Westfälische
Wilhelms-Universität
Münster

Datenanalyse

Dr. Ingolf Terveer

Data Science: Statistik und Optimierung

Sommersemester 2022

WIRTSCHAFTS
INFORMATIK

1. Einführung

- Einordnung DA
- Aufbau der Vorlesung
- Organisation
- Literatur

Rückblick "Daten und Wahrscheinlichkeiten"

I. Deskriptive Statistik

- Analyse eines und zweier Merkmale
- Grafische Aufbereitung von Daten
- Lage, Streuungs- und Zusammenhangsmaße
- Clusteranalyse
- Statistische Software (R)

II. Wahrscheinlichkeitsrechnung

- Zufallsereignisse und Wahrscheinlichkeiten
- Zufallsvariablen
- Erwartungswert, (Ko-)varianz, Verteilungen von Zufallsvariablen
- Zufallsvektoren, Zusammenhangsmaße
- Gesetze großer Zahlen

Einordnung der Veranstaltungen DuW und EDA

Merkmal x
Stichprobe x_1
$\vdots$
x_n

Kennzahl $T(x_1, \ldots, x_n)$

$\hookrightarrow$ **Deskriptive Statistik:** Aggregation von Merkmalsausprägungen der einzelnen Merkmalsträger zu Aussagen über die Stichprobe.

Einordnung der Veranstaltungen DuW und EDA

Merkmal x

Stichprobe x_1

$\vdots$

x_n

beschreibt

Kennzahl $T(x_1, \ldots, x_n)$

$\hookrightarrow$ **Deskriptive Statistik:** Aggregation von Merkmalsausprägungen der einzelnen Merkmalsträger zu Aussagen über die Stichprobe.

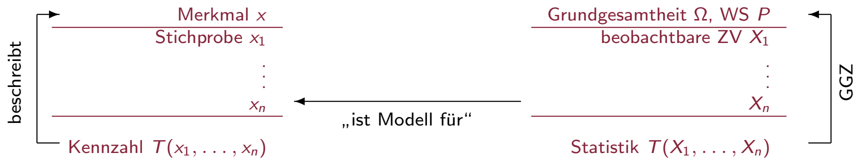
Einordnung der Veranstaltungen DuW und EDA

beschreibt			
	Merkmal x		Grundgesamtheit Ω, WS P
	Stichprobe x_1		beobachtbare ZV X_1
	$\vdots$		$\vdots$
	x_n	$\longleftarrow$ „ist Modell für“	X_n
	Kennzahl $T(x_1, \dots, x_n)$		Statistik $T(X_1, \dots, X_n)$

$\hookrightarrow$ **Deskriptive Statistik:** Aggregation von Merkmalsausprägungen der einzelnen Merkmalsträger zu Aussagen über die Stichprobe.

$\hookrightarrow$ **Wahrscheinlichkeitsrechnung:** WS-Modellierung und -Kalkül.

Einordnung der Veranstaltungen DuW und EDA

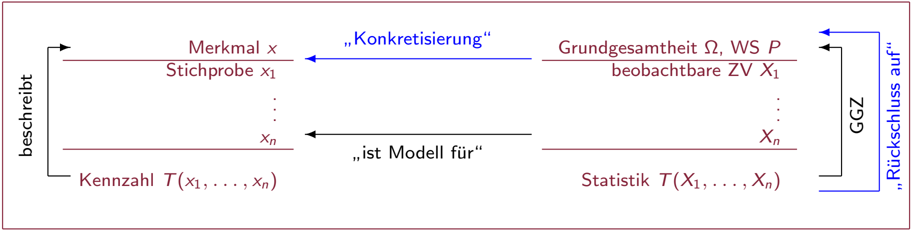


- ↪ **Deskriptive Statistik:** Aggregation von Merkmalsausprägungen der einzelnen Merkmalsträger zu Aussagen über die Stichprobe.
- ↪ **Wahrscheinlichkeitsrechnung:** WS-Modellierung und -Kalkül.
- ↪ **Gesetze großer Zahlen:** Verknüpfung von Daten und WS-Modellen, Zusammenhang zwischen theoretischen und empirischen Kennzahlen.

