02

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

STATISTISCHE GRUNDLAGEN

KURSINHALTE UND TERMINE

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Kursinhalte

Data Science	**1**
Statistische Grundlagen	**2**
Softwareentwicklung (Paradigmen & Projektmanagement)	**3**
Testing / Integration / Deployment	**4**
Ansätze, Methoden und Anwendungen Künstlicher Intelligenz (KI)	**5-8**
Zusammenfassung / Fragen / Klausurvorbereitung	**9**

Termine

#	Wochentag	Datum	von - bis	Räume
1	Freitag	04.04.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
2	Freitag	11.04.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.27 Bothfeld
3	Freitag	25.04.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
4	Freitag	16.05.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
5	Freitag	23.05.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
6	Freitag	06.06.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
7	Freitag	20.06.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
8	Freitag	04.07.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
9	Freitag	11.07.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt

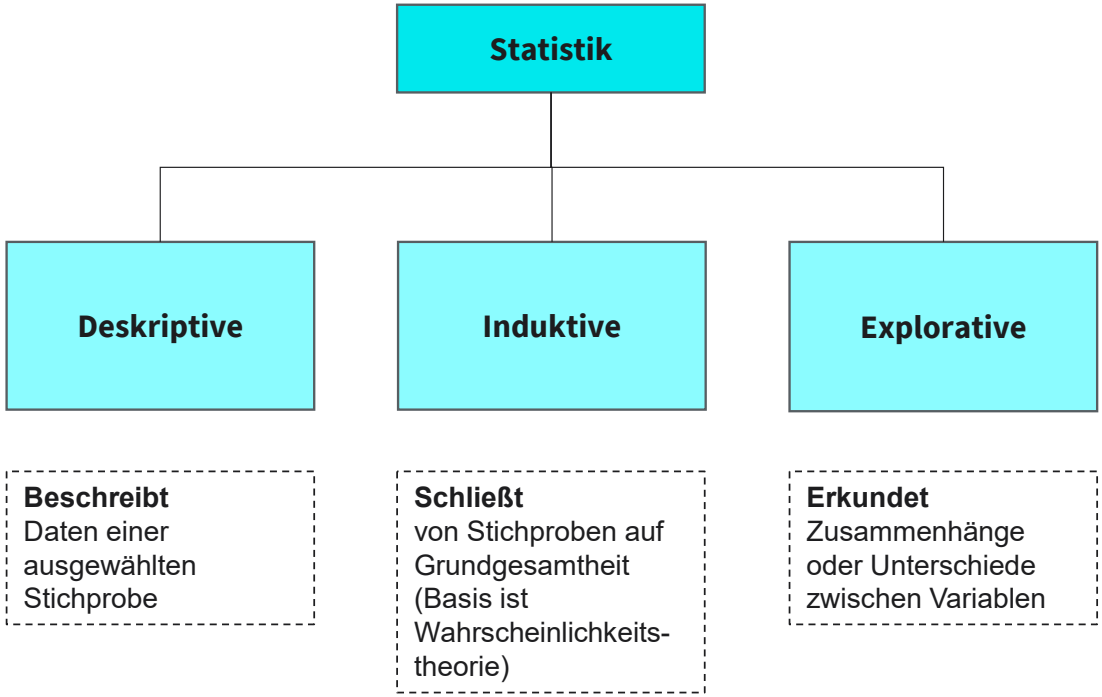
Statistische Grundlagen

Lernziele

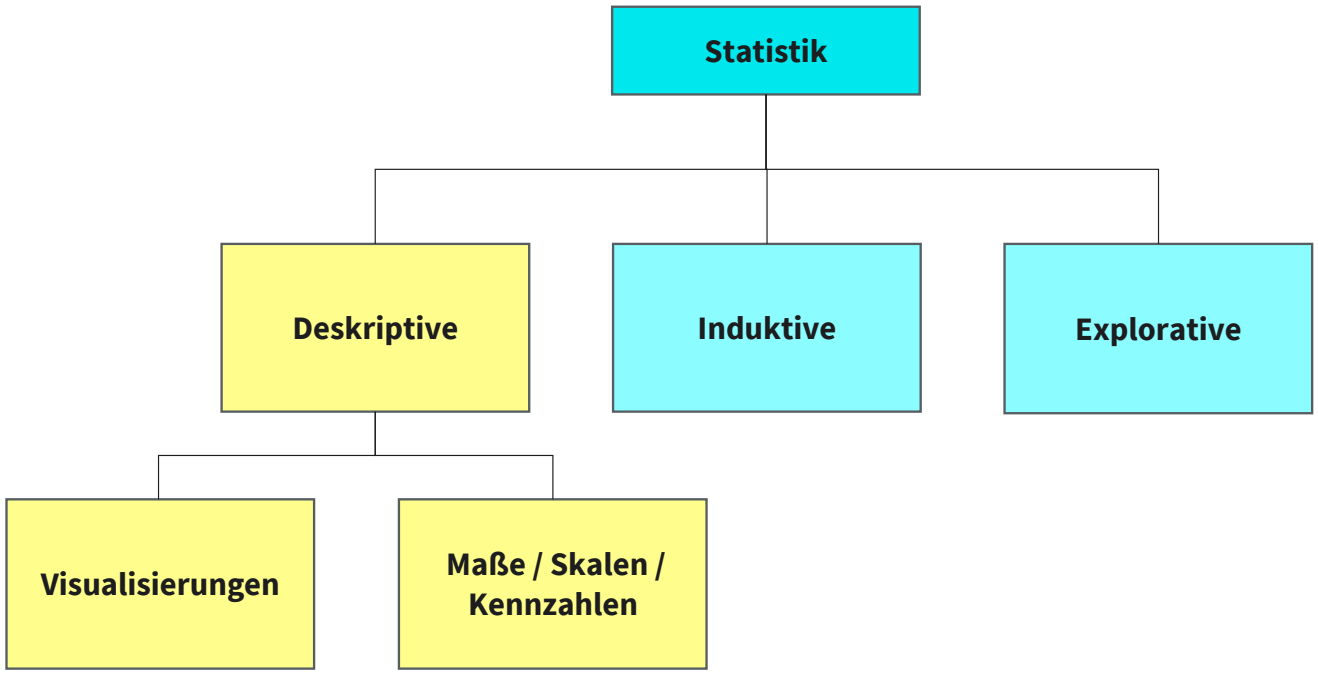
Nach der Bearbeitung dieser Lektion werdet ihr wissen, …

- in welche **Teilbereiche** sich Statistik untergliedern lässt.
- was unter **deskripter Statistik** verstanden wird und welche **Visualisierungen**, **Maße** und **Kennzahlen** dafür verwendet werden.
- was unter **induktiver Statistik** verstanden wird, wie **Hypothesenpaare** aufgestellt werden und wie **statistische Tests** angewendet werden.
- was unter **explorativer Statistik** bzw. **explorativer Datenanalyse (EDA)** verstanden wird.
- Wie sich **univariate** , **bivariate** und **multivariate Analysemethoden** unterscheiden.

STATISTIK
ABGRENZUNG NACH DER ART IN VIER TEILBEREICHE



STATISTIK
ABGRENZUNG DER ARTEN



Darstellung empirischer Daten durch Grafiken und statistische Maße / Skalen / Kennzahlen

DESKRIPTIVE STATISTIK – VISUALISIERUNGEN TABELLEN UND GRAFIKEN

„Mittels **deskriptivstatistischer Methoden** soll eine erste **Visualisierung der Daten** in Form von Tabellen, Diagrammen, einzelnen Kennwerten und Grafiken erfolgen. Es geht dabei in erster Linie um eine **Beschreibung**, einen guten **Überblick zu verschaffen** und **wesentliche Informationen** herauszufiltern – im engeren Sinne um eine Reduktion der Daten. Wichtige Hauptaussagen sollen auf den ersten Blick erkenntlich werden.“
(Raab-Steiner, Benesch, 2015, S. 82)

Tabellarische Darstellung der Daten

- Häufigkeitstabelle
- Kreuztabelle (Kontingenztafeln)

Grafische Darstellung der Daten

- Balkendiagramm
- Histogramm
- Boxplots
- Streudiagramm

Quelle: Raab-Steiner, Benesch, 2015,S 82 -94.

