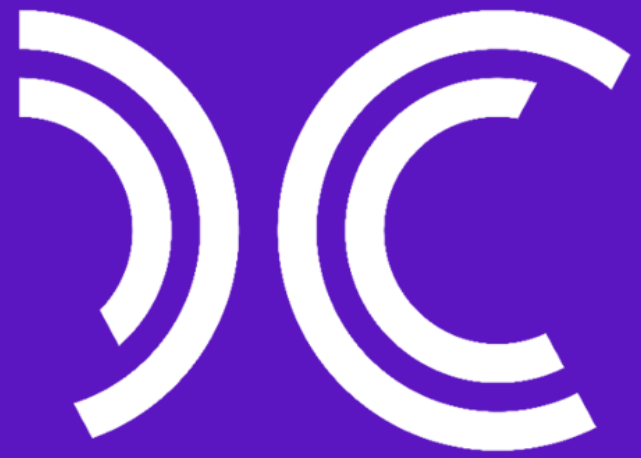


DataCraft

Überblick

DataCraft

Einführung
30 UE

Einführung in Data Analytics
40 UE

Datenauswertung in Excel
150 UE

Präsentation & Kommunikation
50 UE

Grundlagen der Python-Programmierung
200 UE

Python für Datenanalysten
230 UE

Datenbanken & SQL
300 UE

Datenschnittstellen
100 UE

Statistik-Grundlagen
150 UE

Business Analytics
50 UE

Datenvisualisierung mit Python
150 UE

Power BI
200 UE

KI & Machine Learning
100 UE

Low-Code ML (IHK)
100 UE

Arbeitsmarkttransfer und -vorbereitung
150 UE

Abschluss-Projekt
400 UE

DataCraft

Statistik-Grundlagen

Inhalt

DataCraft

1. Skalen- und Merkmalstypen - HEUTE
2. Lage- und Streuungsmaße
3. Visualisierungen
4. Kombinatorik
5. Verteilungen
6. Zusammenhangsmaße
7. Konfidenzintervalle
8. Fehlende Werte
9. Hypothesentests
10. Lineare Regression

DataCraft

Statistik

Die Lehre von Methoden zum Umgang mit quantitativen Informationen (Daten)

deskriptive Statistik
(beschreibende S., empirische S.)

Vorliegende Daten werden in geeigneter Weise beschrieben, aufbereitet und zusammengefasst.

Prädiktiv
Induktive Statistik
(schließende S.)

Herleitung von Eigenschaften einer Grundgesamtheit anhand von Daten einer Stichprobe

klassisch

explorative Statistik
(hypothesen-generierende S.)

Systematische Suche nach möglichen Zusammenhängen/ Unterschieden

Python Modulübersicht

DataCraft

Zusätzlich zu math, numpy, pandas, matplotlib und seaborn benötigen wir folgende Module:

Modul	Beschreibung	Quelle	Installations-Befehl
statistics	Statistische Funktionen	Standardbibliothek	–
scipy	„Scientific Python“	conda	conda install scipy
statsmodels	Statistische Modelle	conda	conda install statsmodels
lmdiag	Visualisierungen	pip	pip install lmdiag

DataCraft

1 Skalen und Merkmalstypen

DataCraft

Um welche Art von Daten handelt es sich bei einer Variable?

- **Nominalskala**
  Daten können nur bennant werden, z.B. Farben
- **Ordinalskala**
  Daten haben eine natürliche Ordnung/Reihenfolge, d.h. man kann die Daten sortieren. Z.B. eine Frage mit Antwortmöglichkeiten immer, häufig, selten, nie)

kategorial

- **Kardinalskala**
  - **Intervallskala**: Abstände zwischen Werten können gemessen werden, z.B. Zeitpunkte
  - **Verhältnisskala**: Zusätzlich zur Intervallskala gibt es einen natürlich Nullpunkt, z.B. Gewicht oder Alter; Temperatur in Kelvin

metrisch

Die Skala gibt an, was man mit den Daten machen darf, z.B. Mittelwert bilden!

