Computational Health Informatics

Vorlesung WiSe 2020/21

- Medizinische Statistik

1. Motivation
2. Medizinische Studien
3. Stochastik
4. Deskriptive Statistik
5. Induktive Statistik

Medizinische Statistik

Motivation

- Erkenntnisse und Entscheidungen in der Medizin sind naturgegeben mit Unsicherheiten behaftet
- Es ist unmöglich alle Einflussfaktoren auf ein Merkmal zu kennen
  - Vorgänge und Zusammenhänge im Menschen sind zu komplex
- Auch bei bekannten und gesicherten Risikofaktoren für eine Krankheit gibt es z.T. extreme Ausreißer

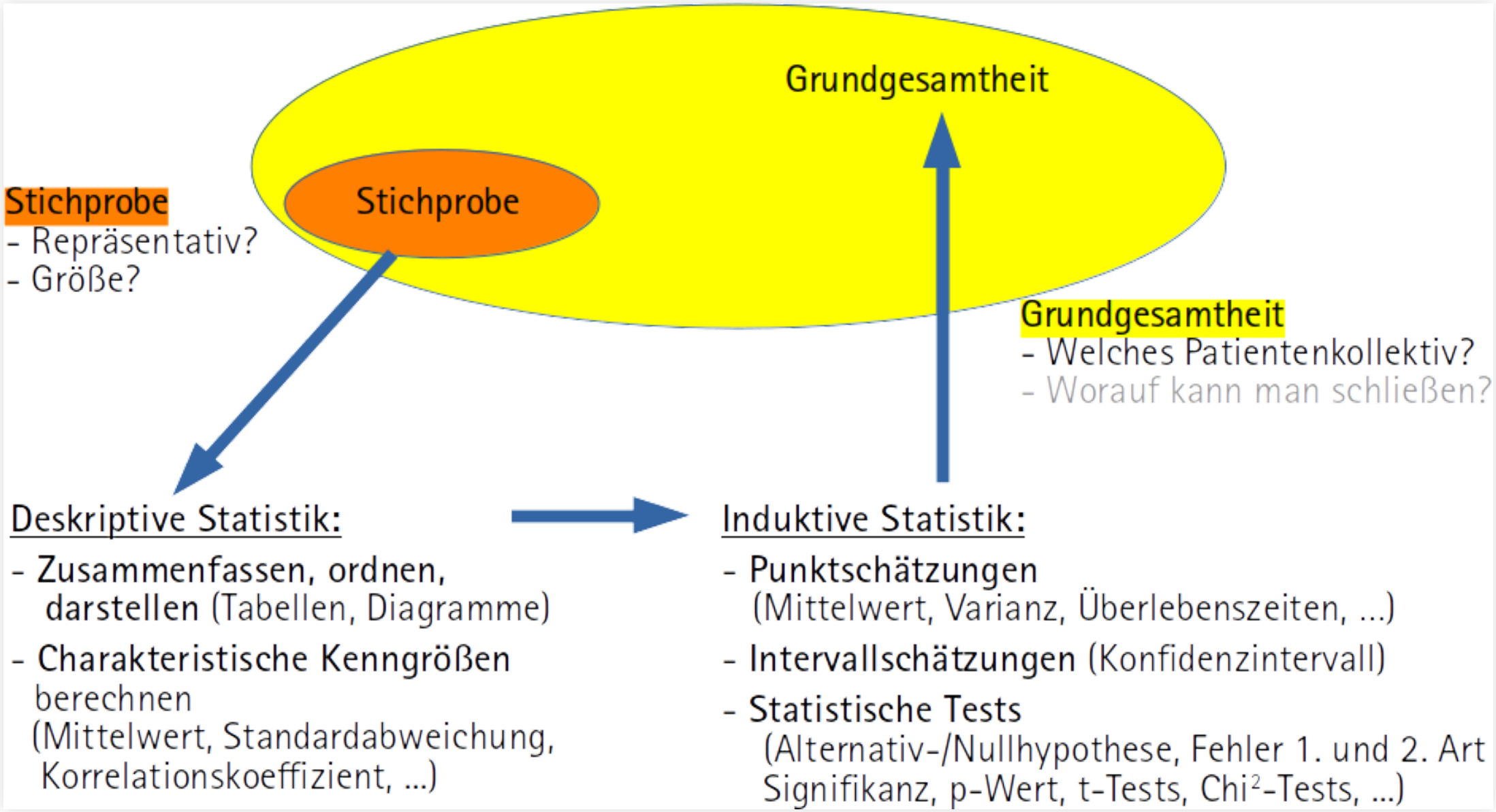
- Statistische Methoden erlauben es, trotz Unberechenbarkeit von Einzelvorgängen allgemein gültige Aussagen herzuleiten
  - Basis für wissenschaftliche Erkenntnis
  - Basis für das ärztliche Handeln