DESKRIPTIVE STATISTIK – VISUALISIERUNGEN TABELLEN UND GRAFIKEN

Tabellarische Darstellung der Daten mittels Häufigkeitstabelle

- Darstellung der absoluten und relativen Häufigkeiten
- „Gültige Prozente“ berücksichtigen gegebenenfalls fehlende Werte
- „Kumulierte Werte“ zeigen bspw. 60 Prozent haben einen positiven Wert gewählt

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
+++	4	20,0	20,0	20,0
++	3	15,0	15,0	35,0
+	5	25,0	25,0	60,0
-	4	20,0	20,0	80,0
--	2	10,0	10,0	90.0
---	2	10,0	10,0	100,0
Gesamt	20	100,0	100,0	

Quelle: Raab-Steiner, Benesch, 2015, S 82.

Quelle: Raab-Steiner, Benesch, 2015,S 82 -94.

DESKRIPTIVE STATISTIK – VISUALISIERUNGEN TABELLEN UND GRAFIKEN

Tabellarische Darstellung der Daten mittels Kreuztabelle (auch Kontingenztafeln genannt)

- Darstellung der absoluten Häufigkeiten bestimmter Ausprägungen von Merkmalen (kategorial bzw. nicht metrisch)
- Beziehungen der Häufigkeitsverteilungen mehrere Merkmale untereinander
- Ablesen einzelner Beziehungen, bspw. Eine Person mit „finanzielle Situation“ +++ und „Wohnsituation“ +++
- Signifikanz kann mittels x^2-Test überprüft werden

	Wohnsituation						
Finanzielle Situation	+++	++	+	-	--	---	**Gesamt**
+++	1	0	2	0	0	0	3
++	0	0	1	0	0	0	1
+	0	1	0	1	1	0	3
-	1	0	0	1	1	0	3
--	1	0	0	1	1	1	4
---	0	2	0	1	1	2	6
Gesamt	3	3	3	4	4	3	20

Quelle: Raab-Steiner, Benesch, 2015, S 86.

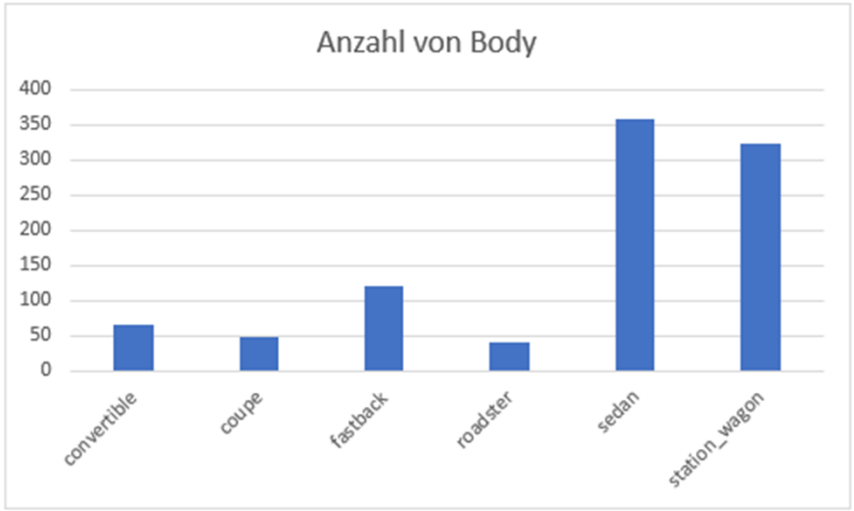
Quelle: Raab-Steiner, Benesch, 2015,S 82 -94.

DESKRIPTIVE STATISTIK – VISUALISIERUNGEN TABELLEN UND GRAFIKEN

Grafische Darstellung der Daten

Balkendiagramm

- Darstellung der Häufigkeiten von **nominal-** oder **ordinalskalierten** Variablen.
- Üblicherweise wird die ausgewählte Variable auf der x-Achse (Abzisse) und die absoluten Werte auf der y-Achse (Ordinate) dargestellt
- Hier verschiedene Fahrzeugtypen und deren Häufigkeit im Datensatz



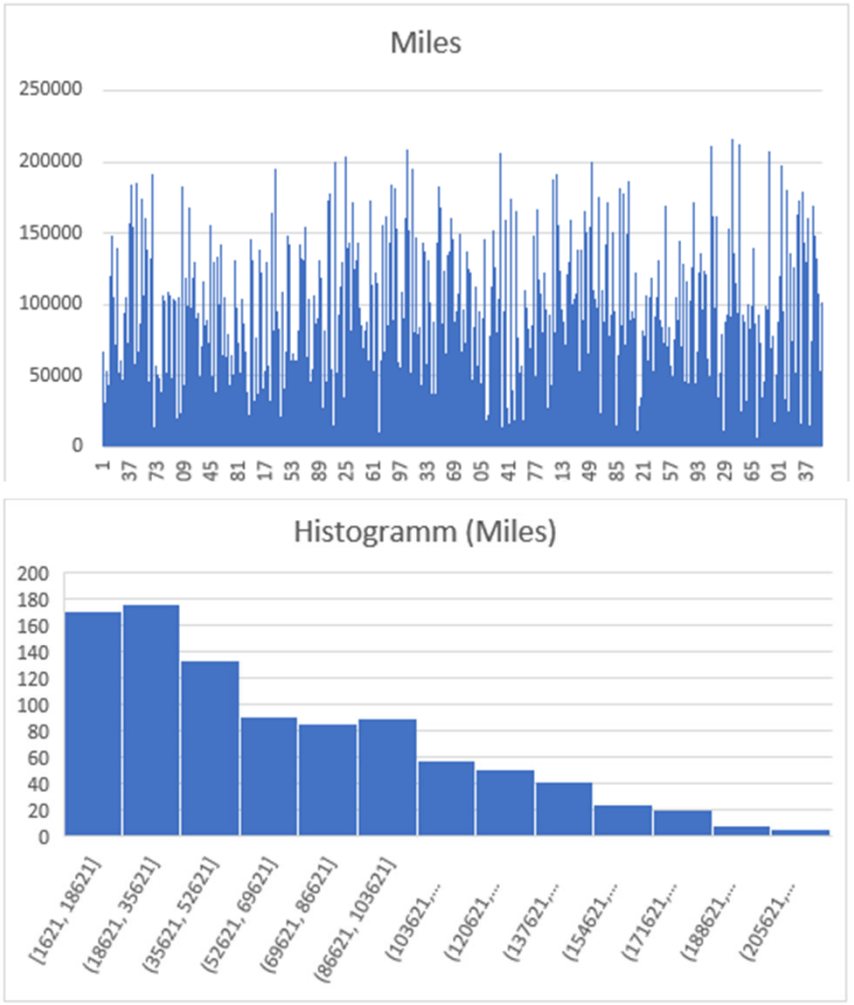
Quelle: Eigene Darstellung.

Quelle: Raab-Steiner, Benesch, 2015,S 82 -94.

DESKRIPTIVE STATISTIK – VISUALISIERUNGEN TABELLEN UND GRAFIKEN

Grafische Darstellung der Daten
Histogramm

- Darstellung der Häufigkeiten von **intervallskalierten** Variablen
- Sofern viele verschiedene Werte vorliegen, wird ein Balken für jeden einzelnen Wert zu unübersichtlich (siehe oben)
- In einem Histogramm werden Werten in Klassen zusammengefasst und die Klassenhäufigkeiten als Balken dargestellt (siehe unten)
- Hier am Beispiel „Miles“



Quelle: Raab-Steiner, Benesch, 2015,S 82 -94.

