# DESKRIPTIVE STATISTIK

# THEORETISCHE GRUNDLAGEN

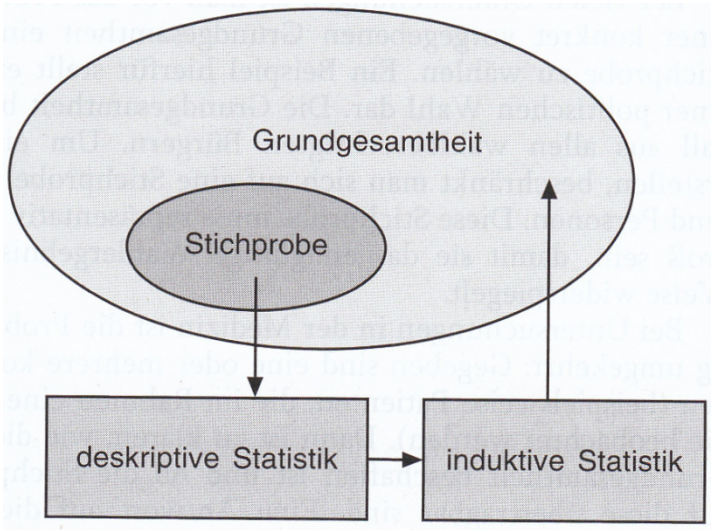
## Grundgesamtheit und Stichprobe

*Total- oder Vollerhebung:* Untersuchung bezieht sich auf gesamte Population (z.B. Todesursachenstatistiken)
*Stichprobe:* Untersuchung einer kleinen Teilmenge und Übertragung der gewonnen Erkenntnisse auf die Grundgesamtheit; Stichprobe muss repräsentativ und ausreichend groß sein; Daten einer Stichprobe $x_1,\ldots,x_k$ = Urliste; n … Stichprobenumfang
*repräsentative Stichprobe:* Charakteristische Eigenschaften der Stichprobe, abgesehen von zufällig bedingten Abweichungen, stimmen mit denen der Grundgesamtheit überein

Bei Untersuchungen in der Medizin sind eine oder mehrere konkrete Stichproben gegeben und man versucht zu klären, wie die Grundgesamtheit beschaffen ist.

## Aufgaben der deskriptiven Statistik

Die statistische Analyse besteht aus 2 Teilen:
- deskriptive Statistik
- induktive Statistik




### Deskriptive Statistik

… Auswertung der Daten der Stichprobe, um die charakteristischen Eigenschaften zu beschreiben; Vorstufe zur induktiven Statistik

*Aufgaben*
- Zusammenfassen und Ordnen der Daten in Tabellen
- Erstellen von Diagrammen
- Berechnen charakteristischer Kenngrößen oder Maßzahlen (Mittelwert, Standardabweichung, etc.)

bei 2 oder mehrere Stichproben:
- jede einzelne Stichprobe auswerten (graphische Darstellung, Kenngrößen berechnen)
- dadurch erhält man einen Überblick, ob und wie sich die Stichproben unterscheiden

### Induktive Statistik

… Benützung von Methoden, die die Ergebnisse, die aus den Stichproben gewonnen wurden, verallgemeinern und statistisch absichern

# Merkmale

## Grundbegriffe

*Untersuchungseinheiten, Merkmalsträger*
= Personen (z.B. Patienten) oder Objekte einer Stichprobe

*Beobachtungseinheiten*
= kleinste Einheiten, an den die einzelnen Beobachtungen registriert werden
z.B. Untersuchung beider Augen bei Patienten → Patienten sind die Untersuchungseinheiten, die einzelnen Augen stellen die Beobachtungseinheiten dar; oft: Untersuchungseinheit = Beobachtungseinheit

*Merkmale, Variablen, Zufallsvariablen*
Beobachtungseinheiten sind durch bestimmte Merkmale charakterisiert; Eigenschaften, die für die untersuchende Fragestellung relevant sind und statistisch ausgewertet werden

*Merkmalsausprägungen*
= Werte oder Ausprägungen, die ein bestimmtes Merkmal annehmen kann

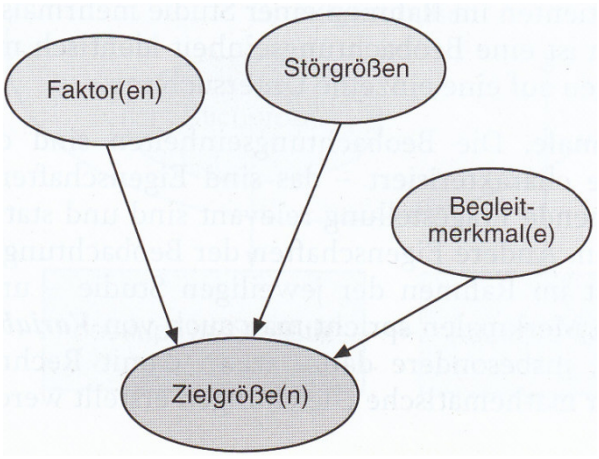
Art der Merkmale ist entscheidend für
- Planung und Durchführung einer Studie
- erforderlichen Stichprobenumfang
- geeignete Analysemethoden

Klassifizierung von Merkmalen
- nach ihrer Funktion bei der statistischen Analyse
- nach ihrem Skalenniveau
- danach, ob diskret oder stetig

## Ziel- und Einflussgrößen

Merkmale lassen sich in Ziel- und Einflussgrößen einteilen.
Zweck einer Studie ist es, Erkenntnisse über eine oder mehrere *Zielgrößen* zu gewinnen.
*Einflussgrößen:* Merkmale, die in einem funktionalen Zusammenhang zu den Zielgrößen stehen und möglicherweise beeinflussen

*Einteilung der Ziel- und Einflussgrößen*
- Faktoren
  … werden erfasst und ausgewertet
- Störgrößen
  … werden nicht berücksichtigt/erfasst
- Begleitmerkmale
  … werden erfasst, aber nicht ausgewertet (z.B. Nebenwirkungen)



*Beispiel:*
„Zigarettenrauchen beeinflusst das Entstehen eines Lungenkarzinoms"
Zielgröße: Entstehen eines Lungenkarzinoms
zu untersuchende Faktor: Zigarettenrauchen

Störgrößen können
- nicht-verzerrend
  … verantwortlich für zufallsbedingte Streuung der Daten
- verzerrend
  … werden fälschlicherweise in einen kausalen Zusammenhang mit der Zielgröße gebracht (Confounder) → Fehlinterpretation

sein.

**Klassifikation nach Skalenniveau**

Merkmale lassen sich in Skalenniveaus einteilen → Auskunft über das Messniveau und wie Daten weiterverarbeitet werden können

*Nominalskala*
- niedrigstes Niveau
- Ausprägungen unterscheiden sich nur begrifflich voneinander
- Beispiel: Augenfarbe, Bluttgruppe
- Spezielle Form: Alternativmerkmale, auch dichotome oder binäre Merkmale genannt
  - z.B. Ausprägungen „männlich" und „weiblich" (Geschlecht), „positiv" und „negativ" (Rhesusfaktor)

*Ordinalskala, Rangskala*
- höheres Niveau als die Nominalskala
- Ausprägungen dieser Merkmale lassen sich in einer natürlichen Rangfolge anordnen
- Beispiel: Noten (1-5), Scores, Therapieerfolg mit möglichen Abstufungen (vollständig geheilt bis Patient verstorben)

*→ Qualitative oder kategoriale Merkmale*
= Überbegriff für nominal und ordinal skalierte Merkmale
- Beispiel: Geschlecht kann mit 0=männlich und 1=weiblich klassifiziert werden
- Differenz zweier Ausprägungen anzugeben ist nicht sinnvoll – die Zahlen haben keine rechnerische Bedeutung (aber sinnvoll: A = B, A<B, A>B)

*Intervallskala, Abstandsskala*
- höherer Informationsgehalt als die Ordinalskala
- Ausprägungen unterscheiden sich zahlenmäßig
- Nullpunkt sowie negative Messwerte existieren
- möglich und sinnvoll Differenz zwischen zwei Ausprägungen A-B anzugeben
- Beispiel: Temperatur in Celsiusgraden