Einordnung der Veranstaltungen DuW und EDA

- ↪ **Deskriptive Statistik:** Aggregation von Merkmalsausprägungen der einzelnen Merkmalsträger zu Aussagen über die Stichprobe.
- ↪ **Wahrscheinlichkeitsrechnung:** WS-Modellierung und -Kalkül.
- ↪ **Gesetze großer Zahlen:** Verknüpfung von Daten und WS-Modellen, Zusammenhang zwischen theoretischen und empirischen Kennzahlen.
- ↪ **Induktive Statistik:** Schluss von Stichprobe auf Grundgesamtheit

Einordnung der Veranstaltungen DuW und EDA



- ↪ **Deskriptive Statistik:** Aggregation von Merkmalsausprägungen der einzelnen Merkmalsträger zu Aussagen über die Stichprobe.
- ↪ **Wahrscheinlichkeitsrechnung:** WS-Modellierung und -Kalkül.
- ↪ **Gesetze großer Zahlen:** Verknüpfung von Daten und WS-Modellen, Zusammenhang zwischen theoretischen und empirischen Kennzahlen.
- ↪ **Induktive Statistik:** Schluss von Stichprobe auf Grundgesamtheit
- ↪ **Konkretisierung:** Rückschluss von der Grundgesamtheit auf den Einzelfall

Beispiel: Zustimmung zu einer Umfrage ($n = 50$)

Merkmal x (Bereitschaft)
Stichprobe $x_1 \in \{0, 1\}$
$\vdots$
$x_n \in \{0, 1\}$
$T(x_1, \ldots, x_n) = \bar{x}$

- 50 Personen gefragt: „Würden Sie an Umfrage zu Corona teilnehmen?“

Beispiel: Zustimmung zu einer Umfrage ($n = 50$)

beschreibt

Merkmal x (Bereitschaft)

Stichprobe $x_1 \in \{0, 1\}$

$\vdots$

$x_n \in \{0, 1\}$

$T(x_1, \ldots, x_n) = \bar{x}$

- 50 Personen gefragt: „Würden Sie an Umfrage zu Corona teilnehmen?“

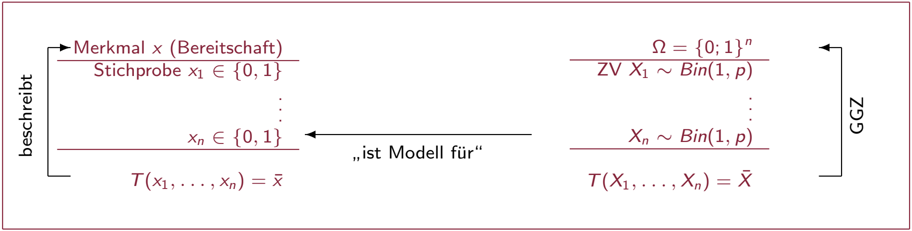
Beispiel: Zustimmung zu einer Umfrage ($n = 50$)

Merkmal x (Bereitschaft)		$\Omega = \{0; 1\}^n$
Stichprobe $x_1 \in \{0, 1\}$		ZV $X_1 \sim Bin(1, p)$
$\vdots$		$\vdots$
$x_n \in \{0, 1\}$	$\longleftarrow$ „ist Modell für“	$X_n \sim Bin(1, p)$
$T(x_1, \ldots, x_n) = \bar{x}$		$T(X_1, \ldots, X_n) = \bar{X}$

beschreibt

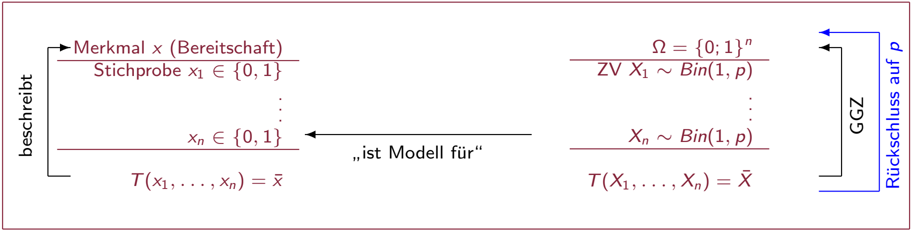
- 50 Personen gefragt: „Würden Sie an Umfrage zu Corona teilnehmen?“
- Modell: Bernoulli-Kette (n st.u. Einzelexperimente, $X_i \sim Bin(1, p)$), $nT(X) = X_1 + \cdots + X_n \sim Bin(n, p)$ (sog. **Stichprobenverteilung**)