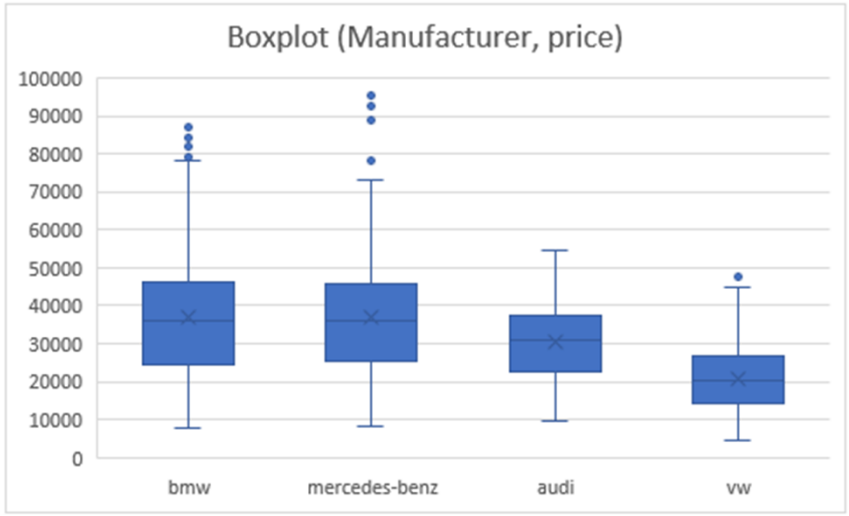
Quelle: Eigene Darstellung.

DESKRIPTIVE STATISTIK – VISUALISIERUNGEN TABELLEN UND GRAFIKEN

Grafische Darstellung der Daten

Boxplots

- Darstellung von **Median** und **Quantile** von **intervallskalierten** Variablen
- Untere und obere Linien markieren den kleinsten und größten Wert
- Untere Begrenzung der Box ist das erste Quartil (Q1, 25% liegen unterhalb)
- Die obere Begrenzung der Box ist das dritte Qaurtil (Q3, 75 % liegen unterhalb)
- Mittlere Linie zeigt den Median (50%)
- Hier am Beispiel „Manufacturer“ und „Price“



Quelle: Eigene Darstellung.

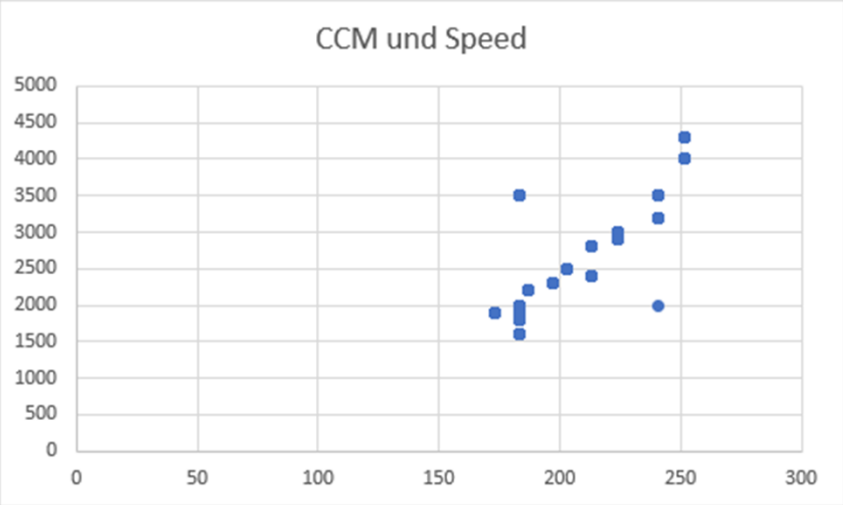
Quelle: Raab-Steiner, Benesch, 2015, S. 82 -94.

DESKRIPTIVE STATISTIK – VISUALISIERUNGEN TABELLEN UND GRAFIKEN

Grafische Darstellung der Daten

Streudiagramm

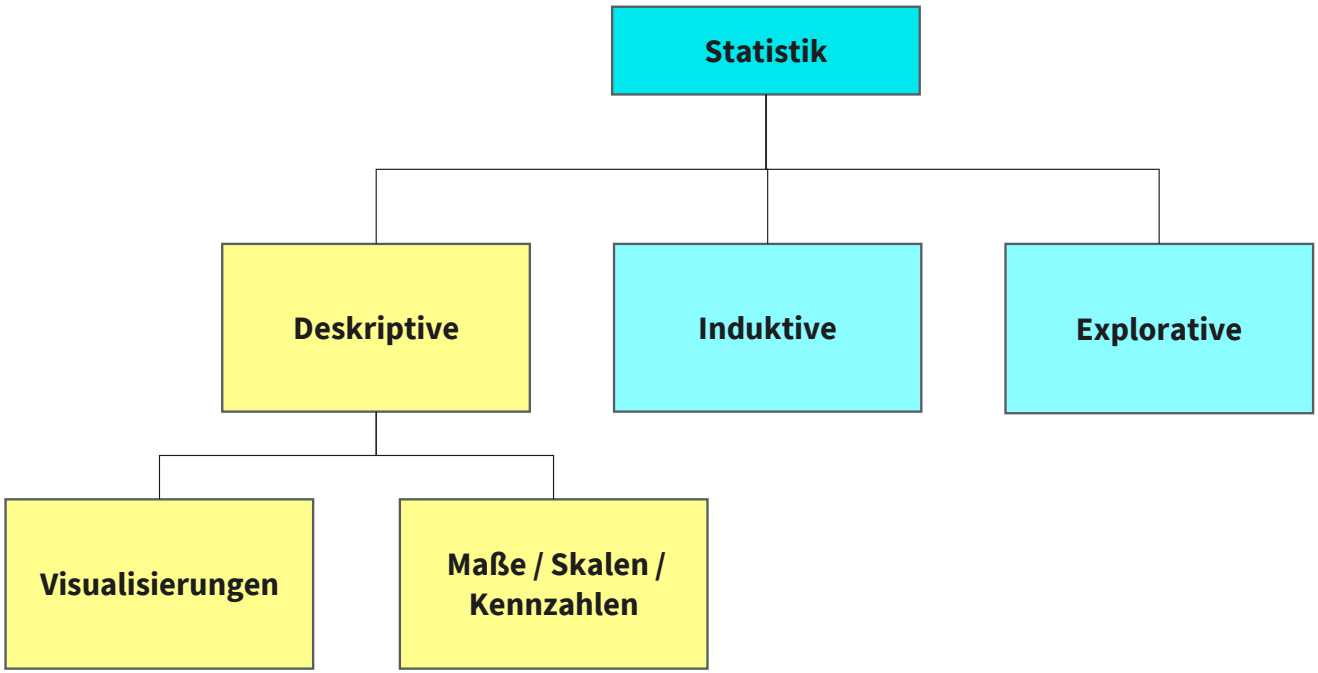
- Grafische Darstellung des **Zusammenhangs von Variablen** über **Punktewolke**
- Betrachtung von Wertepaaren zweier Variablen, jedes Wertepaar ist ein Punkt im Koordinatensystem
- Sofern überdurchschnittlich hohe Werte einer Variablen mit überdurchschnittlich hohen Werten der anderen Variablen einhergehen und überdurchschnittlich niedrige Werte mit überdurchschnittlich niedrigen Werten, spricht man von einem positiven Zusammenhang
- Gegenläufige Beobachtungen nennt man negativer Zusammenhang
- Hier am Beispiel „CCM“ und „Speed“



Quelle: Eigene Darstellung.

Quelle: Raab-Steiner, Benesch, 2015, S. 82 -94.