*Verhältnisskala, Ratioskala*
- hat absoluten Nullpunkt und nur positive Messwerte
- Bestimmung der Differenz sowie das Verhältnis A:B zweier Ausprägungen möglich
- Beispiel: Körpergewicht, Leukozytenanzahl pro µl Blut, Temperaturangabe in Kelvin-Graden

*→ Quantitative oder metrisch skalierte Merkmale*
= Überbegriff für intervall- und verhältnisskalierte Merkmale

| Merk-malsart | Skalenniveau | Beispiele | Hinweise | Vergleich 2er Ausprägungen |
|---|---|---|---|---|
| qualitativ | Nominalskala | Blutgruppe, Rhesusfaktor | niedrigstes Niveau | A=B, A≠B |
| qualitativ | Ordinalskala (Rangskala) | Zensuren, med. Scores | Rangfolge ist definiert | A=B, A≠B, A>B, A<B |
| quantitativ | Intervallskala (Abstandsskala) | Temperatur in °C | Skala mit festgelegtem Null-punkt, Abstand ist definiert | A=B, A≠B, A>B, A<B, d=A-B |
| quantitativ | Ratioskala (Verhältnisskala) | Leukozytenan-zahl/µl Blut, Körpergröße | höchstes Niveau, Skala mit absolutem Nullpunkt, Ver-hältnis ist definiert | A=B, A≠B, A>B, A<B, d=A-B, c=A:B |

**Diskrete und stetige Merkmale**

*Diskret*

… Merkmal kann nur abzählbar viele Werte annehmen

- umfasst alle qualitativen Merkmale
- NUR: quantitative Merkmale, die durch einen Zählvorgang ermittelt wurden (z.B. Anzahl der Schwangerschaften einer Frau, Anzahl richtig gelöster Klausuraufgaben)

*Stetig*

… Merkmal kann alle Werte innerhalb eines bestimmten Intervalls annehmen

- Ermittlung der Ausprägungen durch Messvorgang (z.B. Körpergröße, Blutdruck)
- durch begrenzte Messgenauigkeit wird Merkmal diskret (z.B. Körpergröße in cm wird auf- oder abgerundet)
- stetige Merkmale sind effizienter bei Informationsgewinnung und häufig leichter zu analysieren

→ Unterscheidung von diskreten und stetigen Merkmalen sinnvoll

**Skalentransformationen**

- höheres Skalenniveau auf ein niedriges zu transformieren
- jede Verhältnisskala ist automatisch Intervallskala
- Intervallskala kann als Ordinalskala aufgefasst werden
- Nominalskala kann grundsätzlich jedem Merkmal zugeordnet werden

| Ausprägungen | Merkmalsart | Skala |
|---|---|---|
| Menge des pro Tag konsumierten Tabaks in Gramm | quantitativ, stetig | Verhältnisskala |
| Anzahl der pro Tag gerauchten Zigaretten | quantitativ, diskret | Verhältnisskala |
| Nichtraucher – schwacher R. – mäßiger R. - starker Raucher | qualitativ | Ordinalskala |
| Nichtraucher – Raucher | qualitativ, binär | Nominalskala |

→ einfachere Messtechnik, aber Informationsverlust

Vor der Datenerfassung zu klären (um geeignete Analysemethode wählen zu können)

- Fragestellung der Studie formulieren (Hypothese)
- Geeignete Ziel- und Einflussgrößen wählen
- Spezifische Eigenschaften für jedes Merkmal bestimmen

**Merkmalsausprägungen**

- zu Beginn der Studie: Merkmale erheben und Skalenniveaus definieren
- danach: Ausprägungsliste für jedes Merkmal definieren
  - bei quantitativen Merkmalen: Mess- oder Zählwerte
  - Ausprägungen bei qualitativer Merkmale: numerisch codiert
- Liste muss
  - vollständig sein (für jede Beobachtung eine Ausprägung, bei qualitativen Merkmalen z.B. „Sonstiges" → Vollständigkeit)
  - disjunkt sein (2 Ausprägungen sind unterscheidbar und schließen sich gegenseitig aus, Zuordnung muss eindeutig sein)

## Besondere Problematiken

*Klinische Scores und Skalen*
- quantitative (Intervall- und Verhältnisskala) Merkmale lassen sich effizienter auswerten als qualitative (Nominal, - und Ordinalskala)
- Tendenz Sachverhalte, die eigentlich qualitativ beschreibbar sind, quantitativ messbar zu machen
- Einführung klinischer Scores und Skalen zur Erfassung komplexer Merkmale (z.B. Allgemeinzustand des Patienten) → „weiche Daten" („harte Daten" = exakte Messung)
- Beispiele
  - Apgar-Score (= Beurteilung des Zustands Neugeborener – Erhalt von Punkten für jedes eintreffende Merkmal [Herzfrequenz, Atmung, Reflexe, Hautfarbe, etc.] und Addierung dieser – Summe zw. 0 und 10)
  - Karnofsky-Skala (= Beurteilung des Allgemeinzustands eines Patienten, Werte zw. 0 und 100)
  - visuelle Analog-Skala (= Beschreibung der Schmerzintensität – Patient markiert auf 10cm langer Linie sein Schmerzempfinden)
- zwei gleiche Werte müssen nicht die gleichen Symptome beschreiben, sie sind also nicht unbedingt ident

*Ausreißer*
- extrem hohe oder extrem niedrige Werte
- Erkennung durch graphische Darstellung
- Nachforschung, wie diese Werte entstanden sind
  - Möglichkeit von Mess- oder Dokumentationsfehler → Ausschluss von der Analyse
  - pathologische Besonderheiten
- Ausreißer gerechtfertigt → Analyse 2mal durchführen, einmal mit und einmal ohne Ausreißer
  - kein Unterschied
  - Unterschied – Anwendung von statistischen Verfahren, die unempfindlich gegen Ausreißer sind

*Surrogatmerkmale*
- Diagnose von Zielgrößen kann oft nur mit großem Aufwand geschehen
- Lösung: Einsatz von Surrogatmerkmalen, die eine Funktionsstörung anzeigen und einfach zu bestimmen sind
- Beispiel: Kreatinin-Wert zum Bestimmen von Nierenversagen

*Ungenaue Definitionen*
- Beschreibung und Untersuchung von Zielgrößen, die nicht klar definiert sind
- Beispiel: „Therapieerfolg" – vollständige Heilung, Besserung der Symptome, etc.
- exakte Definition ist notwendig, um praxisrelevante Schlussfolgerungen ziehen und Vergleiche anstellen zu können.

*Falsche oder unvollständige Informationen*
- befragte Personen können unvollständige oder falsche Angaben geben (z.B. aus Scham, anderen Gründen)
- Zurückgreifen auf mangelhafte Dokumentation in Patientenakten

*Zensierte Daten*
- Beispiel: bei Überlebenszeitstudien wird die Zeit untersucht, die bis zum Eintreten eines bestimmten Ereignisses (z.B. Tod) vergeht
- Gründe, weshalb sich die Überlebenszeit nicht exakt feststellen lässt:
  - frühzeitiges Ausscheiden des Patienten aus der Studie
  - Überleben der Studie
- Zeit, die nach der Studie vergeht, ist unbekannt → zensiert
- Eliminierung aller zensierten Daten aus der Analyse → Verzerrung der Ergebnisse
  - Anwendung von Verfahren, die zensierte Daten bei der Analyse berücksichtigen (Kaplan-Meier-Methode, Logrank-Test)

# Listen und Tabellen

## Listen

- Dokumentierung aller relevanten Informationen für jede Beobachtungseinheit
- wichtig für statistische Analysen
- bei Daten, die fehlen, können die Gründe, weshalb diese nicht dokumentiert wurden, später schwer nachvollzogen werden

## Tabellen

- übersichtliche Zusammenfassung der für die statistische Analyse relevanten Daten in einer Tabelle
- Basis für alle nachfolgenden Analysemethoden und daraus resultierenden Erkenntnisse
- Tabellenzeilen – Beobachtungseinheiten mit eindeutiger ID
- Tabellenspalten – Daten eines bestimmten Merkmals
- Legende – weitere Informationen