Skalen Schaubild

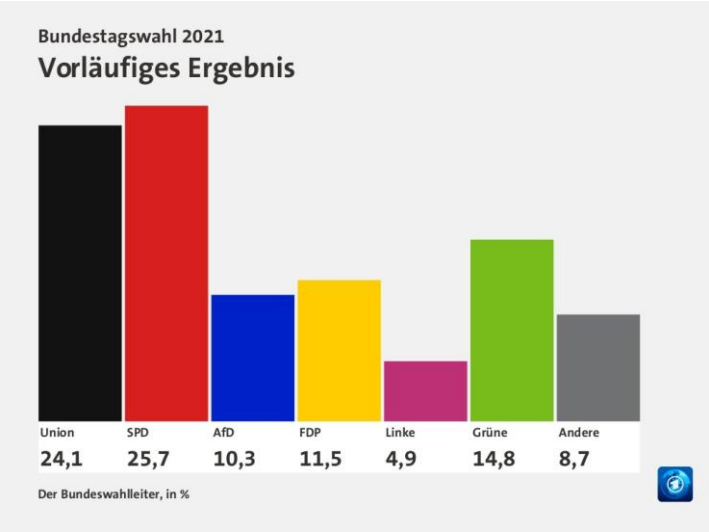
DataCraft

Skala	Eigenschaften			
	Kategorial		Metrisch	
Nominal	Häufigkeit			
Ordinal	Häufigkeit	Rangfolge		
Intervall	Häufigkeit	Rangfolge	Interpretierbare, regelmäßige Abstände	
Verhältnis	Häufigkeit	Rangfolge	Interpretierbare, regelmäßige Abstände	Nullpunkt

Warum Statistik?

- Statistik findet sich überall im Alltag
- Besonders zur Zeit von Corona sah man sich damit Zunehmens konfrontiert
- Um Statistiken zu verstehen, muss man Statistik verstehen!

Welche Skalen haben diese Daten?



Saison
2021–22

	Verein	Sp	S	U	N	T	GT	TD	Pkte	Letzte 5
1	Bayern	7	5	1	1	24	7	17	16	✗ ✓ ✓ ✓ ✓
2	Leverkusen	7	5	1	1	20	7	13	16	✓ ✓ ✓ ✗ ✓
3	Dortmund	7	5	0	2	19	13	6	15	✓ ✗ ✓ ✓ ✓
4	Freiburg	7	4	3	0	11	5	6	15	✓ ✓ – – ✓
5	Wolfsburg	7	4	1	2	9	8	1	13	✗ ✗ – ✓ ✓
6	Köln	7	3	3	1	13	9	4	12	✓ – – – ✓
7	Union Berlin	7	3	3	1	10	9	1	12	✓ ✓ ✗ – ✓
8	RB Leipzig	7	3	1	3	15	7	8	10	✓ ✓ – ✗ ✗
9	Mainz	7	3	1	3	7	5	2	10	✗ ✗ – ✓ ✓
10	Mönchenglad…	7	3	1	3	9	10	-1	10	✓ ✓ ✗ ✓ ✗
11	Hoffenheim	7	2	2	3	12	11	1	8	✗ ✓ – ✗ ✗
12	VfB Stuttgart	7	2	2	3	12	13	-1	8	✓ – ✗ – ✗
13	Eintracht Fran…	7	1	5	1	8	10	-2	8	✓ – – – –

Deskriptive Statistik

- beschreibt eine Stichprobe
- umfasst u.a.
  - Mittelwerte
  - Streuungsmaße
  - Häufigkeitsverteilungen
- **Reine Beschreibung der Stichprobe, macht keine Schlussfolgerungen**

Induktive Statistik

- **zieht statistische Schlüsse** auf Basis einer (möglichst repräsentativen) Stichprobe auf die Grundgesamtheit
- umfasst u.a. statistische Tests, z.B. t-Test, Chi²-Test, p-Test
- jede Schlussfolgerung ist mit einer gewissen Unsicherheit verbunden (**Irrtumswahrscheinlichkeit**) – Ziel ist i.d.R. Irrtums-wahrscheinlichkeit von max. 5 %

Merkmal, Merkmalsträger, Merkmalsausprägung

DataCraft

Was	Beispiel	In unseren Daten wäre das
Merkmalsträger	Eine Person	1 ganze Zeile
Merkmal	Studiengang, Körpergröße, Zufriedenheit einer Person	• Je Merkmal 1 Spalte • Je Merkmal PRO Merkmalsträger 1 Zelle
Beobachtungs-werte	Eine menschliche Körpergröße von 180 cm bei Person X / Merkmalsträger X	Der Wert in einer Zelle (in der Zeile der Person X)
Merkmals-ausprägung	Theoretisch mögliche Werte. (Eine Körpergröße von 1180 cm ist nicht sinnvoll, sondern ein Fehler im Datensatz)	(siehe Beobachtungswerte)