STATISTIK
ABGRENZUNG DER ARTEN



Darstellung empirischer Daten durch Grafiken und statistische Maße / Skalen / Kennzahlen

DESKRIPTIVE STATISTIK – MAßE – SKALEN - KENNZAHLEN
SKALEN (NIVEAU)

Skala		Merkmale	Beispiel	Berechnungen	Maße
Nicht-metrische Skalen (Kategorial)	NOMINAL	Eigenschafts-ausprägungen	Müller, Meier, Schulze	Häufigkeiten, Gleich oder Ungleich (= , ≠)	Modus
	ORDINAL	Rangwert	Sehr gut, gut, befriedigend	Rang und Position (<, >)	Median, Quantile
Metrische Skalen	INTERVALL	Gleich große Abschnitte	20° C, 2020 n Chr.	Addition, Subtraktion (+, -)	Mittelwert, Standard-abweichung
	RATIO (Verhältnis)	Natürlicher Nullpunkt	20 cm, 2 kg, 100km/h, 50 €	Division, Multiplikation (/ , *)	Verallgemeinerter Mittelwert

„Je höher das Skalenniveau ist, desto größer ist auch der Informationsgehalt der betreffenden Daten und desto mehr Rechenoperationen und statistische Maße lassen sich auf die Daten anwenden.“
„Mit der Transformation auf ein niedrigeres Skalenniveau ist natürlich immer auch ein Informationsverlust verbunden“ (Backhaus, et. al., 2011, S. 11, 12)

Quellen: Backhaus et. al, 2011, S- 11, Runkler 2015, S8.

DESKRIPTIVE STATISTIK – MAßE – SKALEN - KENNZAHLEN
LAGEMAßE UND STREUUNGSMAßE

Lagemaße (Kennzeichnung des Zentrums)

> „Von Interesse sind Statistiken, die als **Lagemaße** die Position des Zentrums einer Verteilung in Form eines **zentralen Wertes** beschreiben.“ (Kähler, 2011, S. 37)

Streuungsmaße (Kennzeichnung der Variabilität)

> „**Streuungsmaße** geben darüber Auskunft, wie sehr sich vorliegende **Messwerte voneinander unterscheiden**, wie die **Verteilung** von einzelnen gewonnenen Messwerten aussieht, präziser formuliert, **wie breit eine Verteilung ist**. Es ist daher sehr wichtig , zu einem Lagemaß auch das entsprechende Streuungsmaß anzugeben. Dadurch kann man in Erfahrung bringen, wie sehr einzelne Werte von der Mitte abweichen. „ (Raab-Steiner, Benesch, 2015, S. 99)

DESKRIPTIVE STATISTIK – MAßE – SKALEN - KENNZAHLEN
LAGEMAßE UND STREUUNGSMAßE

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Lagemaße (Kennzeichnung des Zentrums)

Lagemaße werden je nach **Skalenniveau** unterschiedlich gebildet:

- **Modus (Modalwert)**

„Der Modalwert ist der am häufigsten auftretende Wert in einer Stichprobe. Er ist eine passende Kennzahl für **nominalskalierte** Variablen.“ Beispiel Messwerte: 1, **2, 2, 2**, 3, 6, 6, 7, 7 => **Modus: 2**

- **Median**

„Der Median, auch Zentralwert genannt, ist derjenige Punkt der Verteilung, unterhalb und oberhalb dem jeweils die Hälfte der Messwerte liegt. Der Median ist eine passende Kennzahl für **ordinalskalierte** und **nicht normalverteilte** Variablen.“ Beispiel Messwerte: 1, 2, 2, 2, **3**, 6, 6, 7, 7 => **Median: 3**

- **Mittelwert (Arithmetisches Mittel)**

„Der Mittelwert ist eine passende Kennzahl für **intervallskalierte und normalverteilte** Variablen. Der Mittelwert ist das Arithmetische Mittel der Messwerte und berechnet sich daher aus der Summe der Messwerte dividiert durch deren Anzahl.“ Beispiel Messwerte: 1, 2, 2, 2, 3, 6, 6, 7, 7 => **Mittelwert: 36 / 9 = 4**

Quelle: Raab-Steiner, Benesch, 2015, S. 96 -99.

DESKRIPTIVE STATISTIK – MAßE – SKALEN - KENNZAHLEN
LAGEMAßE UND STREUUNGSMAßE

Streuungsmaße (Kennzeichnung der Variabilität)

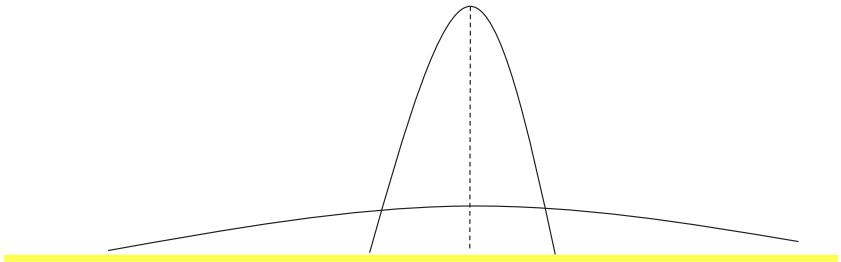
- **Spannweite**

„Die Spannweite… ist die Ausdehnung zwischen dem Maximum (höchster Messwert) und dem Minimum (niedrigster Messwert). Man bildet die Differenz aus dem größten und kleinsten Wert."

- Aussage bezieht sich auf die Ränder der Verteilung, Ausreißer können diese stark beeinflussen.

+ Durch Betrachtung von Perzentilwerten, bspw. nur die mittleren 80% der Werte, kann der Einfluss von Ausreißern entgegengewirkt werden

Beispiel: Verteilung mit gleichem Mittelwert, jedoch unterschiedlicher Streuung.



Quelle: Raab-Steiner, Benesch, 2015, S. 99 -104.

DESKRIPTIVE STATISTIK – MAßE – SKALEN - KENNZAHLEN
LAGEMAßE UND STREUUNGSMAßE

- **Varianz**

Die Varianz ist die **durchschnittliche quadrierte Abweichung vom Mittelwert**. Sie kann nur für **intervallskalierte, normalverteilte Variablen** sinnvoll berechnet werden. Die Differenzen der einzelnen Messwerte vom Mittelwert sind in Summe Null. Eine große Abweichung erhält durch das Quadrieren mehr Gewicht.

- **Standardabweichung**

Die Standardabweichung ist ein Maß für die Streuung der Messwerte, sie ist die **Quadratwurzel aus der Varianz**. Sie hat die ursprüngliche Einheit der Variable.

$\sqrt{6{,}67}$ = 2,5826 cm

Bei kleiner Standardabweichung liegen alle Messwerte nahe am Mittelwert, bei großer hingegen weiter weg vom Mittelwert.

Beispiel Messwerte: 1, 3, 5, 7

1. Berechnung des arithmetischen Mittels
   => 16 / 4 = 4,
2. Subtraktion des arithmetischen Mittels von jedem einzelnen Messwert, quadrieren dieser Differenzen
   1-4 = -3 => quadriert 9
   3-4 = -1 => quadriert 1
   5-4 = 1 => quadriert 1
   7-4 = 3 => quadriert 9
3. Summierung der quadrierten Differenzen
   Summe => 20 (bspw. cm^2);
   Division durch Anzahl Messwerte -1
   => 20/(4-1) = 6,67 cm^2

Quelle: Raab-Steiner, Benesch, 2015, S. 99 -104.

Deskriptive Statistik

Übunsgfragen

1. Wozu dienen **deskriptivstatistische** Methoden?
2. Was ist eine **Kontingenztafel** und welche Zwecke kann sie erfüllen?
3. Welche **grafischen Darstellungsmöglichkeiten** der Daten wurden gezeigt und für welche Daten sind sie geeignet?
4. Welche **Lagemaße** kennt ihr? Bitte kurz beschreiben.
5. Welche **Streuungsmaße** kennt ihr? Bitte kurz beschreiben.

Deskriptive Statistik

20 Minuten Zeit

Aufgabe:

Erstellt einfache **deskriptive Statistiken** für Gebrauchtwagen.

