
  - „Error“ versus „Bias“
    - Studienergebnisse = „Treffer“ auf einer Zielscheibe in der Mitte: Lage des „wahren Wertes“
    - Geringe Präzision ⇒ der zufällige Fehler ist groß
    - Keine Validität ⇒ es liegt ein systematischer Fehler vor
  - Systematische Fehler („Bias“)
    - **Bias**: „Any process at a stage of inference which tends to produce results or conclusions that differ systematically from the truth.“
    - Systematische Fehler können prinzipiell zu jedem Zeitpunkt einer Studie geschehen!
    - Übersicht: Systematische Fehler („*Bias*“)
      - **Selektionsfehler**
        - durch gestörte/s Rekrutierung und/ *Follow-up* auftretend
        - A, B, C und D: Stichproben, für eine unverzerrte Schätzung
        - $\widehat{OR}=\frac{A \bullet D}{B \bullet C}$ ist unverzerrt (*unbiased*)
        - ABER: Sind die Selektionswahrscheinlichkeiten pi in den vier Gruppen unterschiedlich (Ausfallrate.. etc.), liegt ein Selektionsfehler vor ⇒ Veränderung des Schätzers
        - *bei bekannten Selektionsraten pi:* $„OR_{Sel}“=\frac{p_a \cdot p_d}{p_b \cdot p_c}$ $\widehat{OR}_{unbiased}=\frac{\frac{a \bullet d}{b \bullet c}}{OR_{Sel}}$
      - **Informationsfehler**
        - können bei der Erhebung der interessierenden Merkmale (Expositions-, Krankheitsstatus) auftreten.
        - Veranschaulichung: Sensitivität und Spezifität bei der Erhebung < 100% ⇒ falsch Positive, falsch Negative...
        - nicht nur Veränderung einzelner Zellen proportional zu einem Faktor, sondern „Wandern“ von Personen zwischen den Zellen! ⇒komplizierte Veränderung der geschätzten Assoziation
        - Typisch z.B. **Expositions-Fehlklassifikation** in Fall- Kontroll-Studien, z.B. durch *recall bias*
      - **Confounding**
        - ... ist ein systematischer Fehler, der durch die mangelnde Berücksichtigung von Störgrößen (*Confoundern*) entstehen kann
        - Nicht als Confounder gelten „Zwischenstufen“ einer Erkrankung!
        - Maßnahmen gegen Confounding
          - Vorüberlegungen zu potentiellen Confoundern anstellen! dann evtl.: Restriktive Einschlusskriterien oder Matching (und entspr. Analyse)

- Mögliche statistische Verfahren in der Auswertung: Standardisierung, Stratifizierte Analyse, Analyse „gematchter Daten“ und Multiple Verfahren, z.B. logistische Regression
- Den Verfahren gemeinsam ist, dass sie die Vermengung von Effekten bzw. Zusammenhängen für das berechnete Maß „auflösen“ sollen bzw. einen „vergleichbaren“, vom Vorliegen von Kofaktoren unabhängigen, Schätzer der Assoziation liefern.

- logische Regression
  - Das allgemeine Modell der logistischen Regression ist:
    $\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n$
  - Dieses Modell ist aus verschiedenen Gründen in der Epidemiologie besonders geeignet. Hauptvorteil: Koeffizient $\hat{\beta}_i \Longrightarrow \widehat{OR}_i$
  - Zusammenhang zwischen OR und den Koeffizienten aus dem linearen Term: $\exp(\hat{\beta}_i) = \widehat{OR}_i$
  - Schätzung der $\widehat{OR}_i$ aus $\hat{\beta}_i$ Koeffizienten
    - dichotom klassierte Expositionen: $e^{\beta_i} = OR$ der Exponierten gegenüber Nicht-Exp., *wäre zum Beispiel*: $\hat{\beta}_1 = 0,531 \Rightarrow OR_1 = e^{\hat{\beta}_1} \approx 1,7$
    - ordinale Expositionen: $e^{\beta_i} = OR$ der der n-ten Stufe des Risikofaktors Exponierten gegenüber denen der (n-1)-ten Stufe Exponierten
    - stetige Expositionen: $e^{\beta_i} = OR$ zwischen denen, die einer bestimmten „Menge“ des Risikofaktors exponiert sind, gegenüber denen, die eine Einheit weniger exponiert sind

- EBM = evidence based medecine
  - Englisch: “evidence” = Nachweis, Beweis, d.h. Informationen aus klinischen Studien, die einen Sachverhalt belegen oder widerlegen
  - Problem: Informationsflut (> 2 Mio. Wiss. Artikel pro Jahr) vs. ärztliche Lesezeit (oft < 1 h / Woche)
  - 1988 Cochrane Collaboration gegründet, wichtigste EBM-Institution. Ziele: gemeinschaftliche, internationale Erstellung und regelmäßige Überarbeitung von systematischen Übersichten, Entwicklung von Methoden und Infrastruktur dafür
  - Die Anwendung besten verfügbaren Wissens in der täglichen Patientenversorgung - 5 Elemente:
    - Strukturierte klinische Fragestellung
    - Konsultation geeigneter Informationsquellen
    - Kritische Bewertung (“Hierarchy of evidence”)
    - Standardisierte Dokumentation des Ergebnisses
    - Kontinuierliche Verbesserung der Praxis EBM
  - Hierarchie der Beweiskraft: (siehe rechts)
  - systematische Übersicht:
    Einbezug aller zugänglichen Einzelstudien nach klar definierten Kriterien. Kritische Zusammenfassung je nach Evidenzgrad etc.
  - Meta-Analyse:
    - „Analyse von Analysen“ mit quantitativer Zusammenfassung von Einzelstudien-Ergebnissen, sofern ausreichende Homogenität gegeben ist.
    - Definition der Fragestellung
    - Systematische Literatursuche mit Prüfung der Ein-/Ausschlusskriterien für Einzelstudien
    - Beurteilung der Qualität der Studien
    - Strukturierte (grafische) Darstellung der Ergebnisse (z.B. „Forest-Plot“)
    - Bei ausreichender Homogentität: quantitative Zusammenfassung von Studienergebnissen mit Hilfe statistischer Methoden

- Fallstricke
  - Publikationsbias: Nur Studien, die einen(signifikanten) Effekt zeigen, werden veröffentlicht (→ Effekt-Überschätzung)
  - Selektionsbias: meist nur englischsprachige Artikel
  - Evtl. Heterogenität bei Einzelstudien (→besser nach Ursache suchen als zu poolen)
  - Insgesamt ist das Ergebnis der Meta-Analyse genauer (↓ KI) als das einer Einzelstudie, aber ebenso Bias-anfällig

- Quellen
  - Computerisierte bibliographische Datenbanken (MEDLINE, EMBASE)
  - Literaturverzeichnisse in Publikationen
  - Abstracts, Tagungsbeiträge
  - Studienregister (z. B. „Cochrane Controlled Trials Register“)
  - Evtl. Informationen zu nicht publizierten Studien
- gepoolte Auswertung
  - Ausreichende Homogenität der Studienergebnisse
  - Schätzung eines mittleren, gepoolten Effektes (hier: OR) mit KI aus den gewichteten Einzelstudien-Ergebnissen (Gewicht: Zahl der Studienteilnehmer)
  - Robustheit des Ergebnisses kann durch Sensitivitätsanalysen geprüft werden, z.B. durch Ausschluss bestimmter Studien