Beispiel: Zustimmung zu einer Umfrage ($n = 50$)



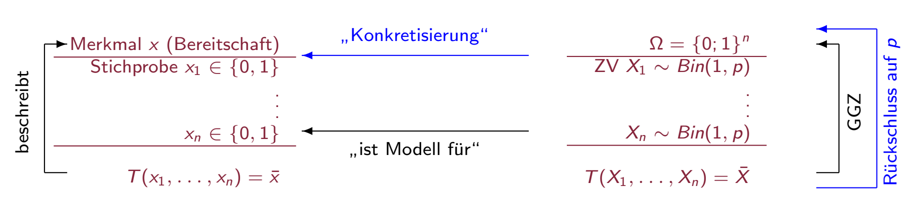
- 50 Personen gefragt: „Würden Sie an Umfrage zu Corona teilnehmen?“
- Modell: Bernoulli-Kette (n st.u. Einzelexperimente, $X_i \sim Bin(1,p)$),
  $nT(X) = X_1 + \cdots + X_n \sim Bin(n,p)$ (sog. **Stichprobenverteilung**)
- Gesetz großer Zahlen: $T(X) \xrightarrow{n\to\infty} p$ fast sicher, aber: p unbekannt, n endlich.

Beispiel: Zustimmung zu einer Umfrage ($n = 50$)



- 50 Personen gefragt: „Würden Sie an Umfrage zu Corona teilnehmen?“
- Modell: Bernoulli-Kette (n st.u. Einzelexperimente, $X_i \sim Bin(1, p)$),
  $nT(X) = X_1 + \dots + X_n \sim Bin(n, p)$ (sog. **Stichprobenverteilung**)
- Gesetz großer Zahlen: $T(X) \overset{n\to\infty}{\longrightarrow} p$ fast sicher, aber: p unbekannt, n endlich.
- Induktive Statistik: Verwende $T(X)$ als Ersatz für p („Schätzer“).
  Quantifiziere Fehlentscheidungen mit Hilfe der Verteilung von $T(X)$.

Beispiel: Zustimmung zu einer Umfrage ($n = 50$)



- 50 Personen gefragt: „Würden Sie an Umfrage zu Corona teilnehmen?“
- Modell: Bernoulli-Kette (n st.u. Einzelexperimente, $X_i \sim Bin(1, p)$), $nT(X) = X_1 + \dots + X_n \sim Bin(n, p)$ (sog. **Stichprobenverteilung**)
- Gesetz großer Zahlen: $T(X) \overset{n \to \infty}{\longrightarrow} p$ fast sicher, aber: p unbekannt, n endlich.
- Induktive Statistik: Verwende $T(X)$ als Ersatz für p („Schätzer“). Quantifiziere Fehlentscheidungen mit Hilfe der Verteilung von $T(X)$.
- Konkretisierung („Plug-In“): Interpretation von $T(X)$ als Rücklaufquote in der späteren größeren Umfrage mit mehr Befragten.

$\hookrightarrow$ Angenommen k von n Befragten sind bereit zu der Umfrage.

- $\hookrightarrow$ Angenommen k von n Befragten sind bereit zu der Umfrage.
- $\hookrightarrow$ Die größere Umfrage sollte nur bei ausreichend hoher Rücklaufquote p durchgeführt werden, diese ist aber unbekannt.
- $\hookrightarrow$ Typische, datenabhängig zu beantwortende Fragestellungen ($\alpha \in]0;1[$):

$\hookrightarrow$ Angenommen k von n Befragten sind bereit zu der Umfrage.

$\hookrightarrow$ Die größere Umfrage sollte nur bei ausreichend hoher Rücklaufquote p durchgeführt werden, diese ist aber unbekannt.

$\hookrightarrow$ Typische, datenabhängig zu beantwortende Fragestellungen ($\alpha \in]0;1[$):

- □ Finde eine Größe, die p möglichst gut „ersetzt“ $\rightsquigarrow$ **Schätzung**

$\hookrightarrow$ Angenommen k von n Befragten sind bereit zu der Umfrage.

$\hookrightarrow$ Die größere Umfrage sollte nur bei ausreichend hoher Rücklaufquote p durchgeführt werden, diese ist aber unbekannt.