Stochastik

- Wissenschaft der mathematischen Behandlung von Zufallserscheinungen
- Wahrscheinlichkeitsrechnung
  - Mathematisch-theoretische Grundlagen für induktive Statistik (Insb. theoretische Verteilungen: Binomial-, Poisson-, Normalverteilung, …)
  - Fachspezifische Anwendungen:
    - Medizinische Statistik
    - Qualitätssicherung in der Medizin

Stochastik

- Statistik
  - Deskriptive Statistik (Strukturierung, Zusammenfassung, Darstellung)
  - Induktive Statistik (Schließen auf → Grundgesamtheit)


Grundgesamtheit
Stichprobe
Stichprobe
- Repräsentativ?
- Größe?
Grundgesamtheit
- Welches Patientenkollektiv?
- Worauf kann man schließen?
Deskriptive Statistik:
- Zusammenfassen, ordnen, darstellen (Tabellen, Diagramme)
- Charakteristische Kenngrößen berechnen (Mittelwert, Standardabweichung, Korrelationskoeffizient, ...)
Induktive Statistik:
- Punktschätzungen (Mittelwert, Varianz, Überlebenszeiten, ...)
- Intervallschätzungen (Konfidenzintervall)
- Statistische Tests (Alternativ-/Nullhypothese, Fehler 1. und 2. Art Signifikanz, p-Wert, t-Tests, Chi²-Tests, ...)

Deskriptive Statistik

- **Merkmale** sind Eigenschaften, die für die Studie relevant sind und statistisch ausgewertet werden
- Klassifikation nach verschiedenen Aspekten:
  - Diskret oder stetig
  - Skalenniveau
    - Nominalskala: Nur begriffliche Unterscheidung
    - Ordinal-/Rangskala: Natürliche Rangfolge
    - Invervallskala: Nullpunkt, negative Werte
    - Verhältnisskala: Nullpunkt, nur positive Werte

- Unterscheidung in Ziel- und Einflussgrößen
  - Ziel einer Studie: Erkenntnisse über **Zielgrößen**
  - Einflussgrößen:
    - **Faktoren:** Werden erfasst und ausgewertet
    - **Begleitmerkmale:** Werden eventuell erfasst, aber in aktueller Studie nicht ausgewertet
    - **Störgrößen:** Werden nicht berücksichtigt

Beispiel

- Hypothese: Rauchen beeinflusst die Entstehung von Lungenkarzinomen
  - Zielgröße: Entstehung von Lungenkarzinomen
  - Zu untersuchender Faktor: Rauchen
  - Begleitmerkmale: Z.B. das Alter
  - Störgrößen: Genetische Veranlagungen, Umwelteinflüsse, …