*Fehlende Daten*

- Kennzeichnung dieser
- Vermeidung wenn möglich → Reduzierung des Stichprobenumfangs, ungenaue Ergebnisse

# HÄUFIGKEITEN

## Häufigkeiten bei diskreten Merkmalen

### Absolute und relative Häufigkeiten

*Häufigkeitsverteilung:* Verteilung, die beschreibt wie häufig die einzelnen Merkmalsausprägungen in der Stichprobe zu finden sind
*Diskrete Merkmale:* alle qualitativen sowie quantitativ-diskreten Merkmale

- Anzahl der Ausprägungen ist in der Regel wesentlich kleiner als der Stichprobenumfang (Beispiel: „Blutgruppe" hat die Ausprägungen 0, A, B und AB)

Ein diskretes Merkmal A habe k verschiedene Ausprägungen $A_1$, …, $A_k$. Die absolute Häufigkeit einer Ausprägung $A_i$ wird mit $n_i$ bezeichnet. Die Summe aller absoluten Häufigkeiten $n_i$ entspricht der Anzahl der Beobachtungseinheiten in der Stichprobe – das ist der Stichprobenumfang n:

$$\sum_{i=1}^{k} n_i = n$$

Unter der relativen Häufigkeit $h_i$ einer Ausprägung $A_i$ versteht man den Quotienten

$$h_i = \frac{n_i}{n}$$

$0 \leq h_i \leq 1$

$$\sum_{i=1}^{k} h_i = \frac{\sum_{i=1}^{k} n_i}{n} = \frac{n}{n} = 1$$

| Ausprägung | absolute Häufigkeiten | relative Häufigkeiten |
|---|---|---|
| $A_1$ = Blutgruppe 0 | $n_1 = 28$ | $h_1 = 39\%$ |
| $A_2$ = Blutgruppe A | $n_2 = 31$ | $h_2 = 44\%$ |
| $A_3$ = Blutgruppe B | $n_3 = 9$ | $h_3 = 13\%$ |
| $A_4$ = Blutgruppe AB | $n_4 = 3$ | $h_4 = 4\%$ |
| *Summe* | *n = 71* | *100%* |

### Graphische Darstellungen

*Kreisdiagramm*

- einzelne Kreissektoren stellen absolute oder relative Häufigkeit dar → Aussehen ist dasselbe
- Darstellung eignet sich nur für nominal skalierte Merkmale

*Rechteckdiagramm, Blockdiagramm*

- Rechteck ist entsprechend der einzelnen Häufigkeiten unterteilt
- Darstellung eignet sich auch für ordinal skalierte Merkmale – kleinste und größte Ausprägung sind erkennbar

*Balkendiagramm*

- eignet sich für alle diskreten Merkmale
- Längen der einzelnen Balken entsprechen $n_i$ oder $h_i$
- jede Ausprägung ergibt eine Säule
- senkrechte Anordnung: Säulendiagramm
- 1-dimensionale Striche: Stabdiagramm

*Punktediagramm*

- Darstellung einfachster Art für quantitative Merkmale
- Stichprobenwerte werden entlang einer Achse als einzelne Punkte eingetragen

# Häufigkeiten bei stetigen Merkmalen

## Prinzip der Klassenbildung

- existieren viele Ausprägungen im Vergleich zur Stichprobengröße und haben viele davon die Häufigkeit 0 → sinnvoll: Zusammenfassung von Ausprägung in Klassen
  - bei stetigen Merkmalen
  - bei einem quantitativ-diskreten Merkmal mit extrem vielen, fein abgestuften Ausprägungen
- viele schmale Klassen → Darstellung unübersichtlich, Verteilungstyp schwer erkennbar
- geringe Anzahl von breiten Klassen → hoher Informationsverlust, Verdeckung charakteristischer Eigenschaften der Verteilung
- Klassenanzahl k richtet sich nach Stichprobenumfang n
  - $k \approx \sqrt{n}$
  - n ≥ 1000, k ≈ 10 * lg n
- weniger als 3 Klassen sind nicht sinnvoll
- übersichtlich – gleiche Klassenbreite
- Vermeidung von Klassen mit Grenzen -∞ oder +∞ vermeiden
- eindeutige Klärung zu welcher Klasse ein Wert gehört (halboffene Intervalle – links offen, rechts abgeschlossen)

Beispiel

| Klassengrenzen cm | absolute H./Besetzungszahl einer Klasse $n_i$ | relative Häufigkeit $h_i$ | absolute Summenhäufigkeit $N_i$ | relative Summenhäufigkeit $H_i$ |
|---|---|---|---|---|
| (152,5;157,5) | 5 | 0,07 | 5 | 0,07 |
| (157,5;162,5) | 2 | 0,03 | 7 | 0,10 |
| (162,5;167,5) | 10 | 0,14 | 17 | 0,24 |
| (167,5;172,5) | 18 | 0,25 | 35 | 0,49 |
| (172,5;177,5) | 12 | 0,17 | 47 | 0,66 |
| (177,5;182,5) | 17 | 0,24 | 64 | 0,90 |
| (182,5;187,5) | 3 | 0,04 | 67 | 0,94 |
| (187,5;192,5) | 1 | 0,01 | 68 | 0,96 |
| (192,5;197,5) | 3 | 0,04 | 71 | 1 |

( , ) … Wert gehört nicht zum Intervall
[ , ] … Grenzwert gehört zum Intervall

F(172) = 0,49
Knapp die Hälfte der Probanden ist 172cm groß oder kleiner.

## Graphische Darstellungen

(Seite 45)
*Histogramm*

- Darstellung jeder Klasse durch ein Rechteck, Flächen sind proportional zu den jeweiligen Klassenhäufigkeiten
- x-Achse … Klassen
- y-Achse … Häufigkeiten der Klassen
- mathematische Funktion, die ein Histogramm beschreibt = empirische Dichte

$$f(x) = \begin{cases} 0 \; für x \leq a_0 \\ \frac{h_i}{a_i - a_{i-1}} \; für \; a_{i-1} < x \leq a_i (i = 1, \dots, k) \\ 0 \; für \; x > a_k \end{cases}$$

- $a_{i-1}$ und $a_i$ … untere bzw. obere Grenze der i
- k … Klassenanzahl
- Histogramm besteht aus k Rechtecken der Fläche $h_i$
- Gesamtfläche = 1

*Häufigkeitspolygon*
- x-Achse … Klassen
- in der Mitte jeder Klasse wird die Häufigkeit (y-Achse) eingetragen
- Punkte werden miteinander verbunden

*Stamm- und Blatt-Diagramm*
- Daten nach Größe geordnet
- von unten nach oben aufgetragen
- Stamm besteht aus den ersten Stellen der Stichprobenwerte
- Blätter stellen die folgenden Ziffern dar
- zur Verschaffung eines schnellen Überblicks über die Daten geeignet

Ablesen vom Diagramm
- Lage
  - Bereich, wo sich die Werte konzentrieren, häufigste/niedrigste Ausprägungen
- Streuung
  - Streuung um den Mittelwert, Ausreißer
- Form
  - symmetrisch, schief, Anzahl der Gipfel

## Empirische Verteilungsfunktion

*(kumulative) Summenhäufigkeiten $N_i$, $H_i$:* Bei quantitativen oder ordinal skalierten Merkmalen können die Häufigkeiten beginnend bei der kleinsten Ausprägung in aufsteigender Reihenfolge aufsummiert werden.

$A_1 < A_2 < \ldots < A_k$

Absolute Summenhäufigkeiten

$$N_i = \sum_{j=1}^{i} n_j \ (\text{für } i = 1, \ldots, k)$$

Relative Summenhäufigkeiten

$$H_i = \sum_{j=1}^{i} h_j \ (\text{für } i = 1, \ldots, k)$$

Die zu den einzelnen Ausprägungen gehörenden relativen Summenhäufigkeiten Hi werden durch die empireische Verteilungsfunktion F(x) mathematisch beschrieben:

$$F(x) = \begin{cases} 0 \text{ für } x < A_1 \\ H_i \text{ für } A_i \leq x < A_{i+1} \ (i = 1, \ldots, k-1) \\ 1 \text{ für } x \geq A_k \end{cases}$$

- F(x) gibt die relativen Häufigkeiten an, wie viele Werte gleich x oder kleiner als x sind.
- F(x) ist eine Treppenfunktion.
- F(x) = 0 für alle x, die kleiner als der kleinste Stichprobenwert $x_{min}$ sind
- F(x) = 1 ab dem größten Wert $x_{max}$

Eine Funktion heißt monoton wachsend, wenn für $x_1 < x_2$ gilt: $F(x_1) \leq F(x_2)$
Eine Funktion heißt streng monoton wachsend, wenn für $x_1 < x_2$ gilt: $F(x_1) < F(x_2)$

Bei fein abgestuften Ausprägungen, ist die Anzahl der Treppen zahlreich, die Stufen sind niedrig → Annäherung an eine glatte Kurve.