Population und Stichprobe

DataCraft

Was	Beispiel	In unseren Daten wäre das
Population	Bei der Bundestagswahl: Die Stimme aller wahlberechtigten Bürger Deutschlands	Nichts, da es fast immer viel zu aufwändig ist, alle zu befragen. Wenn doch: der gesamte Datensatz
Stichprobe	Der Teil der Population, den man befragt hat (z.B. wenn 55% der Bürger befragt werden, ist das meine Stichprobe)	der gesamte Datensatz oder ein Auszug daraus

DataCraft

Häufigkeitsverteilungen

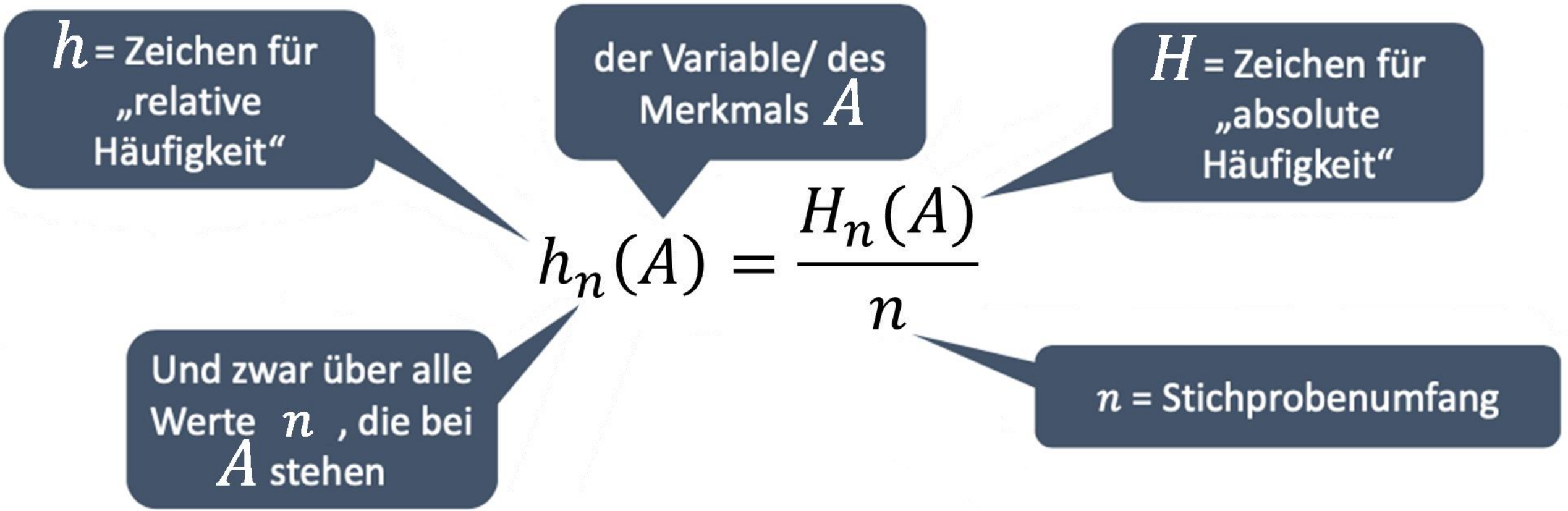
Kurzeinführung statistische Formeln

Symbole der Statistik und Mathematik

DataCraft

Messbare Eigenschaft	Symbol	Beispiel: Bundestagswahl
Stichprobenumfang	n	Gesamt (gültige Stimmen): 46.419.448 Stimmen
Absolute Häufigkeit (Anzahl / Count)	H	CDU: 8.770.980 Stimmen CSU: 2.402.826 Stimmen
Relative Häufigkeit (Anteil)	h	CDU: 18,9% (von 100) Anteil der gültigen Stimmen CSU: 5,2% Anteil der gültigen Stimmen
Kumulierte absolute Häufigkeit	F_{abs}	Union (= CDU & CSU summiert): 11.175.806 Stimmen
Kumulierte relative Häufigkeit	F_{rel}	Union (= CDU & CSU summiert): 24,1 % Stimmenabteil der gültigen Stimmen

Erste Formel: Berechnung relative Häufigkeit



Relative Häufigkeit Beispiel