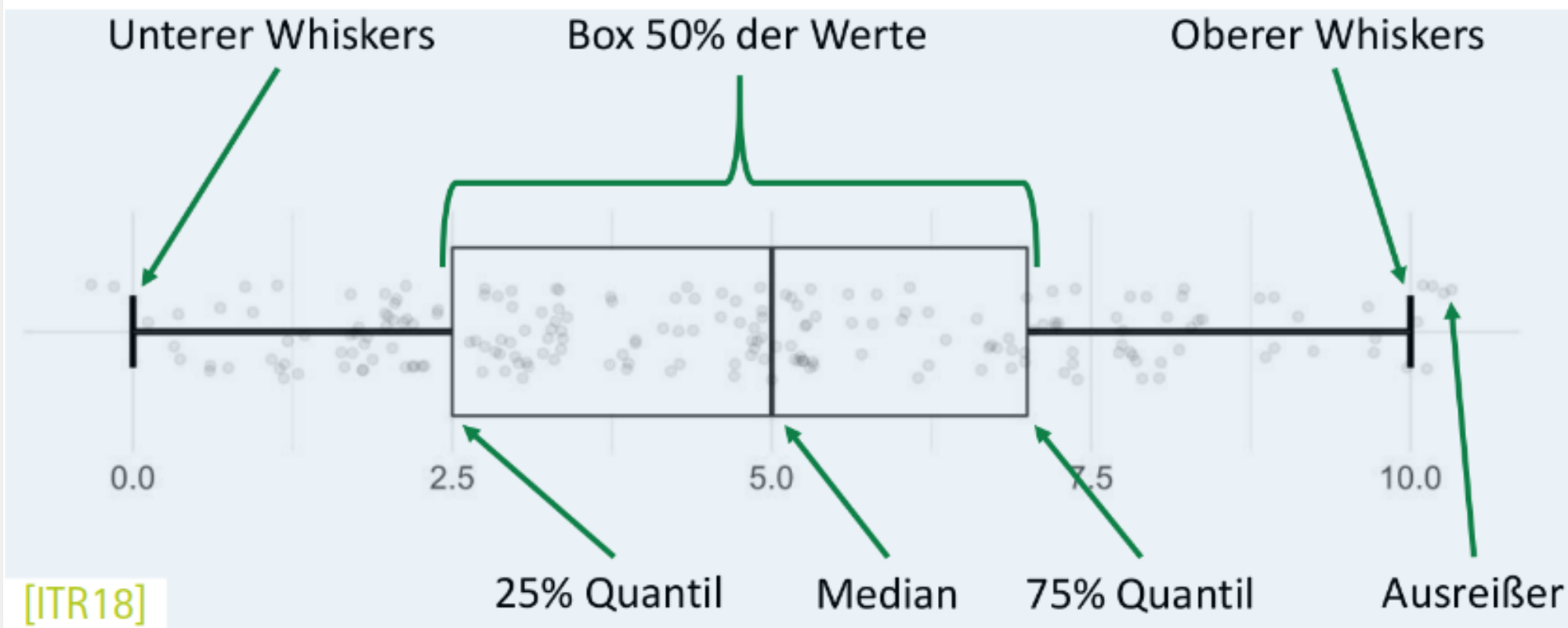
Charkteristische Kenngrößen (Auswahl)

- **Univariate Datenbeschreibung:** Ein Merkmal
  - **Lagemaße:** Arithmetisches Mittel ("Mittelwert", "Durchschnitt"), Median, Quartile und Quantile, Modus, geometrisches Mittel, …
  - **Streuungsmaße:** Varianz, Standardabweichung, Variationskoeffizient, Spannweite, …
  - **Formmaße:** Schiefe, Wölbung, …

Box-Plot

Box-Plot

- Innerhalb der **Box** liegen 50% der Datenwerte
- Kleinsten 25% bzw. 75% der Datenwerte sind ≤ **25%-** bzw. **75%-Quantil**
- Mittlerer Strich gibt den **Median** an
  - Kleinsten 50% der Datenwerte sind ≤ **Median**
- Whisker zeigen als Wert **Median ± 1,5 · IQA** an
  - IQA: **Interquartilsabstand** = Bereich zwischen 25%- und 75%-Quantil



[ITR18]