1. Ladet die Daten für Gebrauchtwagen (BMW.csv).
2. Ermittelt das passende **Skalenniveau** pro Attribut (Nominal, Ordinal, Intervall, Ratio).
3. Berechnet das passende **Lagemaß** pro Attribut (Modus, Median, Mittelwert)
4. Berechnet das passende **Streumaß** pro Attribut (Spannweite, Varianz, Standardabweichung)
5. Erstellt ein **Balkendiagramm** (für ein beliebiges Attrtibut)
6. Erstellt ein **Histogramm** (für ein beliebiges Attrtibut)
7. Erstellt ein **Streudiagramm** (für zwei beliebige Attribute)
8. Erstellt einen **Boxplot** (für ein beliebiges Attribut)

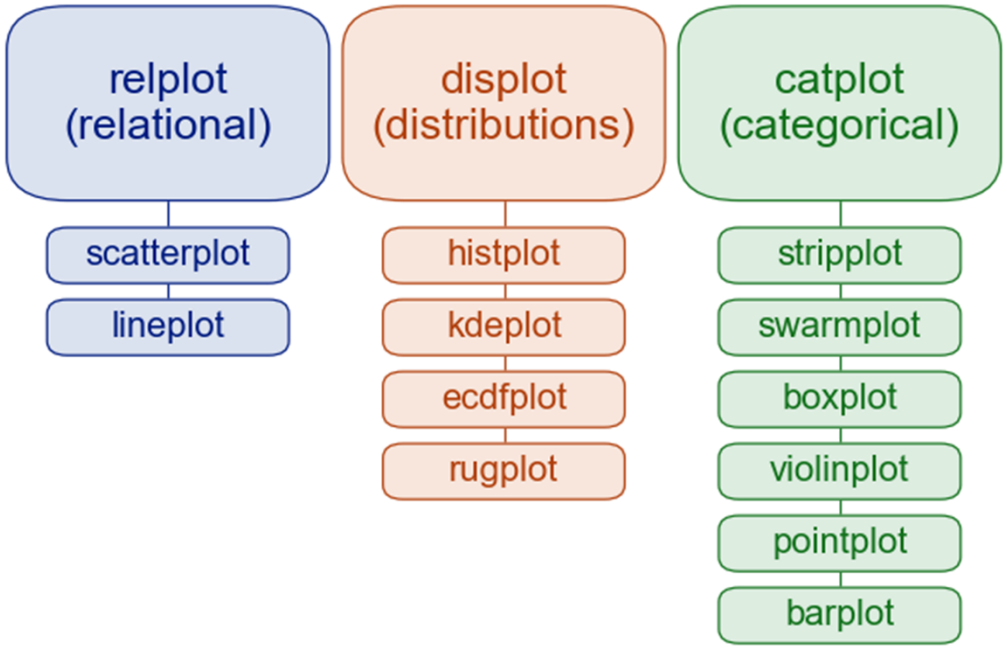
SEABORN - FIGURE LEVEL FUNCTIONS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

“Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.”



https://seaborn.pydata.org/tutorial/function_overview.html

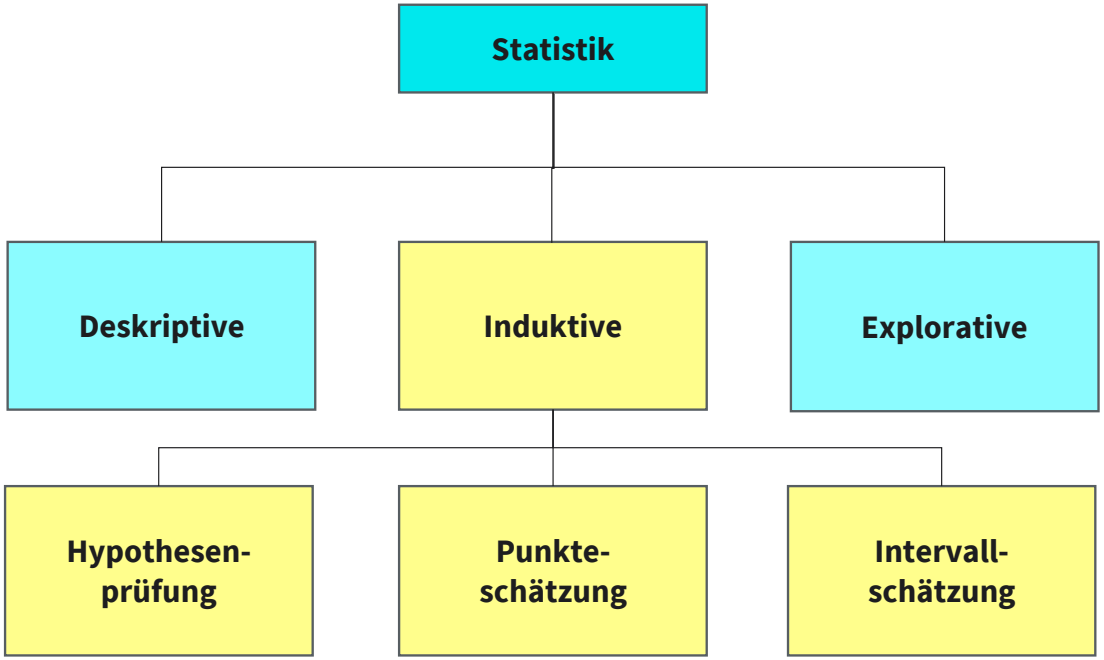
Statistische Grundlagen

Lernziele

Nach der Bearbeitung dieser Lektion werdet ihr wissen, …

- in welche **Teilbereiche** sich Statistik untergliedern lässt.
- was unter **deskripter Statistik** verstanden wird und welche **Visualisierungen**, **Maße** und **Kennzahlen** dafür verwendet werden.
- was unter **induktiver Statistik** verstanden wird, wie **Hypothesenpaare** aufgestellt werden und wie **statistische Tests** angewendet werden.
- was unter **explorativer Statistik** bzw. **explorativer Datenanalyse (EDA)** verstanden wird.
- Wie sich **univariate** , **bivariate** und **multivariate Analysemethoden** unterscheiden.

STATISTIK
ABGRENZUNG DER ARTEN



Rückschlüsse auf Grundgesamtheit durch Entnahme einer Zufallsstichprobe
Fungiert als Verbindung der deskriptiven Statistik und Wahrscheinlichkeitsrechnung

INDUKTIVE STATISTIK
SCHLÜSSE AUS DER STICHPROBE

Normalerweise ist eine deskriptive Beschreibung von Daten nicht das primäre Ziel einer Untersuchung, sondern von Aussagen über die Stichprobe auf die **Grundgesamtheit** zu schließen. Dazu werden Hypothesen über die Grundgesamtheit aufgestellt.

„**Statistische Hypothesen** werden stets als Hypothesen paar formuliert: Die sogenannte „**Nullhypothese**“ steht der „**Alternativhypothese**“ gegenüber, und es ist die Aufgabe der Signifikanztests, diese Hypothesen zu überprüfen.„ (Raab-Steiner, Benesch 2015, S. 108)

- **Nullhypothese** behauptet **es gibt keine Zusammenhänge** zwischen Gruppen oder Variablen
- **Alternativhypothese** behauptet **es gibt Zusammenhänge** zwischen Gruppen oder Variablen

„Mittels der **Inferenzstatistik** werden also konkurrierende Hypothesen, die Null- und die Alternativhypothese, geprüft.“(Raab-Steiner, Benesch 2015, S. 109)

- **Nullhypothese (H0):** „Der Preis für einen Gebrauchtwagen **ist nicht** vom Hersteller abhängig“
- **Alternativhypothese (H1):** „Der Preis für einen Gebrauchtwagen **ist** vom Hersteller abhängig“

Quelle: Raab-Steiner, Benesch, 2015, S. 108 f.

INDUKTIVE STATISTIK
STATISTISCHER TEST

Prüfung der **aufgestellten Hypothesen** auf Allgemeingültigkeit, über die untersuchte Stichprobe hinaus.

„Ein **statistischer Test** ist das Mittel, um diese **Prüfung auf Allgemeingültigkeit** vorzunehmen.“ (Raab-Steiner, Benesch 2015, S. 110)

Es gibt mehrere Arten statistischer Tests, jedoch bietet sich folgende Grundstruktur an:

1. Hypothesen formulieren und Untersuchungsdesign festlegen
2. Erhebung empirischer Daten (bspw. mittels Fragebogen)
3. Berechnung von deskriptiven Statistiken aus den Daten (bspw. Mittelwert)
4. Berechnung einer „Teststatistik“
5. Berechnung, wie wahrscheinlich die Teststatistik ist, unter der Annahme dass in der Population die Nullhypothese gilt
6. Sofern Wahrscheinlichkeit gering => „glaube“ an Alternativhypothese (signifikant)
   Sofern Wahrscheinlichkeit groß => „glaube“ an Hypothese (insignifikant)

Als Standardwerte haben sich hierfür Signifikanzniveaus von 5 %, 1% und 0,1 % (0,05; 0,01; 0,001) etabliert.