# 2-dimensionale Häufigkeiten

## Kontingenztafel

Wird der Zusammenhang zwischen zwei qualitativen Merkmalen betrachtet, spricht man von Assoziation oder Kontingenz.

- 2 diskrete Merkmale $A_i$ (i=1,…,k) und $B_j$ (j=1,…,l)
- Anzahl aller denkbaren Kombinationen k * l
- absolute Häufigkeiten $n_{ij}$ = Anzahl der Beobachtungseinheiten, bei denen die Ausprägungen $A_i$ und $B_j$ gemeinsam auftreten

*relative Häufigkeiten*

$$h_{ij} = \frac{n_{ij}}{n} \qquad mit\ i = 1, \dots, k \ \ und \ \ j = 1, \dots, l$$

$h_{ij}$ … zwischen 0 und 1

*Aufaddieren aller Häufigkeiten*

$$\sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} = n$$

$$\sum_{i=1}^{k} \sum_{j=1}^{l} h_{ij} = 1$$

*Randhäufigkeiten, Randsummen:* Häufigkeiten, die sich nur auf die Ausprägungen $A_i$ oder $B_j$ beziehen

*Kontingenztafel*

- genaue Informationen über und Darstellung aller Häufigkeiten
- Auflistung der Ausprägungen der beiden Merkmale in Kopf und Vorspalte
- Inneres: Felder mit den jeweiligen Häufigkeiten
  - absolute Häufigkeiten $n_{ij}$ = a, b, c, d
  - relative Häufigkeiten $h_{ij}$
  - relative Reihen- oder Spaltenhäufigkeiten
- letzte Tabellenspalte, letzte Zeile: Randsummen (Reihen- bzw. Spaltensummen)
- Vierfeldertafel: Betrachtung zweier Alternativmerkmale → vier Felder

| | **Raucher** | **Nichtraucher** | |
|---|---|---|---|
| **männlich** | a = 4<br>$h_{ij}$= (0,06)<br>(0,17)<br>(0,31) | b = 19<br>(0,27)<br>(0,83)<br>(0,33) | 23<br>(0,32) |
| **weiblich** | c = 9<br>(0,13)<br>(0,19)<br>(0,69) | d = 39<br>(0,55)<br>(0,81)<br>(0,67) | 48<br>(0,68) |
| | 13<br>(0,18) | 58<br>(0,82) | **n = 71** |

- 23 Männer, 48 Frauen → Gesamt: 71
- 4 männliche Raucher, 9 weibliche Raucher → Gesamt Raucher: 13 (18% Raucher)
- 19 männliche Nichtraucher, 39 weibliche Nichtraucher → Gesamt Nichtraucher: 58 (82% Nichtraucher)
- 17%(Reihenhäufigkeit: 4/23) der Männer und 19% der Frauchen rauchen
- Raucher sind zu 31% (Spaltenhäufigkeit: 4/13) männlich, Nichtraucher zu 33%
- OR = (4 * 39) / (19 * 9) = 0,912 (→ kein wirklicher Zusammenhang zw. Raucher und Geschlecht)
  - a/c … Anzahl der männlichen Raucher im Verhältnis zu den weiblichen Rauchern

## Beschreibung einer Assoziation

Kontingenztafel ist wenig geeignet um den Grad eines Zusammenhangs zu erfassen.

*Balkendiagramm*

- Darstellung der Zusammenhänge zweier qualitativer Merkmale
- Längen der Balken repräsentieren Häufigkeiten der Ausprägungen des ersten Merkmals (z.B. Nichtraucher, Raucher)
- Balken sind unterteilt in die Häufigkeiten der Ausprägungen des zweiten Merkmals (z.B. weiblich, männlich)

*Odds Ratio*

… Assoziationsmaß, das den Grad eines Zusammenhangs zwischen zwei Alternativmerkmalen quantifiziert

- Bildung durch Kreuzprodukt der Häufigkeiten im Innern der Vierfeldertafel
- Odds: Verhältnis aus zwei zusammen gehörenden Häufigkeiten

$$OR = \frac{ad}{bc}$$

OR = 1 … kein Zusammenhang zwischen den beiden Merkmalen

*Assoziationskoeffizient nach Yule*

- $-1 \leq Q \leq +1$
- Q = 0, falls ad = bc (→ vollkommene Unabhängigkeit)

$$Q = \frac{ad - bc}{ad + bc}$$

# BESCHREIBUNG EINES MERKMALS

## Methoden der univariaten Statistik

- zur Beschreibung charakteristischer Eigenschaften eines einzelnen Merkmals
- Methoden sind abhängig von der Art es Merkmals (Skalenniveau)
- statistische Kenngrößen (Maßzahlen) zur quantitativer Analyse eines Merkmals
  - Lagemaße
  - Streuungsmaße
  - Formmaße
- empirische Größen: Kenngrößen werden aus den Daten der Stichprobe ermittelt und dienen als Schätzwerte für die Parameter der Grundgesamtheit

## Lagemaße

= Lokalisationsmaße
… geben an, in welchem Bereich sich die Stichprobenwerte konzentrieren

### Arithmetisches Mittel

= Durchschnitt
- Addieren aller Stichprobenwerte und Division durch den Stichprobenumfang n
- Nachteil: von Ausreißern stark betroffen; gibt bei schiefen Verteilungen ein verzerrtes Bild wider
- Berechnung nur dann sinnvoll, wenn Differenz zwischen zwei Ausprägungen definiert → Voraussetzung von quantitativen Merkmalen
- vor allem bei symmetrischen, eingipfeligen Verteilungen sinnvoll
- nutzt im Gegensatz zu anderen Lagemaßen alle Informationen der Stichprobenwerte

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

### Median

= empirischer Median, Zentralwert
- Teilung der Stichprobenwerte in zwei Hälften
- Sortierung der Werte nach der Größe = Rangliste

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

  - $x_{(1)}$ … kleinster Wert $x_{min}$
  - $x_{(n)}$ … größter Wert $x_{max}$

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{für } n \text{ ungerade} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{für } n \text{ gerade} \end{cases}$$

- Median =
  - Wert aus der Urliste, falls n ungerade
  - Durchschnittswert der beiden mittleren Werte, falls n gerade
- gegenüber Ausreißern robust
- Ausreißer bewirken, dass Median und Mittelwert stark voneinander abweichen → Verteilung schief
- stimmen Mittelwert und Median in etwa überein → Verteilung symmetrisch
- Median kann berechnet werden, sobald die Hälfte der Daten bekannt sind (beim Mittelwert nicht möglich)
- Angabe des Medians ist sinnvoll
  - bei ordinal skalierten Daten
  - bei quantitativen Merkmalen, die schief verteilt sind
  - bei Verdacht auf Ausreißer
  - bei zensierten Daten

## Quartile und Quantile

*Quartile*

- teilen die Stichprobe in vier Viertel
- unteres / erstes Quartil $Q_1$ … 25% der Stichprobenwerte sind kleiner oder gleich $Q_1$
- mittleres /zweites Quartil = Median
- oberes / drittes Quartil $Q_3$ … 75% der Werte sind kleiner oder gleich $Q_2$

*Quantile, Fraktile*

- Verfeinerung der Häufigkeitsverteilung
- α … reelle Zahl mit 0 < α < 1
- Ermittlung der Rangzahl $k = \alpha * n$ (auf- bzw. abrunden wenn nötig)
- falls $\alpha * n$ keine ganze Zahl, sei k die direkt auf $\alpha * n$ folgende ganze Zahl und
$$\tilde{x}_\alpha = x_{(k)}$$
- falls $\alpha * n$ eine ganze Zahl ist, sei $k = \alpha * n$ und
$$\tilde{x}_\alpha = \frac{x_{(k)} + x_{(k+1)}}{2}$$