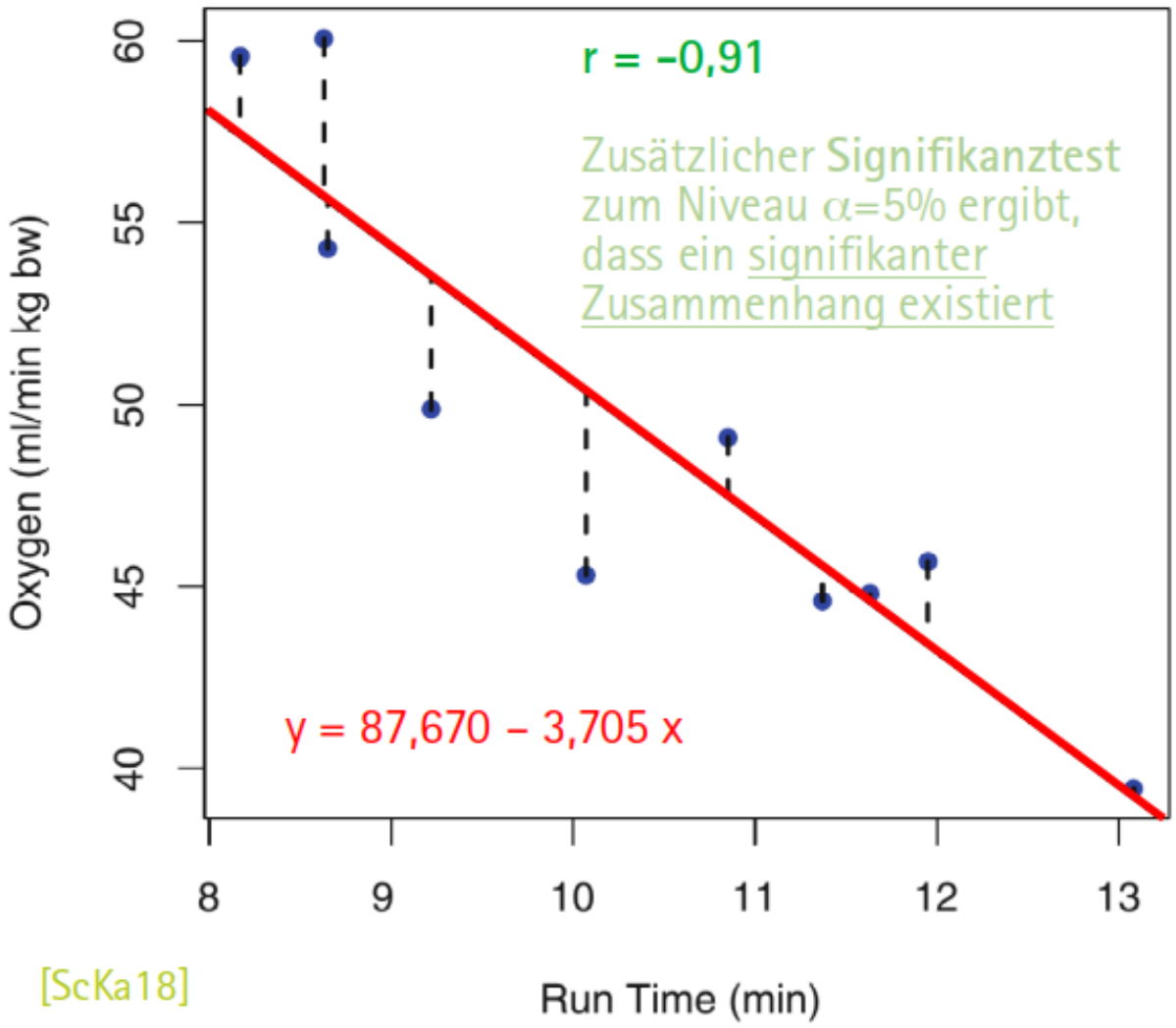
Charkteristische Kenngrößen (Auswahl)

- **Bivariate Datenbeschreibung:** Zwei Merkmale
  - Gibt es einen *stochastischen* (nicht funktionalen) Zusammenhang?
    - Darstellung durch **Punktwolken**, …
    - **Korrelationsanalyse**: Kovarianz, Pearson’scher Korrelationskoeffizient, …
    - Vorsicht bei Interpretationen!

- Bivariate Datenbeschreibung (Forts.)
  - **Regressionsanalyse**
  - Häufig eingesetzt (Ursache/Wirkungsanalysen)
  - Regressionsgrade bei linearem Zusammenhang
  - Nichtlineare Regression sonst
  - Regression 1. Art: x gegeben, y Zufallsvariable
  - Regression 2. Art: x und y Zufallsvariable
  - Herleitung einer mathematischen Gleichung für den Zusammenhang (→ Theorie!)

Beispiel

- Fitness eines Menschen messen über die Fähigkeit, Sauerstoff aufzunehmen
  - Schwer messbar, daher Untersuchung, ob stattdessen ein Surrogat-Parameter genommen werden kann
  - → Gibt es einen Zusammenhang zwischen *run time* bei einem 3-km-Lauf und *oxygen*?



