JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

7. Vorlesung am 7.12.23 (Thema Kreuztabelle)

Plan für heute

- **Wo stehen wir im Plan?**
- **Klärung etwaiger Fragen, kurze Wiederholung mit Übungen**
- **Einstieg bivariate Datenanalyse: Kreuztabelle**

Planung

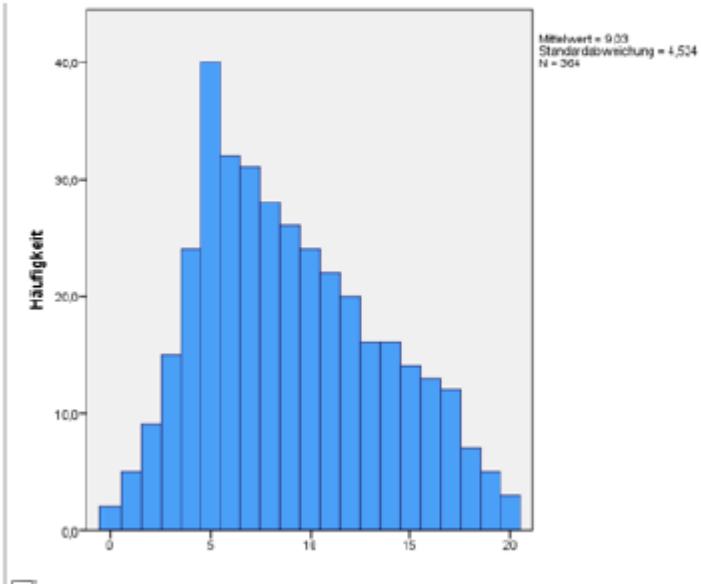
Lehrbrief	Kapitel 3, Abschnitt 1
7. Sitzung 7.12. Bivariate Statistik Teil 2	
Inhalt	• Zusammenhangmaße für nominale Merkmale: Chi-Quadrat und χ2-basierte Zusammenhangsmaße (C, Cramer's V) • Zusammenhangsmaß für ordinale Merkmale: Spearman's ρ (Rho)
WBT	Modul 3, Abschnitt 2
Lehrbrief	Kapitel 3, Abschnitt 2 – 3
8. Sitzung 14.12 Bivariate Statistik Teil 3	
Inhalt	• Zusammenhangmaße für metrische Merkmale (Pearson's r und PRE-Maß η2 (Eta-Quadrat) • PRE-Maß λ (Lambda)
WBT	Modul 3, Abschnitt 2
Lehrbrief	Kapitel 3, Abschnitt 4 - 6
9. Sitzung ACHTUNG findet online am 20.12. 12 bis 14 Uhr statt	
Inhalt	• Vortrag Ringvorlesung von Mical Gerezgiher und mir zum Thema „Demokratie leben lernen – Erste empirische Ergebnisse"
10. Sitzung 11.1. Grundlagen Inferenzstatistik	
Inhalt	• Statistische Verteilungen
WBT	Modul 3
Lehrbrief	Kapitel 5

Lernziel

- Kenntnis der Funktionsweise und Interpretation von Kreuztabellen

Übungen Wiederholung

- **Welche Eigenschaften treffen auf die folgende Verteilung zu?**



Übungen Wiederholung

- **Welches Skalenniveau hat die Variable „Berufsstatus“ (angestellt, verbeamtet, selbstständig)?**

Übungen Wiederholung

- **Welches Skalenniveau hat die Variable „Berufsstatus“ (Angestellt, verbeamtet, selbstständig)**

Übung Beispiel IQR

- Wie lautet im Beispiel der Interquartilsabstand?