*Dezile*

- $\alpha = 0{,}1\ ;\ 0{,}2\ ;\ \ldots\ ;\ 0{,}9$

*Perzentile*

- $\alpha = 0{,}01\ ;\ \ldots\ ;\ 0{,}99$

*Beispiele*

1. Quartil: $\alpha * n = 0{,}25 * 48 = 12$; also k = 12
$$Q_1 = \frac{x_{(12)} + x_{(13)}}{2} = 165cm$$
9. Dezil: : $\alpha * n = 0{,}90 * 48 = 43{,}2$; also k = 44
$$\tilde{x}_{0{,}90} = x_{(44)} = 178cm$$

## Modus

= Modalwert M, Dichtemittel D

- Ausprägung mit der größten Häufigkeit
- kann bei allen Skalenniveaus ermittelt werden
- bei Klassen: Angabe der modalen Klasse = Klasse mit der größten Besetzungszahl; deren Mitte = Modus
- Verteilung kann sein
  - eingipfelig (unimodal)
  - zweigipfelig (bimodal)
  - mehrgipfelig (multimodal)
- U-förmige Verteilungen
  - 2 Modalwerte an ihren Rändern
  - Tiefpunkt in der Mitte
  - Mittelwert repräsentiert einen atypischen Wert
- Angabe von Modalwerten
  - bei nominalen Merkmalen, da andere Lagemaße bei diesem Skalenniveau nicht zulässig sind
  - bei ordinalen und quantitativen Merkmalen, bei „ausgeprägtem Gipfel" (bei hohem Stichprobenumfang)
  - bei einer U-Verteilung
- Angabe eines Modalwertes ist nicht empfehlenswert
  - bei Alternativmerkmalen (z.B. Geschlecht)
  - kein „ausgeprägter Gipfel"

## Minimum und Maximum

- extremste Werte eines ordinal oder metrisch skalierten Merkmals
- grober Überblick über die Streuung der Daten
- Plausibilitätsprüfung der Daten (Fehler, die bei der Dateneingabe entstehen können)

### Geometrisches Mittel

- Verwendung bei relativen Änderungen, bei denen sich der Unterschied zweier Merkmalswerte sinnvoller durch einen Quotienten als durch eine Differenz beschreiben lässt
- $x_i$ … relative Änderungen, $x_i > 0$, dimensionslos

$$\bar{x}_G = \sqrt[n]{x_1 * \ldots * x_n}$$

### Harmonisches Mittel

- Beobachtungswerte $x_i$ sind Verhältniszahlen (also Quotienten), die sich nur in ihren Nennern unterscheiden
- z.B. Durchschnittsgeschwindigkeit, durchschnittliche Dichte

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

## Streuungsmaße

= Dispersionsmaße
… geben Auskunft über die Variabilität der Stichprobenwerte

### Varianz und Standardabweichung

- bei quantitativen Merkmalen ist der Mittelwert das am häufigsten benutzte Lagemaß

*Varianz*

- Streuungsmaß, das die Abweichungen der Stichprobenwerte vom Mittelwert quantifiziert
- mittlere quadratische Abweichung der Daten vom Mittelwert
- Varianz der Stichprobe: Summe der Abstandsquadrate dividiert durch n
- empirische Varianz:

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

*Standardabweichung*

- Streuungsmaß mit gleicher Dimension wie die Stichprobendaten
- Maß für die Homogenität bzw. Heterogenität der Stichprobe
- nur bei quantitativen Merkmalen sinnvoll
- wenn alle Werte identisch sind und Stichprobe vollkommen homogen ist → s = 0
- Charakterisierung von quantitativen, annähernd symmetrisch verteilten Daten durch $\bar{x} \pm s$ unter Angabe des Stichprobenumfangs n

$$s = \sqrt{Var}$$

Normalverteilung

- 2/3 aller Werte liegen zwischen $\bar{x} \pm s$
- 95% aller Werte liegen zwischen den Grenzen $\bar{x} \pm 2s$

symmetrische, eingipfelige Verteilungen

- 8/9 aller Werte sind innerhalb der Grenzen $\bar{x} \pm 2s$
- 95% aller Werte sind im Bereich $\bar{x} \pm 3s$

alle Verteilungen (auch schiefe)

- mindestens ¾ aller Werte liegen im Intervall $\bar{x} \pm 2s$
- 8/9 aller Daten liegen im Bereich $\bar{x} \pm 3s$

### Variationskoeffizient

$$V = \frac{s}{\bar{x}} \qquad (falls\ \bar{x} > 0)$$

- dimensionslos
- nur für verhältnisskalierte Merkmale geeignet
- Maximum beträgt $\sqrt{n}$
- relativer Variationskoeffizient liegt zwischen 0 und 1

$$V_r = \frac{\frac{s}{\bar{x}}}{\sqrt{n}}$$

## Spannweite

= Variationsbreite

- R = 0, wenn alle Stichprobenwerte identisch sind
- lediglich Berücksichtigung der beiden extremsten Werte → stark von Ausreißern beeinflusst
- hauptsächliche Verwendung bei diskreten Merkmalen mit wenigen Ausprägungen

$$R = x_{max} - x_{min} = x_{(n)} - x_{(1)}$$

## Weitere Streuungsmaße

*Dezilabstand*

- weniger empfindlich als die Spannweite
- erhält man, wenn man an beiden Rändern der Verteilung jeweils 10% abschneidet und die Länge dieses Interdezilbereichs berechnet

$$I_{80} = \tilde{x}_{0,90} - \tilde{x}_{0,10}$$

*Quartilabstand*

- Länge des Interquartilbereichs [$Q_1$, $Q_3$], der die mittleren 50% der Stichprobenwerte enthält

$$I_{50} = Q_3 - Q_1 = \tilde{x}_{0,75} - \tilde{x}_{0,25}$$

*Mittlere Abweichung vom Median*

- Verwendung bei ordinal skalierten oder schief verteilten Daten

$$MA_{\tilde{x}} = \frac{\sum_{i=1}^{n} |x_i - \tilde{x}|}{n}$$

*Variation Ratio*

- Streuungsmaß für nominal skalierte Merkmale
- relative Häufigkeit der Beobachtungen, die nicht in die modale Kategorie fallen
- $h_{modal}$ … relative Häufigkeit des Modalwertes
- VR = 0 … alle Beobachtungen sind identisch
- $0 \leq VR \leq 1$
- VR ≈ 1 … je größer die Anzahl der Merkmalsausprägungen, desto weniger unterscheiden sich die Häufigkeiten der einzelnen Kategorien

$$VR = 1 - h_{modal}$$

## Lagemaße und Streuungsmaße müssen zusammen passen

- *Mittelwert und Standardabweichung*
  bei symmetrisch verteilten Daten
- *Median und Quartilabstand* (Dezilabstand oder mittlere Abweichung vom Median)
  bei schief verteilten Daten oder Verdacht auf Ausreißer
- *Modus und Spannweite*
  bei diskreten Merkmalen mit wenigen Ausprägungen
- *Modus und Variation Ratio*
  bei nominal skalierten Merkmalen

# Formmaße

… quantitative Beschreibung der Verteilungsform

## Schiefe

- kennzeichnet die Symmetrie bzw. Asymmetrie einer Verteilung
- dimensionslos
- kann sowohl positive als auch negative Werte annehmen

- bei symmetrischen Verteilungen ist $g_1$=0 und $\bar{x} = \tilde{x} = D$
- bei rechtsschiefen Verteilungen ist $g_1 > 0$ und $\bar{x} > \tilde{x} > D$
- bei linksschiefen Verteilungen ist $g_1 < 0$ und $\bar{x} < \tilde{x} < D$