Quelle: Raab-Steiner, Benesch, 2015, S. 110.

INDUKTIVE STATISTIK
STATISTISCHER TEST

Es gibt mehrere Arten statistischer Tests, jedoch bietet sich folgende Grundstruktur an:

	Schritt	Beispiel
1	Hypothesen formulieren und Untersuchungsdesign festlegen	**Nullhypothese (H0):** „Der Preis für einen Gebrauchtwagen **ist nicht** vom Hersteller abhängig“ **Alternativhypothese (H1):** „Der Preis für einen Gebrauchtwagen **ist** vom Hersteller abhängig“
2	Erhebung empirischer Daten (bspw. mittels Fragebogen)	Datensatz zu Gebrauchtwagendaten
3	Berechnung von deskriptiven Statistiken aus den Daten (bspw. Mittelwert)	• Mittelwert Price Audi: 22.000 € • Mittelwert Price Ford: 12.000 €
4	Berechnung einer „Teststatistik“	• Subtrahieren der Mittelwerte, also 22 TSD – 12 TSD =10 TSD. • **Wenn Nullhypothese** gilt, dann sollte der **Wert nahe 0** sein (kein Unterschied), Ergebnis wird als **Auftretenswahrscheinlichkeit** errechnet (auch als „Signifikanz“ oder “p-Wert“ bezeichnet, von lat. probabilitas: Wahrscheinlichkeit) • Wenn diese **Wahrscheinlichkeit gering** ist, entscheidet man sich **für die Alternativhypothese**.
5	Berechnung, wie wahrscheinlich die Teststatistik ist, unter der Annahme dass in der Population die Nullhypothese gilt	„Der **P-Wert** ist die Wahrscheinlichkeit, mit der man sich irrt, wenn man die Nullhypothese ablehnt.“ (Sachs, 1999, S.188)
6	• Sofern Wahrscheinlichkeit gering => „glaube“ an **Alternativhypothese** (signifikant) • Sofern Wahrscheinlichkeit groß => „glaube“ an **Hypothese** (insignifikant)	Als Standardwerte haben sich hierfür **Signifikanzniveaus von 5 %, 1%** und **0,1 %** (0,05; 0,01; 0,001) etabliert.

Quelle: Raab-Steiner, Benesch, 2015, S. 110.

INDUKTIVE STATISTIK
FEHLER ERSTER UND ZWEITER ART

- Fehler erster Art (**Alpha Fehler**)
  wenn wir an einen Unterschied in der Population glauben, also die Alternativhypothese annehmen, obwohl sie in der Population nicht gilt.

- Fehler zweiter Art (**Beta Fehler**)
  wenn wir annehmen, es gäbe keinen Effekt in der Population, also die Nullhypothese beibehalten, obwohl sie in der Population nicht gilt.

- Konventionen:
  - Signifikanzniveau Alpha: Zwischen 0,1 % bis 5 %
  - Beta Fehlerniveau: Nicht größer als 20 %

Test / Wirklichkeit	Nullhypothese	Alternativhypothese
Nullhypothese	Korrekte Entscheidung	**Beta-Fehler**
Alternativhypothese	**Alpha-Fehler**	Korrekte Entscheidung

Quelle: Eigene Darstellung nach Raab-Steiner, Benesch 2015, S. 112.

Quelle: Raab-Steiner, Benesch, 2015, S. 112.

INDUKTIVE STATISTIK
STATISTISCHER TEST

Für die Wahl des statistischen Tests stellt man sich zunächst folgende Fragen:

1. Handelt es sich um **unabhängige** oder um **abhängige** Stichproben?
2. Möchte man **zwei** oder **mehr als zwei Stichproben** vergleichen?
3. Auf welchem **Skalenniveau** wurden die interessierenden Merkmale erhoben?

1. **Unabhängig** sind Stichproben, wenn sie unterschiedliche Objekte enthalten und bspw. verschieden groß sind. (bspw. Stichprobe für Männer und für Frauen)
2. „**Abhängig** sind Stichproben, wenn jeweils zwei oder mehrere Werte aus verschiedenen Stichproben eindeutig einander zugeordnet werden können.“
   (Raab-Steiner, Benesch, 2015, S. 115) (bspw. wiederholte Befragung nach 2 Wochen)

Quelle: Raab-Steiner, Benesch, 2015, S. 110.

INDUKTIVE STATISTIK
STATISTISCHE TESTS

Zur Wahl eines passenden **statistischenTests** ist die **Art der Abhängigkeit**, das **Skalenniveau** und das **Vorliegen von Normalverteilung** der interessierenden Variablen zu berücksichtigen.

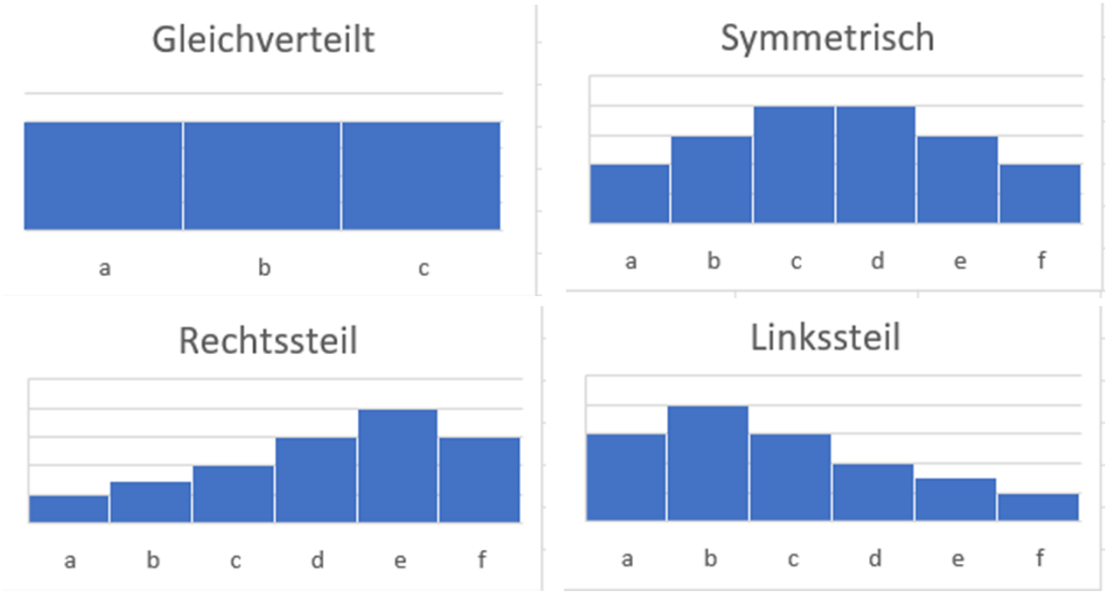
Anzahl der Stichproben	Art der Abhängigkeit	Skalenniveau	Normal-verteilung	Verfahren
2	unabhängig	metrisch	ja	T-Test für **unabhängige** Stichproben
2	abhängig	metrisch	ja	T-Test für **abhängige** Stichproben
2	unabhängig	ordinal	nein	U-Test nach Mann & Whitney
2	abhängig	ordinal	nein	Wilcoxon-Test
>2	unabhängig	metrisch	ja	Einfaktorielle Varianzanalyse
>2	abhängig	metrisch	ja	Einfaktorielle Varianzanalyse mit Messwiederholung
>2	abhängig	ordinal	nein	Friedman-Test

Quelle: Raab-Steiner, Benesch, 2015, S. 117.

