

## Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

6. Einheit

- **Beginn bivariate Datenanalyse: Kreuztabelle**

- Kenntnis der Funktionsweise und Interpretation von Kreuztabellen

## Hintergrund:

- An einer Beobachtungseinheit werden i.d.R. mehrere Merkmale erfasst
- Quantitative sozialwissenschaftliche Analyse ist nicht nur an der Verteilung einzelner Merkmale bzw. Variablen interessiert
- Ziel: Zusammenhänge und Beziehungen zwischen Merkmalen untersuchen, um Hypothesen zu überprüfen

Auch „Kontingenztafel“

- Werkzeug der deskriptiven Statistik
- 2 Merkmale werden in der (absoluten und relativen) Häufigkeit ihres gemeinsamen Auftretens dargestellt

## Voraussetzung:

- Nominales bzw. ordinales Skalenniveau
- Metrische Daten gruppiert genutzt werden (bspw. Altersgruppen, Einkommensgruppen)

→ Kreuztabellen umfassen formal k-Zeilen und l-Spalten

Aber: Nicht zu viele Ausprägungen, da sonst unübersichtlich

- Kreuztabellen erlauben erste empirische Aussagen zum Verhältnis zweier Merkmale:
- gibt es Zusammenhänge oder sind die Merkmale „statistisch unabhängig“ voneinander?

Beispiele:

- Haben Raucher häufiger schwere Corona-Krankheitsverläufe als Nichtraucher?
- Gehen höher Gebildete eher zur Wahlurne als Niedriggebildete?
- Wählen Ostdeutsche häufiger die AfD als Westdeutsche?

## Abendliche Bibliotheksnutzung und Studiengang, Befragung, Urliste mit 9 Studierenden aus 100 Befragten

Befragten-ID	Studiengang	Nutzung am Abend
1	BA	Nein
2	MA	Ja
3	MA	Nein
4	BA	Nein
5	BA	Ja
6	MA	Ja
7	MA	Ja
8	BA	Nein
9	MA	Ja



**4 Kombinationen möglich:**

**1) BA + Nutzung abends: I**

**2) BA – Nutzung abends: III**

**3) MA + Nutzung abends: IV**

**4) MA – Nutzung abends: I**

Befragten-ID	Studiengang	Nutzung am Abend
1	BA	Nein
2	MA	Ja
3	MA	Nein
4	BA	Nein
5	BA	Ja
6	MA	Ja
7	MA	Ja
8	BA	Nein
9	MA	Ja

**4 Möglichkeiten → 2x2-Tabelle (Vierfeldertafel) als einfachster Form der Kreuztabelle**

- **Spalte: Studiengang**
- **Zeile: Abend-Nutzung Ja/Nein**

Studiengang	BA	MA	Gesamt
Nutzung			
Ja	1	4	5
Nein	3	1	4
Gesamt	4	5	9

Studiengang	BA	MA		Gesamt	
Nutzung					
Ja	1	4		5	
Nein	3	1		4	
Gesamt	4	5		9	

- **Randhäufigkeiten:** rechter und unterer „Rand“ der Kreuztabelle  
Diese Informationen sind allgemein deskriptiver Natur und hätten wir auch durch univariate Häufigkeitsauszählungen herausbekommen

## 2x2-Tabelle, Vierfeldertafel

Studiengang	BA	MA	Gesamt
Nutzung			
Ja	1	4	5
Nein	3	1	4
Gesamt	4	5	9

- Bedingte (absolute) Häufigkeiten in den übrigen Feldern → Berechnung der relativen prozentualen Häufigkeiten, um die Zellen besser miteinander vergleichen zu können
- 3 Möglichkeiten zur Prozentuierung: 1) Gesamtprozente, 2) Zeilenprozente, 3) Spaltenprozente

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten,  $n=100$

**Frage:** Wieviel Prozent der Befragten sind im BA-Studiengang eingeschrieben und nutzen das Abendangebot?

→ Ermittlung der **Gesamtprozente**: bedingter Anteil der Zelle wird im Hinblick auf alle Beobachtungseinheiten berechnet

Studiengang	BA	MA	Gesamt
Nutzung			
Ja	13 $13/100=13\%$	43 $43/100=43\%$	56 $56/100=56\%$
Nein	17	27	44
Gesamt	30 $30/100=30\%$	70	100

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten, n=100

Frage: Wieviel Prozent der Befragten sind im BA-Studiengang eingeschrieben und nutzen das Abendangebot?

Studiengang	BA	MA	Gesamt
Nutzung			
Ja	13 13/100=13%	43 43/100=43%	56 56/100=56%
Nein	17	27	44
Gesamt	30 30/100=30%	70	100

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten, n=100

**Frage:** Wieviel Prozent der abendlichen Nutzer sind im BA-Studiengang eingeschrieben?

→ **Ermittlung der Zeilenprozente:** bedingter Anteil der Zelle wird im Hinblick auf die jeweilige Zeile berechnet (Achtung: im Beispiel gerundet)

Studiengang	BA	MA	Gesamt
Nutzung			
Ja	13 13/56=23%	43 43/56=77%	56 100%
Nein	17	27	44
Gesamt	30 30/100=30%	70	100