Semesterzahl	Absolute Häufigkeit	%	Kumulierte %
10	1	9.1	9.1
11	2	18.2	27.3
12	3	27.3	54.6
13	2	18.2	72.8
14	1	9.1	81.9
15	1	9.1	91
20	1	9.1	100
$\sum$	*11*	*100*	*100*

Beispiel IQR

- Wie lautet im Beispiel der IQR → 14-11=3 Semester

→ Interpretation: (Etwas mehr als) 50% der Befragten haben zwischen 11 und 14 Semester für ihr Studium benötigt

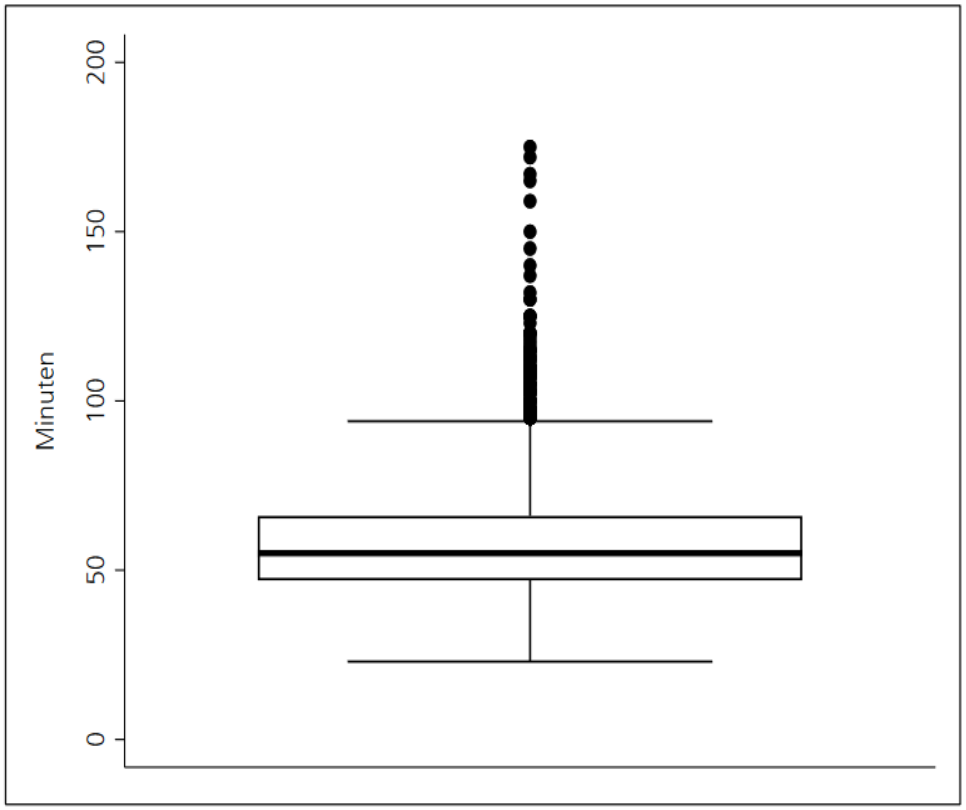
Semesterzahl	Absolute Häufigkeit	%	Kumulierte %
10	1	9.1	9.1
11	2	18.2	27.3
12	3	27.3	54.6
13	2	18.2	72.8
14	1	9.1	81.9
15	1	9.1	91
20	1	9.1	100
∑	*11*	*100*	*100*

Wiederholung Lagemaße/Streumaße

- Was sind Lagemaße?
- Was sind Streumaße?
- Warum sollte man sowohl Lage- als auch Streumaße bei der univariaten Datenanalyse ermitteln?
- Was ist ein Boxplot und was ermöglicht es Ihnen?

Übungen Wiederholung

- Interpretieren Sie folgendes Boxplot der Interviewdauer beim Allbus Survey in Minuten



Daten: ALLBUS 2016. Eigene Berechnungen

Übung: Boxplot

Sie haben für eine Verteilung folgende Kennwerte ermittelt:

	Wert
Minimum	=8
1. Quartil	=11
Median	=12
3. Quartil	=14
Maximum	= 16

Bitte skizzieren Sie auf dieser Basis ein einfaches Boxplot (senkrecht oder waagrecht)

Übung: Standardabweichung

Für welche Werte erwarten Sie die höchste Standardabweichung?

a) **19, 21, 31, 36**

b) **6, 11, 35, 21**

c) **146, 142, 141, 149**

d) **23, 201, 15, 167**

Übung z-Werte

Lisa und Bart haben jeweils an einem Leistungstest teilgenommen. Wer hat „besser“ abgeschnitten?