INDUKTIVE STATISTIK – VERTEILUNGSVERLÄUFE

Gleichverteilung
alle Ausprägungen treten gleich häufig auf, die Verteilungskurve formt eine Waagerechte

Rechtssteile Verteilung
überwiegender Teil der Verteilungsfläche konzentriert sich rechtsseitig



Quelle: Eigene Darstellung nach Kähler, 2015, S. 27f.

Symmetrische Verteilung
Symmetrieachse, so dass sich die rechte und die linke Verteilungsfläche spiegelt

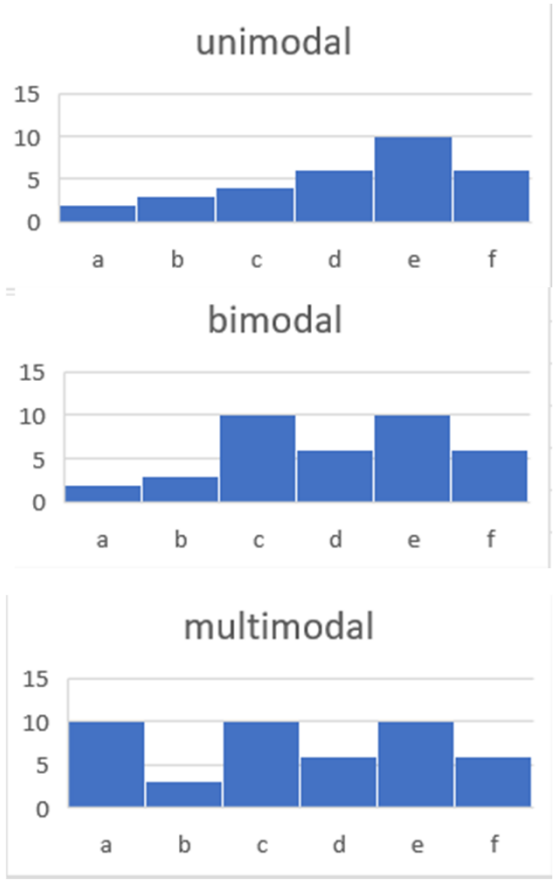
Linkssteile Verteilung
überwiegender Teil der Verteilungsfläche konzentriert sich linksseitig

Quelle: Kähler, 2015, S. 27f.

INDUKTIVE STATISTIK – VERTEILUNGSVERLÄUFE

Mit Blick auf **die Anzahl der vorliegenden Gipfel** ist eine Verteilung entweder:

- **Unimodal** (eingipflig)
- **Bimodal** (zweigipflig)
- **Multimodal** (mehrgipflig)



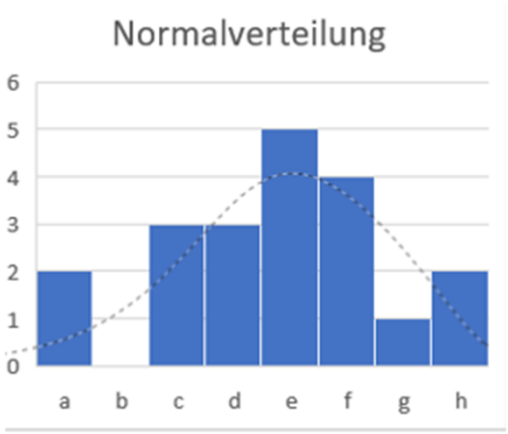
Quelle: Eigene Darstellung nach Kähler, 2015, S. 28.

Quelle: Kähler, 2015, S. 26f.

INDUKTIVE STATISTIK – VERTEILUNGSVERLÄUFE

„Die **Normalverteilung** ist eine mathematische Basisverteilung, von der sich andere theoretische Verteilungen ableiten. Sie ist dadurch charakterisiert, dass sie eingipflig und symmetrisch ist.“
(Raab-Steiner, Benesch, 2015, S. 95)

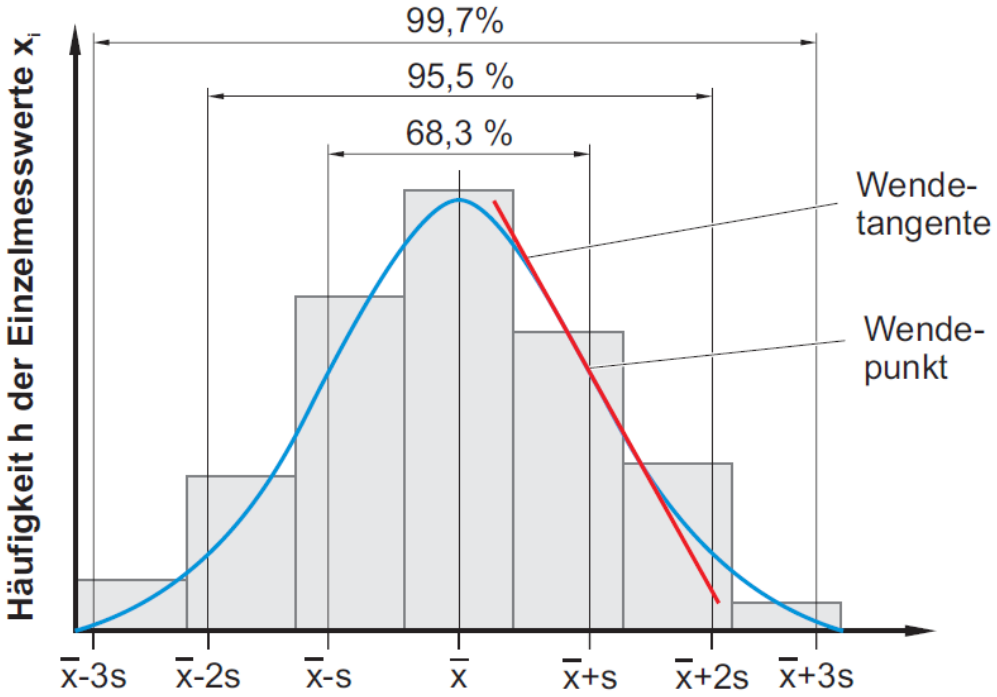
- In vielen **Grundgesamtheiten** der realen Welt haben Merkmale eine Verteilung, die gut durch die **Normalverteilung** approximiert werden kann (Kubinger, 2006, S 113.)
- Wenn die empirische Verteilung eines Merkmals nur geringfügig abweicht, spricht man von einem **„annähernd“ normalverteiltem** Merkmal
- „Bei der **Standardnormarlverteilung** liegt die Symmetrieachse bei dem Wert 0, so dass 0 als die Mitte angesehen werden kann.“ (Kähler, 2015, S. 30)



Quelle: Eigene Darstellung nach Raab-Steiner, Benesch, 2015, S.90.

Quelle: Raab-Steiner, Benesch, 2015, S. 117.

INDUKTIVE STATISTIK – KONFIDENZINTERVALL



$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad \mu = \lim_{n\to\infty} \bar{x}; \qquad s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}; \qquad \sigma = \lim_{n\to\infty} s$$

INDUKTIVE STATISTIK – STATISTISCHE TESTS

Prüfung auf Normalverteilung durch Kolmogorov-Smirnov-Test (KS-Test)

- Vergleich der **empirischen Verteilungsfunktion** mit der **theoretischen Normalverteilung**
- Verteilungsunabhängig
- Besonders für kleine Stichproben geeignet
- Kann **Abweichungen von der Normalverteilung** entdecken

Hypothesenpaar für den Test:
Nullhypothese:
„Die Stichprobe entstammt einer **normalverteilten Grundgesamtheit**."
Alternativhypothese:
„ Die Stichprobe entstammt **nicht** einer **normalverteilten Grundgesamtheit**."

$n \backslash \alpha$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$

Quelle: Kolmogorov-Smirnov Table | Real Statistics Using Excel (real-statistics.com)

Quelle: Raab-Steiner, Benesch, 2015, S. 119f.

INDUKTIVE STATISTIK
STATISTISCHE TESTS

Zur Wahl eines passenden **statistischenTests** ist die **Art der Abhängigkeit**, das **Skalenniveau** und das **Vorliegen von Normalverteilung** der interessierenden Variablen zu berücksichtigen.