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten,  $n=100$

**Frage:** Wieviel Prozent der abendlichen Nutzer sind im BA-Studiengang eingeschrieben?

→ **Ermittlung der Zeilenprozente:** bedingter Anteil der Zelle wird im Hinblick auf die jeweilige Zeile berechnet (Achtung: im Beispiel gerundet)

Studiengang Nutzung	BA	MA	Gesamt
Ja	13 $13/56=23\%$	43 $43/56=77\%$	56 100%
Nein	17	27	44
Gesamt	30 $30/100=30\%$	70	100



Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten, n=100

**Frage:** Wieviel Prozent der BA-Studierenden nutzen das Abendangebot?

→ Ermittlung der **Spaltenprozente**: bedingter Anteil der Zelle wird im Hinblick auf die jeweilige Spalte berechnet

Studiengang	BA	MA	Gesamt
Nutzung			
Ja	13 13/30=43%	43 43/70=61%	56 56/100=56%
Nein	17 17/30=57%	27	44
Gesamt	30 100%	70	100

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten, n=100

**Frage:** Wieviel Prozent der BA-Studierenden nutzen das Abendangebot?

→ Ermittlung der **Spaltenprozente**: bedingter Anteil der Zelle wird im Hinblick auf die jeweilige Spalte berechnet

Studiengang	BA	MA	Gesamt
Nutzung			
Ja	13 13/30=43%	43 43/70=61%	56 56/100=56%
Nein	17 17/30=57%	27	44
Gesamt	30 100%	70	100

Sinnvolle und konventionelle Erstellung:

- **Spalte:** „unabhängige“ Variable, **Zeile:** „abhängige“ Variable
- Als Basis der Prozentuierung dabei die unabhängige Variable wählen und interpretieren: Spaltenprozente

- **Aussagen über Merkmalszusammenhänge – meistens:  
Beziehung zwischen unabhängiger/n und abhängiger Variablen**

*„Wenn Eltern über eine hohe Bildung verfügen, dann haben auch die Kinder einen hohen Bildungsabschluss“*

## Abhängige Variable (aV)

- „Das zu erklärende“,
- Beispiel: Höhe des Bildungsabschlusses einer Person
- („Y“)

## Unabhängige Variable (uV)

- (mögliche) Erklärungsfaktoren, z.B. Bildung der Eltern, Intelligenz, etc.
- („X“)

## Sinnvolle und konventionelle Erstellung:

- **Spalte:** „unabhängige“ Variable, **Zeile:** „abhängige“ Variable
- Als Basis der Prozentuierung dabei die unabhängige Variable wählen und interpretieren: Spaltenprozente

Grundlegende Idee bei der Überprüfung der „Unabhängigkeit“ von Variablen:

- Bei Unabhängigkeit muss die prozentuale Verteilung der unabhängigen Variablen in jeder Kategorie der abhängigen Variablen (annähernd) gleich sein
- Abweichungen von diesen Verteilungen lassen darauf schließen, dass die Variablen nicht unabhängig voneinander sind

→ „Es besteht ein Zusammenhang“

## Sinnvolle und konventionelle Erstellung:

- **Spalte:** „unabhängige“ Variable, **Zeile:** „abhängige“ Variable
- Als Basis der Prozentuierung dabei die unabhängige Variable wählen und interpretieren: Spaltenprozente

## Lesen“ und Interpretieren einer (konventionell erstellten) Kreuztabelle:

- Spaltenprozente zeilenweise vergleichen,
  - „Prozentsatzdifferenz“ ermitteln
- Beispiel: Gender gap im politischen Interesse? Hängt das Geschlecht mit dem politischen Interesse zusammen?



Geschlecht Politisches Interesse	Männliche Befragte	Weibliche Befragte	Gesamt
Sehr stark	311 17,6%	116 6,7%	427 12,2%
Stark	537 30,3%	345 20,1%	882 25,3%
Mittel	634 35,8%	795 46,2%	1429 40,9%
Wenig	207 11,7%	349 20,3%	556 15,9%
Überhaupt nicht	81 4,6%	115 6,7%	196 5,6%
Gesamt	1770 100,0%	1720 100,0%	3490 100,0%

Daten: ALLBUS 2016. Eigene Berechnungen

- liegt vor, wenn sich die Spaltenprozentage in einer Zeile nicht oder nur kaum unterscheiden
- **Faustregel (nach Kühnel/Krebs 2007)**
  - Differenzen unter 5 Prozentpunkte kaum interpretierbar
  - Differenzen unter 10 Prozentpunkte gelten als gering
  - Differenzen von 25 und mehr Prozentpunkten pro Zelle) weisen auf einen starken Zusammenhang hin

Dabei: auf Besetzung der einzelnen Zellen achten (mind. 15 Fälle)

- Kreuztabellen ermöglichen die kombinierte Betrachtung der Häufigkeiten
- Aussagekräftige bedingte prozentuale Häufigkeiten anzeigen lassen!
- Aber Hinweis: In den Sozialwissenschaften betrachten wir meistens komplexe Merkmale, die in Zusammenhang mit einer Vielzahl von Merkmalen stehen

- Erstellung einer Indifferenztabelle → Basis der
- Zusammenhangsmaße

- Erstellung einer Indifferenztabelle → Basis der
- Zusammenhangsmaße

- Kenntnis und Verständnis der Funktionsweise und Interpretation von Kreuztabellen