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

*Formeln von Pearson für unimodale Verteilungen*

- schiefe Verteilung, wenn $g_1$ stark von 0 abweicht und Stichprobenumfang groß (mind. n ≥ 100)

$$g_1 \approx \frac{3\,(\bar{x} - \tilde{x})}{s}$$

$$g_1 \approx \frac{\bar{x} - D}{s}$$

## Wölbung

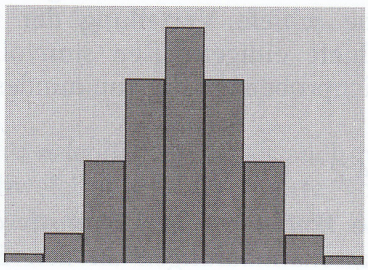
= Kurtosis, Exzess

- Beschreibung der Massenanhäufungen an den Enden bzw. um den Mittelwert der Verteilung
- nur größere Abweichungen von 0 lassen den Schluss zu, dass die Daten nicht normalverteilt sind

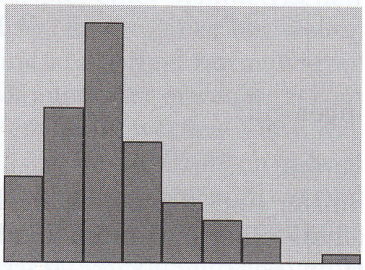
$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^4}$$

Für symmetrische, eingipfelige Verteilungen gilt

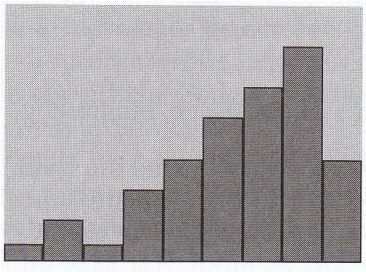
- $g_2$=0
  normalverteilte Daten
- g2 > 0
  Verteilung ist schmaler und steilgipfeliger als Glockenkurve der Normalverteilung mit gleicher Standardabweichung, Maximum ist größer, Werte häufen sich in der Umgebung des Mittelwerts und an den Ausläufern
- g2 > 0
  Verteilung ist flacher als die Glockenkurve der Normalverteilung, Maximum ist kleiner



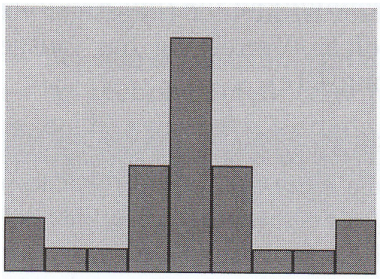
**a.** symmetrische Verteilung (Schiefe=0)



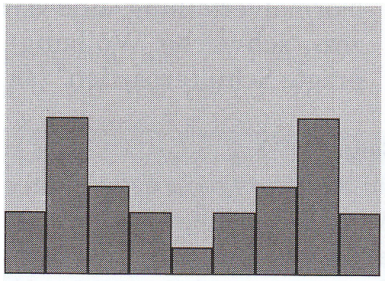
**b.** rechtsschiefe Verteilung (Schiefe>0)



**c.** linksschiefe Verteilung (Schiefe<0)



**d.** symmetrische Verteilung (Wölbung>0)



**e.** symmetrische Verteilung (Wölbung<0)

| Skala | Lagemaße | Streuungsmaße | Formmaße | graphische Darstellungen |
|---|---|---|---|---|
| *Nominalskala* | Modus | Variation Ratio | -- | Kreisdiagramm<br>Rechteckdiagramm<br>Balkendiagramm |
| *Ordinalskala* | Modus<br>Median<br>Quartile<br>Quantile | Varation Ratio<br>Spannweite<br>Quartilsabstand<br>Interdezilabstand | -- | Rechteckdiagramm<br>Balkendiagramm |
| *Intervallskala* | Modus<br>Median<br>Quartile<br>Quantile<br>Mittelwert | Spannweite<br>Quartilsabstand<br>Interdezilabstand<br>Standardabweichung | Schiefe<br><br>*symmetrische Verteilungen:*<br>Wölbung | *diskrete Daten:*<br>Balkendiagramm<br>Punktediagramm<br><br>*stetige Daten:*<br>Histogramm<br>Häufigkeitspolygon<br>Stamm- und Blatt-Diagramm |
| *Verhältnisskala* | | Spannweite<br>Quartilsabstand<br>Interdezilabstand<br>Standardabweichung<br>Variationskoeffizient | | |

## Vergleich mehrerer Stichproben

### Beispiele für Gruppenvergleiche

- Untersuchen und Vergleichen von zwei oder mehreren Stichproben
- Ziel der statistischen Analyse: Unterschied zwischen diesen Gruppen nachzuweisen

Beispiele

- klinischkontrollierte Studie
  mehrere Therapieformen
- Fall-Kontroll-Studie
  eine Gruppe von erkrankten Patienten wird zur Klärung eines ätiologischen Faktors einer Gruppe von gesunden Personen gegenübergestellt
- Kohortenstudie
  Personen, die einem bestimmten Risiko ausgesetzt sind, und Personen, die diesem Risiko nicht ausgesetzt sind, werden eine zeitlang gemeinsam beobachtet

### Graphische Darstellungen

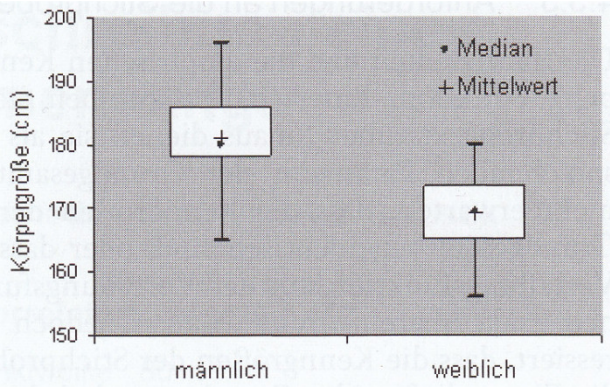
- Anfertigung eines Diagramms für jede Stichprobe und Vergleich dieser

<u>Box-and-Whiskers-Plots</u>

- bei quantitativen Merkmalen
- Repräsentation jeder Stichprobe durch eine rechteckige Box
- unten und oben/rechts und links vom 1. und 3. Quartil begrenzt
- innerhalb der Box ist der Median gekennzeichnet (für Mittelwert auch möglich)
- von Box ausgehende Striche („whiskers") zeigen die Lage des Minimums und Maximums
- Hinweise über

- Lagemaße (Mittelwerte, Mediane, Quartile, Maxima, Minima)
- Streuungsmaße (Spannweite, Quartilsabstand)

- Informationen über die Schiefe
  - je weiter der Mittelwert und der Median voneinander entfernt sind, desto schiefer ist die Verteilung



Box-and-Whisker-Plots

## Anforderungen an die Stichproben

Häufigkeiten und empirischen Kenngrößen

- beschreiben die Charakteristika der Stichprobe
- dienen als Schätzwerte für die entsprechenden Parameter der Grundgesamtheit
- sollen in brauchbarer Weise die Eigenschaften der Grundgesamtheit beschreiben
  - Stichprobe soll repräsentativ für die jeweilige Grundgesamtheit sein
  - Stichprobe und Stichprobenumfang muss hinreichend groß sein
- eine große Stichprobe ermöglicht bessere Schätzungen →
  - sollte jedoch nicht größer sein als nötig
  - Festlegung der Größe vor der Datenerhebung

# BESCHREIBUNG EINES ZUSAMMENHANGS

## Methoden der bivariaten Statistik

- Beschreibung des Zusammenhangs zweier Merkmale
- z.B. Körpergewicht wird von der Größe mitbestimmt, Auftreten bestimmter Krankheiten ist abhängig von diversen Risikofaktoren
- funktionale Zusammenhänge
  - aus der Mathematik: U = 2πr
  - Berechnung einer Größe mittels math. Gleichung
- stochastische Zusammenhänge
  - in der Medizin
- Aufgaben der bivariaten Statistik
  - Zusammenhang zwischen zwei Merkmalen aufzuzeigen und zu beschreiben
  - Methoden sind vom Skalenniveau der beiden Merkmale abhängig
  - Zusammenhang zwischen 2 quantitativen Merkmalen: Korrelationsanalyse, Regressionsanalyse