$\hookrightarrow$ Typische, datenabhängig zu beantwortende Fragestellungen ($\alpha \in]0;1[$):

- □ Finde eine Größe, die p möglichst gut „ersetzt“ $\rightsquigarrow$ **Schätzung**
- □ Finde einen Bereich $[p_1(k); p_2(k)]$, in dem p mit $100(1-\alpha)\%$ Wahrscheinlichkeit liegt $\rightsquigarrow$ **Konfidenzintervall**

$\hookrightarrow$ Angenommen k von n Befragten sind bereit zu der Umfrage.

$\hookrightarrow$ Die größere Umfrage sollte nur bei ausreichend hoher Rücklaufquote p durchgeführt werden, diese ist aber unbekannt.

$\hookrightarrow$ Typische, datenabhängig zu beantwortende Fragestellungen ($\alpha \in]0; 1[$):

- □ Finde eine Größe, die p möglichst gut „ersetzt“ $\rightsquigarrow$ **Schätzung**
- □ Finde einen Bereich $[p_1(k); p_2(k)]$, in dem p mit $100(1-\alpha)\%$ Wahrscheinlichkeit liegt $\rightsquigarrow$ **Konfidenzintervall**
- □ Die größere Umfrage wird nur bei ausreichendem Interesse durchgeführt. Kann (z.B.) $p > 0,2$ aus der Vorumfrage mit $n = 50$ statistisch (d.h. zu einem Signifikanzniveau $100\alpha\%$) gesichert werden? $\rightsquigarrow$ **Hypothesentests**

$\hookrightarrow$ Angenommen k von n Befragten sind bereit zu der Umfrage.

$\hookrightarrow$ Die größere Umfrage sollte nur bei ausreichend hoher Rücklaufquote p durchgeführt werden, diese ist aber unbekannt.

$\hookrightarrow$ Typische, datenabhängig zu beantwortende Fragestellungen ($\alpha \in]0;1[$):

- □ Finde eine Größe, die p möglichst gut „ersetzt“ $\rightsquigarrow$ **Schätzung**
- □ Finde einen Bereich $[p_1(k); p_2(k)]$, in dem p mit $100(1-\alpha)\%$ Wahrscheinlichkeit liegt $\rightsquigarrow$ **Konfidenzintervall**
- □ Die größere Umfrage wird nur bei ausreichendem Interesse durchgeführt. Kann (z.B.) $p > 0,2$ aus der Vorumfrage mit $n = 50$ statistisch (d.h. zu einem Signifikanzniveau $100\alpha\%$) gesichert werden? $\rightsquigarrow$ **Hypothesentests**

Die Probleme hängen zusammen, werden mit der **Stichprobenverteilung**, d.h. der Verteilung von $T(X) = \bar{X}$ bzw. $nT(X) = X_1 + \cdots + X_n$ gelöst.

$\hookrightarrow$ Angenommen k von n Befragten sind bereit zu der Umfrage.

$\hookrightarrow$ Die größere Umfrage sollte nur bei ausreichend hoher Rücklaufquote p durchgeführt werden, diese ist aber unbekannt.

$\hookrightarrow$ Typische, datenabhängig zu beantwortende Fragestellungen ($\alpha \in]0;1[$):

- □ Finde eine Größe, die p möglichst gut „ersetzt" $\rightsquigarrow$ **Schätzung**
- □ Finde einen Bereich $[p_1(k); p_2(k)]$, in dem p mit $100(1-\alpha)\%$ Wahrscheinlichkeit liegt $\rightsquigarrow$ **Konfidenzintervall**
- □ Die größere Umfrage wird nur bei ausreichendem Interesse durchgeführt. Kann (z.B.) $p > 0,2$ aus der Vorumfrage mit $n = 50$ statistisch (d.h. zu einem Signifikanzniveau $100\alpha\%$) gesichert werden? $\rightsquigarrow$ **Hypothesentests**

Die Probleme hängen zusammen, werden mit der **Stichprobenverteilung**, d.h. der Verteilung von $T(X) = \bar{X}$ bzw. $nT(X) = X_1 + \cdots + X_n$ gelöst.

$\hookrightarrow$ Modellerweiterungen: Die Verteilungen von X_i hängen von weiteren beobachtbaren Merkmalen ab (z.B. sozialer Status), d.h. Modellierung von p abhängig von diesen Merkmalen $\rightsquigarrow$ **bedingte Verteilungen, Regression**

Aufbau der Vorlesung DA

I. Stichprobenverteilungen

II. Bedingte Erwartung

III. Schließende Statistik

- Punktschätzung
- Intervallschätzung
- statistische Tests

IV. Regression

Zusammenhang zwischen abhängiger Variable Y u. erklärenden Variablen X.