Stichprobenkovarianz quantifiziert die gemeinsame Abweichung

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

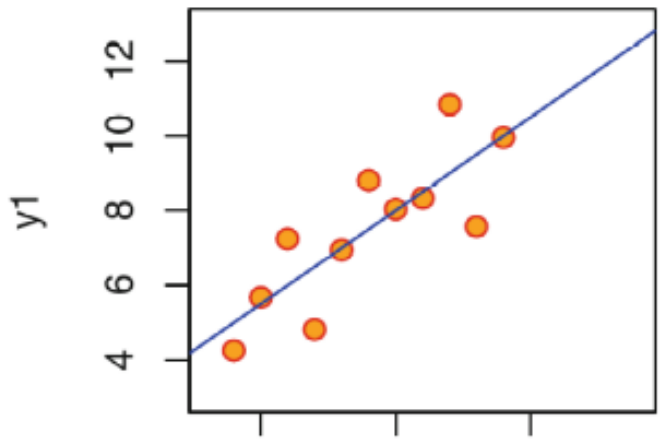
Vergleichbarkeit (Unabhängigkeit von Einheiten) durch Normierung:
→ Korrelationskoeffizient nach Pearson

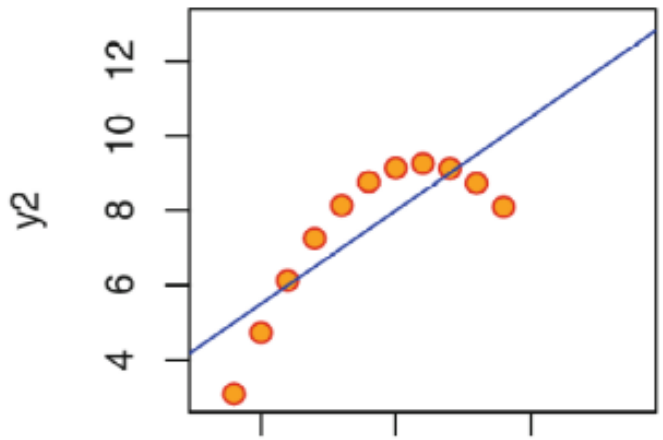
$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

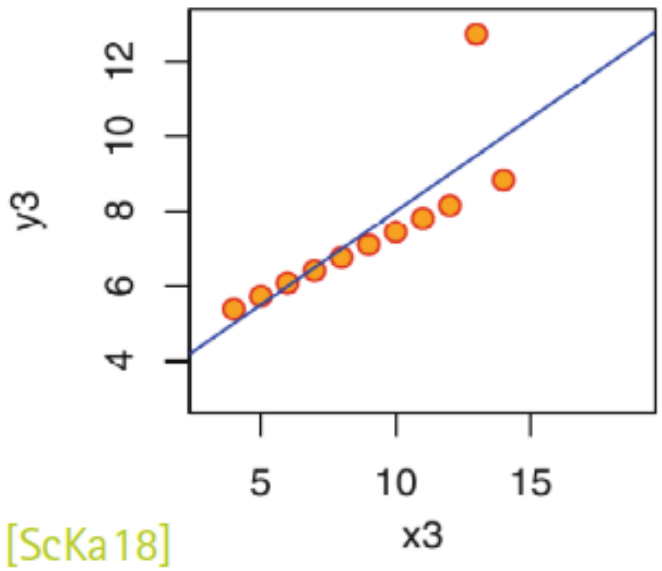
$r \in [-1, 1]$ $r = \pm 1$ bedeutet vollständiger positiver/negativer Zusammenhang

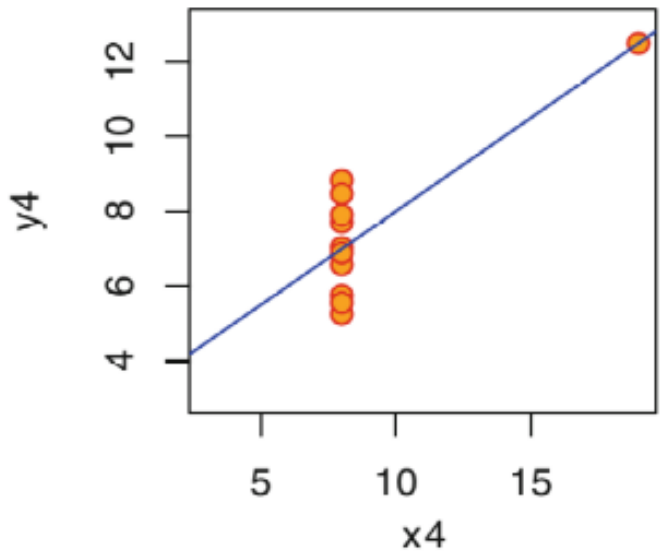
$\bar{x}$, $\bar{y}$: Mittelwerte der n Messwerte, s_x und s_y deren Standardabweichung (→ später Induktive Statistik)