## Korrelationsanalyse

### Punktwolke

- graphische Darstellung im Koordinatensystem – jeder Beobachtungseinheit wird ein Wertepaar ($x_i$, $y_i$) zugeordnet → Erhalt einer Punktwolke (Punkteschar)
  - x sollte das unabhängige und y das abhängige Merkmal sein (ansonsten unterscheiden x und y lediglich die Merkmale)
- Punktwolke lässt zwei charakteristische Eigenschaften eines Zusammenhangs erkennen
  - Stärke des Zusammenhangs
    je dichter die Punkte beieinander liegen, desto stärker ist der Zusammenhang
  - Art des Zusammenhangs
    - Regressionsanalyse: math. Funktion finden, die den Zusammenhang beschreibt
    - Voraussetzung: metrische Skalenniveaus der beider Merkmale

*Regressionsgerade*

- Gerade beschreibt den Zusammenhang der beiden Merkmale → linearer Zusammenhang
- Steigung der Regressionsgerade → gleichsinniger Zusammenhang
  - z.B. je größer, desto mehr Gewicht hat man
- gegensinniger Zusammenhang → negativer Anstieg

### Voraussetzungen der Korrelationsanalyse

- Berechnung des Korrelationskoeffizienten nach Pearson zur Quantifizierung der Stärke eines Zusammenhangs
- Voraussetzungen
  - beide Merkmale x und y sind quantitativ
  - Zusammenhang ist annähernd linear
  - Beobachtungseinheiten sind unabhängig voneinander
- linearer Zusammenhang
  - Gerade kann durch Punktwolke gezogen werden, an der die Punkte ellipsenförmig liegen
- Abhängigkeit der Zusammenhänge
  - z.B. Geschwister werden in die Beobachtungseinheiten aufgenommen, Mehrfacherfassung
- beide Merkmale müssen bivariat normalverteilt sein, wenn die empirischen Maßzahlen der Stichprobe als Schätzer für die entsprechenden Parameter der Grundgesamtheit dienen sollen

### Kovarianz

- Maß für das „Miteinander-Variieren" zweiter quantitativer Merkmale
- basiert auf den Mittelwerten $\bar{x}$ und $\bar{y}$
- anhand des Vorzeichens kann lediglich ermittelt werden, ob ein gleichsinniger oder gegensinniger Zusammenhang der beiden Merkmale besteht

- Korrelationskoeffizient und Parameter der Regressionsgerade bauen darauf auf

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

- Erfassung des durchschnittlichen Produkts der Abweichungen $(x_i - \bar{x})^2$ und $(y_i - \bar{y})^2$
- Erhalt eines optimalen Schätzwerts für die Kovarianz der Grundgesamtheit durch Division durch n-1

- positive Kovarianz $s_{xy} > 0$
  - gleichsinniger Zusammenhang
  - wenn Messwerte beider Beobachtungseinheiten größer oder kleiner sind als die jeweiligen Mittelwerte
- negative Kovarianz $s_{xy} < 0$
  - gegensinniger Zusammenhang
  - unterschiedliche Vorzeichen von $(x_i - \bar{x})$ und $(y_i - \bar{y})$
- Kovarianz $s_{xy} \approx 0$
  - kein linearer Zusammenhang
  - nahe beieinander liegende x-Werte korrelieren sowohl mit positiven als auch mit negativen Abweichungen, sodass die Produkte $(x_i - \bar{x})$ $(y_i - \bar{y})$ ausgeglichen werden und sich in ihrer Summe ein Wert nahe bei 0 ergibt

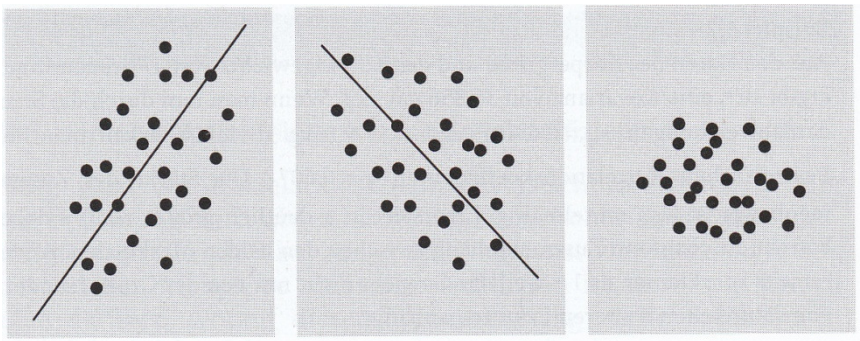


**Abb. 5.2a** gleichsinniger Zusammenhang, positive Kovarianz

**Abb. 5.2b** gegensinniger Zusammenhang, negative Kovarianz

**Abb. 5.2c** kein linearer Zusammenhang, Kovarianz ≈ 0

## Korrelationskoeffizient nach Pearson

- Pearson'sche Korrelationskoeffizient, Produkt-Moment-Korrelationskoeffizient
- normiertes Maß zur Quantifizierung eines linearen Zusammenhangs
- Division der Kovarianz $s_{xy}$ durch die beiden Standardabweichungen $s_x$ und $s_y$
- Werte zwischen -1 und +1
- dimensionslos

$$r = \frac{s_{xy}}{s_x * s_y}$$

- $r > 0$
  - gleichsinniger Zusammenhang
- $r < 0$
  - gegensinniger Zusammenhang

*Betrag von r*

- je näher r bei 0 liegt,
  desto schwächer ist der Zusammenhang, desto weiter streut die Punktwolke um die Gerade
- je näher r bei 1 liegt,
  desto stärker ist der Zusammenhang und desto dichter liegen die Punkte $(x_i, y_i)$ an der Regressionsgeraden

- Extremfälle r=1, r=-1 ergeben sich bei einem funktionalen Zusammenhang, der durch eine lineare Gleichung der Form y=a+bx exakt beschrieben werden kann; alle Punkte ($x_i$, $y_i$) liegen auf der Regressionsgeraden

### Interpretation eines Korrelationskoeffizienten

- empirischer Koeffizient, dessen Betrag größer als 0 ist → Zusammenhang aufgrund der Stichprobe ist nicht auszuschließen
  - keine Aussage worauf dieser Zusammenhang zurückzuführen ist
  - keine Aussage welche Schlussfolgerungen gezogen werden können

*Schein/Nonsens-Korrelationen*

- Formale Korrelation
  zwei relative Anteile werden miteinander in Beziehung gesetzt und addieren sich zu 100%;
  z.B. Eiweiß und Fett in Nahrungsmitteln → Summe 100%

- Selektionskorrelation
  Stichprobe muss die gesamte Variationsbreite der zu untersuchenden Merkmale repräsentieren;
  z.B. Selektion von Risikopatienten einer Spezialklinik (Studie, ob Raucherinnen Geburtsgewicht ihrer Babies beeinflussen)

- Korrelation durch Ausreißer
  Ausreißer, Punkt der weit vom Punkteschwarm entfernt ist

- Inhomogenitätskorrelation
  für zwei inhomogene Gruppen wird ein gemeinsamer Korrelationskoeffizient berechnet, die isoliert betrachtet keinen Zusammenhang offenbaren;
  graphische Darstellung: zwei Punktwolken, die sich nicht oder nur wenig überlappen;
  z.B. Schuhgrößen und Gehälter → Korrelationskoeffizient ist größer 0, weil Männer größere Schuhe haben und meist mehr verdienen als Frauen

- Gemeinsamkeitskorrelation
  zwei Merkmale werden durch ein drittes beeinflusst;
  z.B. Zusammenhang Entwicklung des Storchenbestands und Geburtenrate → Erzeugung der dritten Größe, der allgemeinen zeitlichen Tendenz, die beide Merkmale beeinflusst → Vortäuschung einer Nonsens-Korrelation

*Vermeidung von Fehlinterpretationen*

- Theoretische Herleitung
  - Begründung des zu quantifizierenden Zusammenhangs
  - Erarbeiten des theoretischen Hintergrunds

- Erstellen der Punktwolke
  - Beurteilung, ob der Zusammenhang linear ist
  - Aufdeckung von
    Ausreißern
    inhomogenen Gruppen

- Überprüfen der Voraussetzungen

- Interpretation
  Überprüfung folgender Möglichkeiten
  - x beeinflusst y
  - y beeinflusst x
  - x und y bedingen sich gegenseitig
  - beide Merkmale werden durch eine dritte Größe beeinflusst
  - Zusammenhang kam zufällig zustande