Vorlesung/Präsenzübung: I. Terveer

- ↪ Vorlesung: Videos (Folien), werden sukzessive bereitgestellt
- ↪ Übung: Zum Selbststudium enthalten die Folien Übungsteile, die für die folgende Präsenzübung vorbereitet werden müssen.
  Termine der Präsenzübung (Mi im J2, Di im Leo 1):

Mi	06.04.2022	10:00	12:00	V1: Einleitung
Mi	13.04.2022	10:00	12:00	V2: Stichprobenverteilungen
Mi	20.04.2022	10:00	12:00	V3: Bedingte Verteilung und bedingte Erwartung
Mi	27.04.2022	10:00	12:00	V4: ML-Schätzer und MM-Schätzer
Mi	04.05.2022	10:00	12:00	V5: Intervall- und Bereichsschätzer
Mi	11.05.2022	10:00	12:00	V6: Tests, Fehler 1./2. Art, Konstruktion. Einstichprobentests für EW
Mi	18.05.2022	10:00	12:00	V7: Anteilswerttests, Gütefunktion
Mi	25.05.2022	10:00	12:00	V8: Zweistichprobentests: Mittelwerttests, Wilcoxon-Test ( )
Mi	01.06.2022	10:00	12:00	V9: Verteilungstests: Chi-Quadrat-Unabh, Anp, KS
Mi	15.06.2022	10:00	12:00	V10: Regressionsanalyse theor. Lösung, Modell, Schätzansatz, einfache LR
Mi	22.06.2022	10:00	12:00	V11: KQ-Schätzer, Bestimmtheitsmaß, Hypothesentests
Mi	29.06.2022	10:00	12:00	V12: Parametertests, Bereichsschätzung, Residualanalyse
Di	05.07.2022	16:00	18:00	V13: Indikatorvariablen ( )
Mi	06.07.2022	10:00	12:00	V14: Fragestunde

- ↪ Bereitstellung von Lösungen: nach der Präsenzübung
- ↪ Notwendige Vorkenntnisse: MAWIWI (Differential- und Integralrechnung, Matrizenrechnung), DuW

Tutorium („Übungen zur Vorlesung Datenanalyse“)

↪ Wöchentlich: Übungszettel

- u.a. Fortsetzung der R-Übungsaufgaben aus DuW
- u.a. Aufgaben zur Klausurvorbereitung
- Ausgabe nach VL-Übung,
- Bearbeitung im Selbststudium

↪ In der Folgewoche:

- Übungstermin, Di 16-18: Besprechung von Fragen
- Bereitstellung von Musterlösungen (Ende der Woche)
- Ggf. Besprechung von Fragen zu Übung der Vorwoche.

Weitere Informationen:

↪ DA im Learnweb (Materialien, Foren,Evaluation,...):

- https://sso.uni-muenster.de/LearnWeb/learnweb2/course/view.php?id=61044
- kein Einschreibeschlüssel bis Ende April, danach Einschreibung nur mit pM

↪ Informationen zur Prüfung:

- keine Anrechnung von Übungsaufgaben
- Klausur „Datenanalyse und Simulation“ 18.7.
  (120 min, voraussichtlich 80P DA, 40P Sim)

  Prüfungsformat: voraussichtlich Präsenzklausur.
- R ist Bestandteil der Klausur

Literatur

[1] Auer, B./Rottmann, H.: Statistik und Ökonometrie für Wirtschaftwissenschaftler, 2. Aufl. Gabler. 2011

[2] Durbin, J.: Distribution Theory for Tests Based on the Sample Distribution Function. Philadelpha: Society for Industrial and Applied Mathematics. 1973.

[3] Lilliefors, H.W.: On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. JASA, 62(318), 299-402. 1967

[4] Mood, A.M./Graybill, F.A./Boes, D.C.: Introduction to the theory of Statistics. McGraw-Hill, Singapore. 1974.

[5] Montgomery, D.C./Peck, E.A./Vining, G.G.: Introduction to Linear Regression Analysis, 5th ed. Wiley, Hoboken, New Jersey. 2012.