Person	Wert (x_i)	Arithmetisches Mittel ($\bar{x}$)	Standardabweichung (s_x)
Lisa	45	25	10
Bart	60	50	25

Übung z-Werte

Sie erreichen bei einem Leistungstest einen z-Wert von 0. Das bedeutet, dass…

1) Keine Ihrer Antworten zutreffend war, so dass Sie keinen Punkt erhalten haben
2) Ihr Testwert genau eine Standardabweichung über dem MW liegt
3) Ihr Testergebnis genau eine Standardabweichung unter dem MW liegt
4) Sie genau den durchschnittlichen Testwert erreicht haben.

Abschluss univariate Datenanalyse

- **Was ermöglicht uns die univariate Datenanalyse?**
- **Für welche Art der Fragestellungen ist sie geeignet?**
- **An welcher Stelle im Datenanalyseprozess steht sie?**
- **Was ermöglicht uns die univariate Datenanalyse NICHT?**

Bivariate Datenanalyse

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

Kreuztabellen

Auch „Kontingenztafel“

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

Kreuztabellen

Voraussetzung:

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

→ Kreuztabellen umfassen formal k-Zeilen und l-Spalten

Aber: Gestaltung sinnvoll, wenn nicht zu viele Ausprägungen vorhanden (da sonst unübersichtlich)

Kreuztabellen

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?
- Sind höher Gebildete eher politisch interessiert als niedriger Gebildete?
- Nutzen bestimmte Studiengänge eher das Abendangebot der UB als andere?

Beispiel

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

Beispiel

4 Kombinationen der beiden Merkmale möglich: welche?

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

Beispiel

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

Beispiel: Kreuztabelle

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

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

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

Kreuztabelle: Randhäufigkeiten

Studiengang / Nutzung	BA	MA	Gesamt
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

Kreuztabelle: Bedingte Häufigkeiten

2x2-Tabelle, Vierfeldertafel

Studiengang / Nutzung	BA	MA	Gesamt
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 Gesamtprozentuierung

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten, n=100

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

Studiengang / Nutzung	BA	MA	Gesamt
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 Gesamtprozentuierung

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 / Nutzung	BA	MA	Gesamt
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 Gesamtprozentuierung

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten, n=100

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

Studiengang / Nutzung	BA	MA	Gesamt
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 Zeilenprozentuierung

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten, n=100

Frage: Wieviel Prozent der abendlichen Nutzer sind im BA A-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	70	100

Beispiel Zeilenprozentuierung

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten, n=100

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

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 Zeilenprozentuierung

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	70	100

Beispiel Spaltenprozentuierung

Beispiel Befragung Bibliotheksnutzung, absolute Häufigkeiten, n=100

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

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

Beispiel Spaltenprozentuierung

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

Beispiel Spaltenprozentuierung

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 / Nutzung	BA	MA	Gesamt
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

Kreuztabelle

Sinnvolle und konventionelle Erstellung:

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

Exkurs: Hypothesen

- **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“

Exkurs: Abhängige und unabhängige Variable

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“)

Kreuztabelle

Sinnvolle und konventionelle Erstellung:

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

Kreuztabelle: statistische Unabhängigkeit

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“

Kreuztabelle

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? (aV? uV?)

Kreuztabelle, Beispiel

Politisches Interesse \ Geschlecht	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

Kreuztabelle: statistische Unabhängigkeit

- liegt vor, wenn sich die Spaltenprozente 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)

Zwischenfazit Kreuztabelle

- 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

Ausblick nächste Einheit

- Erstellung einer Indifferenztabelle → Basis der bivariaten Zusammenhangsmaße

Lernziel

- Kenntnis und Verständnis der Funktionsweise und Interpretation von Kreuztabellen