Warnendes Beispiel









[ScKa18]

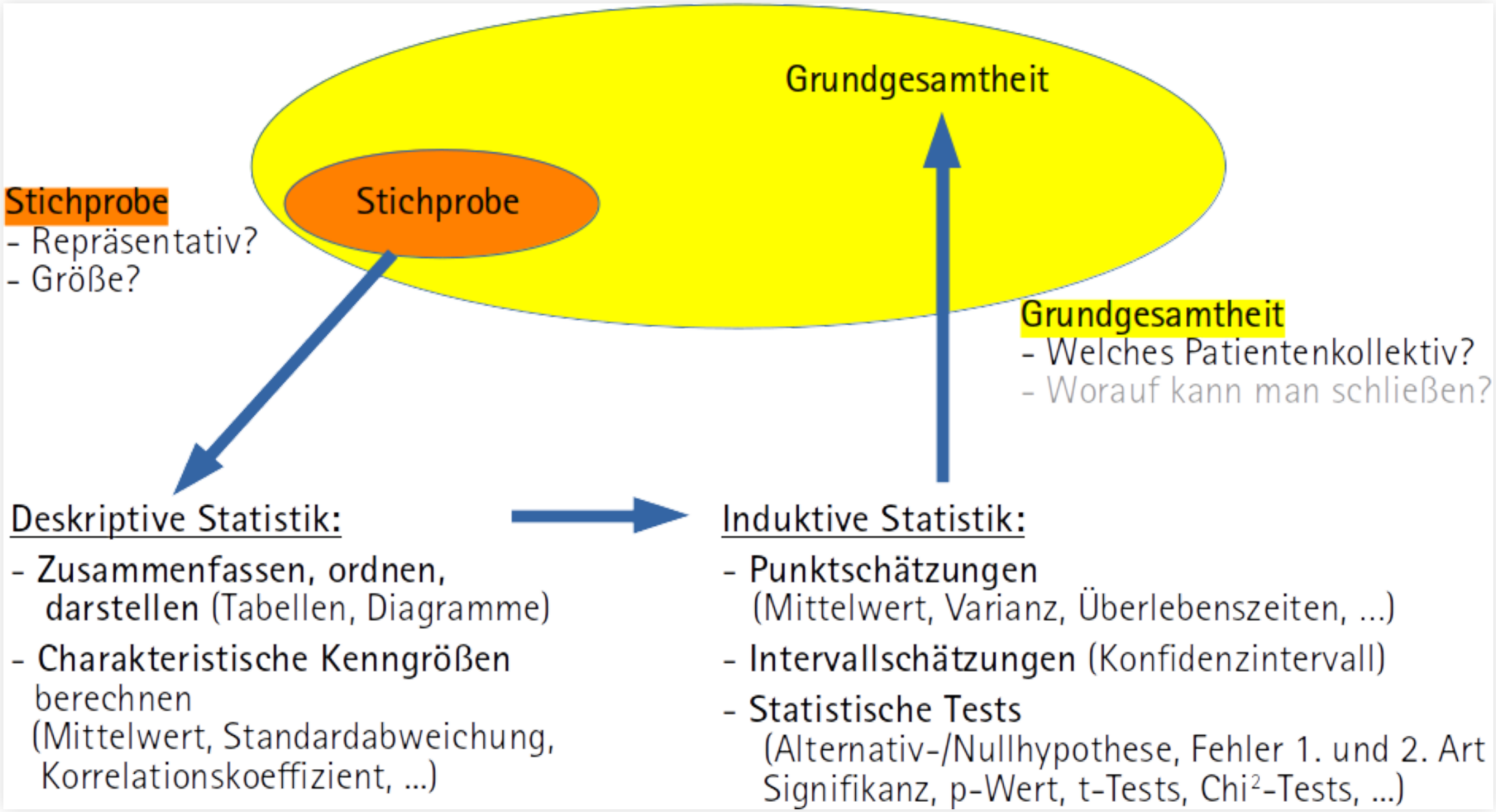
Anscombe-Quartett

- Vier Datensätze aus jeweils 11 Datenpunkten
- Alle vier Datensätze haben die gleichen statistischen Eigenschaften:
  - Mittelwert x = 9,00
  - Mittelwert y = 7,50
  - Standardabw. x = 3,32
  - Standardabw. y = 2,03
  - Korrelation x, y = 0,816
  - Lineare Regression:
    y = 3,00 + 0,50 x

=> Wichtigkeit, Daten vor der statistischen Auswertung grafisch darzustellen!