# Regressionsanalyse

## Herleitung der Regressionsgeraden

- Anwendung der Regressionsanalyse für
  - Ursachen- und Wirkungsanalysen
  - Zeitreihenanalysen
- Multiple Regressionsanalyse
  - y-Merkmal wird von mehreren x-Variablen bestimmt
- Aufgabe
  - Herleitung einer mathematischen Gleichung, welche die Art des Zusammenhangs zwischen zwei quantitativen Merkmalen optimal beschreibt (→ Prognose von y anhand von x)
  - Finden einer Geraden, die die Punktwolke optimal repräsentiert → Regressionsgerade
- wichtig: Klärung, welches der beiden Merkmale das abhängige (y) bzw. unabhängige (x) ist
  - x-Merkmal: dasjenige, das einfacher oder billiger ist, oder früher erfasst werden kann
- einfachste Form: Beschreibung des Zusammenhangs durch eine Gerade
  - vorab: Klärung, ob zu beschreibende Zusammenhang annähernd linear
  - es kann keine Gerade geben, auf der alle Punkte liegen (y wird nicht nur von x beeinflusst sondern auch von anderen Faktoren)
- Regressionsgerade
  - durchschnittliches Abstandsquadrat der Beobachtungspunkte von der Geraden muss minimal sein
  - eindeutig durch Steigung bestimmt

*Steigung der Regressionsgeraden*

$$b = \frac{s_{xy}}{{s_x}^2}$$

$s_{xy}$ … Kovarianz
$s_x^2$ … Varianz der x-Werte
b … Regressionskoeffizient

Regressionskoeffizient
- Variationsbereich zwischen $-\infty$ und $+\infty$
- Vorzeichen von b stimmen mit Korrelationskoeffizienten r überein
  - gleichsinniger Zusammenhang → Steigung der Regressionsgeraden positiv
  - gegensinniger Zusammenhang → Steigung der Regressionsgeraden negativ

*y-Achsenabschnitt der Regressionsgeraden*

$$a = \bar{y} - b\bar{x}$$

*Prognostizieren eines Wertes $y_i$ bei Vorliegen von a und b für $x_i$*
- Mittelwert der berechneten $\hat{y}_i$ Werte = Mittelwert der beobachteten $y_i$-Werte (also $\bar{y}$)
- Punkt ($\bar{x}$, $\bar{y}$) liegt auf der Regressionsgeraden → Schwerpunkt der Punktwolke
- Zusammenhang funktional (r = ±1) → alle Punkte liegen auf der Regressionsgeraden

$$\hat{y}_i = a + bx_i = \bar{y} + \frac{s_{xy}}{s_x^2}(x_i - \bar{x})$$

## Regression 1. Art und 2. Art

- Unterscheidung anhand der Eigenschaften der x-Variablen
- nützlich: Bestimmung der Regressionsgleichung
- Korrelationskoeffizient nach Pearson als Maß der Stärke des Zusammenhangs nur für 2. Art sinnvoll

- Regression 1. Art
  - Ausprägungen der x-Variablen explizit vorgegeben
  - zu jedem x-Wert existieren mehrere, zufällig bedingte y-Werte
  - z.B. Dosis eines Medikaments und dessen Wirkung

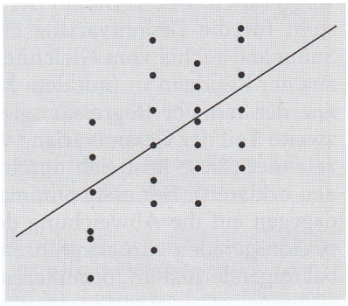


**Abb. 5.4**
Regression 1. Art

- Regression 2. Art
  - beide Merkmale stellen Zufallsvariablen dar
  - z.B. Gewicht und Körpergröße

## Bestimmtheitsmaß

- Problem der Regressionsanalyse: Verlässlichkeit der Schätzung
- meist wird prognostizierte ^$y_i$-Wert zu einem Messwert $x_i$ vom Beobachtungswert $y_i$ abweichen
- Maß dieser Abweichung: Residuum

$$e_i = y_i - \hat{y}_i$$

*Beurteilung der Schätzung durch die Regressionsgerade*
(Residuen gleichen sich gegenseitig aus → Abweichungsquadrate ei²)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Gesamtvarianz = Residualvarianz + erklärte Varianz

- Gesamtvarianz der Beobachtungswerte $y_i$
- Residualvarianz: Varianz der Residuen ei (mit Mittelwert 0)
- erklärte Varianz: Varianz der mit der Regressionsgleichung berechneten Werte ^$y_i$

Eine gute Schätzung ergibt sich, wenn
- Residualvarianz möglichst klein
- erklärte Varianz möglichst groß

*Bestimmung der Güte des Regressionsmodells*
- $r^2$ … Bestimmungsmaß, Determinationskoeffizient
- Wertebereich zwischen 0 und 1
- Extremfall $r^2=1$ → Residualvarianz=0

$$r^2 = \frac{{s_{\hat{y}}}^2}{{s_y}^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{erklärte\ Varianz}{Gesamtvarianz}$$

## Nicht-lineare Regression

- z.B. exponentielle oder logarithmische Zusammenhänge
- graphische Darstellung der Wertepaare als Punktwolke
- Finden der Art des Zusammenhangs und der allgemeinen Regressionsgleichung mit den Parametern a und b
  - Transformierung einer nichtlinearen Regressionsgleichung in eine lineare
    z.B. $y = a * e^{bx}$ → $\ln y = \ln a + bx$
  - Methode der kleinsten Quadrate
    z.B. $f(x) = a * e^{bx}$ → Ableitung von $\square(y_i - a * e^{bxi})^2$ nach a und b bilden und gleich 0 setzen

# Weitere Techniken

## Korrelationskoeffizient nach Spearman

- Bedingungen des Korrelationskoeffizienten nach Pearson
  - quantitative Merkmale
  - Zusammenhang muss annähernd linear sein
- Korrelationskoeffizient nach Spearman (auch Rangkorrelation)
  - Voraussetzungen sind schwächer als beim Korrelationskoeffizienten nach Pearson
  - Maß für die Stärke eines monotonen Zusammenhangs
  - basiert auf den Rangzahlen der Beobachtungswerte ($x_i$, $y_i$)
  - Sortierung aller x-Werte wie auch y-Werte und Versehen von Rangzahlen
  - kleinster Wert erhält den Rang 1, der größte den Rang n
  - Ermittlung von mittleren Rangzahlen (Addierung der Ausprägungen und Division durch deren Anzahl), wenn mehrere Ausprägungen übereinstimmen → verbundene Ränge
  - Differenz der beiden Rangzahlen von x und y → $d_i$

*Spearman'sche Korrelationskoeffizient*

$$r_s = 1 - \frac{6 * \sum_{i=1}^{n} {d_i}^2}{n * (n^2 - 1)}$$

- Bereich von $r_s$ zwischen -1 und +1
- maximaler Betrag $r_s = 1$ → Zusammenhang streng monoton
- positives Vorzeichen - gleichsinnigen Zusammenhang
- negatives Vorzeichen - gegensinnigen Zusammenhang
- $r_s = 0$ → kein monotoner Zusammenhang nachweisbar

Rangkorrelation nach Spearman eignet sich für folgende Konstellationen
- beide Merkmale sind ordinal skaliert
- Merkmal ist metrisch, anderes ordinal skaliert
- beide Merkmale sind quantitativ; der Zusammenhang ist monoton, aber nicht linear

## Zusammenhang zwischen einem quantitativen und einem Alternativmerkmal

- punktbiseriale Korrelation $r_{pb}$
  - z.B. Zusammenhang zwischen Gewicht und Geschlecht
  - Alternativmerkmale erhalten die Werte 0 oder 1
  - Berechnung des Korrelationskoeffizienten nach Pearson dadurch möglich

## Zusammenhang zwischen qualitativen Merkmalen

Je höher das Skalenniveau der zugrunde liegenden Merkmale ist, desto präziser lassen sich die Stärke und die Art eines Zusammenhangs beschreiben.