Vergleich mehrerer Stichproben

- Berechnung von Kenngrößen für jede Stichprobe
- Prüfung mit **statistischen Tests**, ob die Unterschiede der Mittelwerte zufällig bedingt sind, oder ob von einem **signifikaten Unterschied** zwischen den Stichproben ausgegangen werden kann; Bsp.:
  - Mehrere Therapieformen
  - Vor und nach einer Therapie
  - Erkrankte/gesunde Personen
  - Personen mit/ohne Risiko

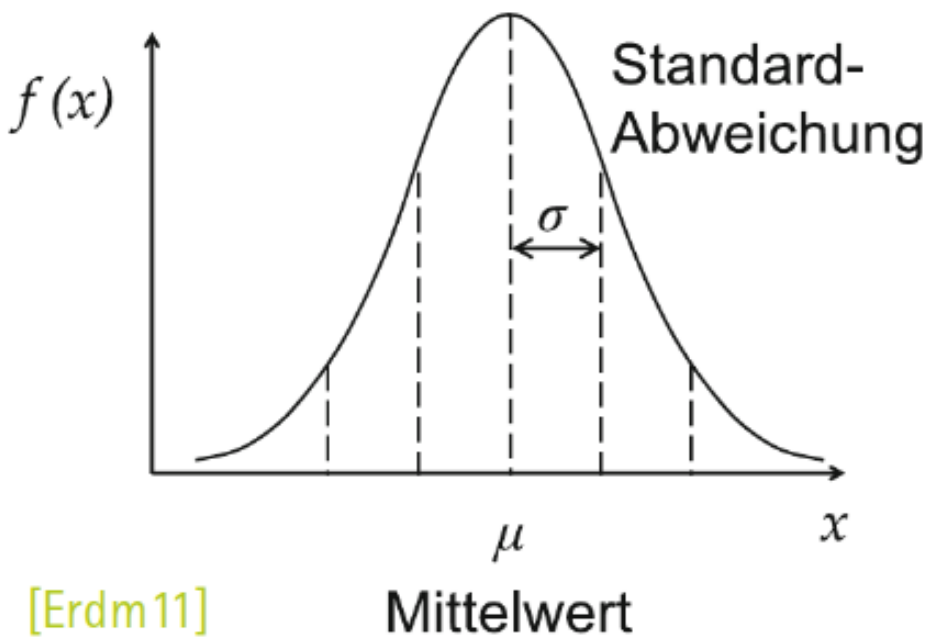

Grundgesamtheit
Stichprobe
Stichprobe
- Repräsentativ?
- Größe?
Grundgesamtheit
- Welches Patientenkollektiv?
- Worauf kann man schließen?
Deskriptive Statistik:
- Zusammenfassen, ordnen, darstellen (Tabellen, Diagramme)
- Charakteristische Kenngrößen berechnen (Mittelwert, Standardabweichung, Korrelationskoeffizient, ...)
Induktive Statistik:
- Punktschätzungen (Mittelwert, Varianz, Überlebenszeiten, ...)
- Intervallschätzungen (Konfidenzintervall)
- Statistische Tests (Alternativ-/Nullhypothese, Fehler 1. und 2. Art Signifikanz, p-Wert, t-Tests, Chi²-Tests, ...)

Induktive Statistik

- Ziel: Aus Stichprobenwerten Informationen bezüglich der Grundgesamtheit und der betrachteten Zufallsvariablen X gewinnen
- Punktschätzungen
  - Wahrscheinlichkeiten aus relativen Häufigkeiten berechnen

- (Punktschätzungen (Forts.))
  - Annahme: Nur statistische Fehler → Normal-verteilung (Gauß-Verteilung)

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



[Erdm11]

Normierte Gauss-Verteilung

- Gesamtfläche = 1
- Zentriert um μ
- Bei den Werten $\mu \pm \sigma$ ist die Gauss-Verteilung auf das 1/e – fache des Maximums abgefallen
- Die Fläche im Bereich zwischen diesen beiden Werten $\mu \pm \sigma$ beträgt 0,6826

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-\sigma}^{\mu+\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 68.26\%$$

Messwert innerhalb des Intervalls	Wahrscheinlichkeit
$\lvert x-\mu\rvert < 1\sigma$	68.26%
$\lvert x-\mu\rvert < 2\sigma$	95.45%
$\lvert x-\mu\rvert < 3\sigma$	99.73%

- (Punktschätzungen (Forts.))
  - Erwartungswert der Grundgesamtheit aus **arithmetischem Mittel $\bar{x}$ der Stichprobe** mit n (n»1) Werten:

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

- (Punktschätzungen (Forts.))
  - Varianz der Grundgesamtheit als Quadrat der **Standardabweichung s_x der Stichprobe**:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

  - In der Praxis wird diese Standardabweichung auch kurz (& ungenau) als "Fehler" bezeichnet

- (Punktschätzungen (Forts.))
  - Ergebnis: Schätzung des **"wahren" Werts** x_w der zugrunde liegenden Grundgesamtheit:

$$x_w = \bar{x} \pm s_x$$

  - Hier wurde der **Vertrauensbereich** s_x genutzt
    - Mit einer Wahrscheinlichkeit von 68,26% weicht der wahre Wert x_w von dem Mittelwert $\bar{x}$ um nicht mehr als s_x (absolut) ab
  - Manchmal wird als Vertrauens-/Konfidenzintervall $\pm 3 s_x$ angegeben ⇒ 99,73% statt 68,26%

- (Punktschätzungen (Forts.))
  - Der Mittelwert $\bar{x}$ und die Standardabweichung s_x der Stichprobe sind selbst Zufallsvariablen und damit Gauß-verteilt
  - Statt s_x wird oft auch die **Standardabweichung des arithmetischen Mittels** $\bar{x}$ angegeben: $\pm \frac{s_x}{\sqrt{n}}$
    - Diese Unsicherheit des Mittelwertes wird mit steigender Stichprobengröße n beliebig klein, allerdings steigt der Aufwand enorm an

- (Punktschätzungen (Forts.))
  - Bei kleiner Stichprobe (wenig Meßwerte n) erfolgt eine Aufweitung bei der Angabe des Konfidenz-intervalls durch Multiplikation von s_x mit einem zu n gehörenden Wert der "Student-" / t-Verteilung $t_{n-1;1-\frac{\alpha}{2}}$ (aus Tabelle ablesbar bzw. berechnen)
    - α ist die Irrtumswahrscheinlichkeit, häufig α=5% → Konfidenzintervall 95% (oder auch α=1%)

Fehlerfortpflanzung

- Wenn die zu bestimmende Größe f von zwei Zufallsvariablen x und y abhängt, dann kann aus der Funktionsbeziehung f(x,y) und den Standardabweichungen s_x und s_y der beiden Messgrößen x und y die Standardabweichung von f berechnet werden:

$$s_f = \sqrt{s_x^2 \left(\frac{\partial f}{\partial x} \right)^2 + s_y^2 \left(\frac{\partial f}{\partial y} \right)^2}$$

Quellen und weiterführende Literatur

- [ITR18] U. Hübner, M. Esdar, J. Hüsers, J.-D. Liebe, J. Rauch, J. Thye, J.-P. Weiß: **IT-Report Gesundheitswesen — Wie reif ist die IT in deutschen Krankenhäusern?**, Forschungsgruppe Informatik im Gesundheitswesen, 2018
- [Demt15] W. Demtröder: **Experimentalphysik 1**, 7. Auflage, Springer Verlag 2015

- [Erdm11] M. Erdmann, T. Hebbeker: **Experimentalphysik 1**, Springer 2011
- [ErHe13] M. Erdmann, T. Hebbeker: **Experimentalphysik 5**, Springer Spektrum 2013
- [WeHe13] K. Weltner, H. Wiesner, P. Engelhard, H. Schmidt: **Mathematik für Physiker und Ingenieure 1**, 17. Auflage, Springer Verlag 2013

- [TrWi00] H. J. Trampisch, J.Windeler, B. Ehle, S. Lange: **Medizinische Statistik**, 2. Auflage, Springer Verlag 2000
- [Weis08] C. Weiß: **Basiswissen Medizinische Statistik**, 4. Auflage, Springer Verlag 2008