Anzahl der Stichproben	Art der Abhängigkeit	Skalenniveau	Normal-verteilung	Verfahren
2	unabhängig	metrisch	ja	T-Test für **unabhängige** Stichproben
2	abhängig	metrisch	ja	T-Test für **abhängige** Stichproben
2	unabhängig	ordinal	nein	U-Test nach Mann & Whitney
2	abhängig	ordinal	nein	Wilcoxon-Test
>2	unabhängig	metrisch	ja	Einfaktorielle Varianzanalyse
>2	abhängig	metrisch	ja	Einfaktorielle Varianzanalyse mit Messwiederholung
>2	abhängig	ordinal	nein	Friedman-Test

Quelle: Raab-Steiner, Benesch, 2015, S. 117.

INDUKTIVE STATISTIK – STATISTISCHE TESTS

T-Test für unabhängige Stichproben

- Vergleicht die **Mittelwerte** zweier Stichproben
- Messwerte müssen **Normalverteilt** sein
- Die **Varianzen** dürfen sich **nicht signifikant unterscheiden**

Hypothesenpaar:

- Nullhypothese: „Der wahre **Mittelwert** der **Differenzen** ist **Null**."
- Alternativhypothese: „ Der wahre **Mittelwert** der **Differenzen** ist **ungleich Null**."

Quelle: Raab-Steiner, Benesch, 2015, S. 119.

INDUKTIVE STATISTIK – STATISTISCHE TESTS

U-Test nach Mann & Whitney

- Für Daten die **nicht mindestens intervallskaliert** sind
- **Normalverteilung keine** Voraussetzung
- Es werden keine Mittelwerte verglichen sondern **Rangplätze**
- Es werden alle Messwerte der betrachteten Gruppen in eine gemeinsame Rangreihe gebracht

Hypothesenpaar:

- Nullhypothese: „Die wahren mittleren **Rangplätze** zwischen den beiden Gruppen **unterscheiden sich nicht**.“
- Alternativhypothese: „Die wahren mittleren **Rangplätze** zwischen den beiden Gruppen **unterscheiden sich**.“

Quelle: Raab-Steiner, Benesch, 2015, S. 124.

INDUKTIVE STATISTIK – MANN-WHITNEY-U-TEST

	n_1																																							
n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
2		-	-	-	-	-	-	0	0	0	0	1	1	1	1	1	2	2	2	2	3	3	3	3	3	4	4	4	4	5	5	5	5	5	6	6	6	6	7	7
3			-	-	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	13	13	14	14	15	15	16	16	17	17	18	18
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	17	18	19	20	21	22	23	24	24	25	26	27	28	29	30	31	31
5					2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20	22	23	24	25	27	28	29	30	32	33	34	35	37	38	39	40	41	43	44	45
6						5	6	8	10	11	13	14	16	17	19	21	22	24	25	27	29	30	32	33	35	37	38	40	42	43	45	46	48	50	51	53	55	56	58	59
7							8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	50	52	54	56	58	60	62	64	66	68	70	72	74
8								13	15	17	19	22	24	26	29	31	34	36	38	41	43	45	48	50	53	55	57	60	62	65	67	69	72	74	77	79	81	84	86	89
9									17	20	23	26	28	31	34	37	39	42	45	48	50	53	56	59	62	64	67	70	73	76	78	81	84	87	89	92	95	98	101	103
10										23	26	29	33	36	39	42	45	48	52	55	58	61	64	67	71	74	77	80	83	87	90	93	96	99	103	106	109	112	115	119
11											30	33	37	40	44	47	51	55	58	62	65	69	73	76	80	83	87	90	94	98	101	105	108	112	116	119	123	127	130	134
12												37	41	45	49	53	57	61	65	69	73	77	81	85	89	93	97	101	105	109	113	117	121	125	129	133	137	141	145	149
13													45	50	54	59	63	67	72	76	80	85	89	94	98	102	107	111	116	120	125	129	133	138	142	147	151	156	160	165
14														55	59	64	69	74	78	83	88	93	98	102	107	112	117	122	127	131	136	141	146	151	156	161	165	170	175	180
15															64	70	75	80	85	90	96	101	106	111	117	122	127	132	138	143	148	153	159	164	169	174	180	185	190	196
16																75	81	86	92	98	103	109	115	120	126	132	137	143	149	154	160	166	171	177	183	188	194	200	206	211
17																	87	93	99	105	111	117	123	129	135	141	147	154	160	166	172	178	184	190	196	202	209	215	221	227
18																		99	106	112	119	125	132	138	145	151	158	164	171	177	184	190	197	203	210	216	223	230	236	243
19																			113	119	126	133	140	147	154	161	168	175	182	189	196	203	210	217	224	231	238	245	252	258
20																				127	134	141	149	156	163	171	178	186	193	200	208	215	222	230	237	245	252	259	267	274

Quelle: Wilcoxon-Mann-Whitney-Test – Wikipedia

INDUKTIVE STATISTIK – STATISTISCHE TESTS

Chi-Quadrat-Test

- Prüfung ob es signifikant auffällige Kombinationen in den Kategorien gibt
- Wird für **nominal skalierte Variablen** angewendet
- Der Vierfelder-Chi-Quadrat-Test setzt dichotome Variablen voraus

Hypothesenpaar:

- Nullhypothese: „Zwischen den beiden Variablen besteht **Unabhängigkeit**.“
- Alternativhypothese: „Zwischen den beiden Variablen besteht **Abhängigkeit**.“

Quelle: Raab-Steiner, Benesch, 2015, S. 117.

INDUKTIVE STATISTIK – KORRELATION

Korrelation liegt vor, wenn zwei Variablen zusammenhängen, so dass die Ausprägungen der einen Variablen die Ausprägungen der anderen Variablen mitbestimmt. (vgl. Raab-Steiner, Benesch, 2015, S. 135)

- **Positive Korrelation** liegt vor, wenn höhere Werte auf der x-Achse mit höheren Werten auf der y-Achse einhergehen.
- **Negative Korrelation** liegt vor, wenn ein höherer Wert auf der x-Achse mit einem niedrigeren Wert auf der y-Achse einhergeht.
- Meistens ist eine Korrelation nicht perfekt, sondern es gibt eine nicht unerhebliche **Variabilität**.
- Je nach Skalenniveau und Verteilungsform sind unterschiedliche **Korrelationsarten** anzuwenden.
- „Der **Korrelationskoeffizient** liegt im Bereich -1 bis +1 und drückt aus, wie stark ein Zusammenhang ist und in welche Richtung er geht. Je näher der Korrelationskoeffizient dem Betrag nach bei 1 liegt, desto stärker der Zusammenhang.“ (Raab-Steiner, Benesch, 2015, S. 137)

„Vom Bestehen einer **Korrelation** darf **nicht** automatisch auf ursächliche **Zusammenhänge** geschlossen werden!“ (Raab-Steiner, Benesch, 2015, S. 144)

Mögliche **Kausalzusammenhänge** können sehr komplex sein. So kann bspw. X=> Y direkt beeinflussen oder über eine andere Variable Z, die nicht betrachtet wurde.

Quelle: Raab-Steiner, Benesch, 2015, S. 135ff.

Induktive Statistik

Übunsgfragen

1. Erkläre den Unterschied zwischen **Deskriptiver** und **Induktiver Statistik**.
2. Erkläre die Begriffe „**Nullhypothese**“ und „**Alternativhypothese**“
3. Was versteht man unter einem **statistischen Test**?
4. Was versteht man unter einem „**Fehler erster Art“** und „**Fehler zweiter Art**“?

Induktive Statistik

20 Minuten Zeit

Aufgabe:

Erstellt einfache **induktive Statistiken** für Gebrauchtwagen.

1. Stellt eine **Nullhypothesen** und **Alternativhypothese** auf Basis der Daten zu Gebrauchtwagen auf.
2. Bestimmt das **Skalenniveau** der **Zielvariablen.**
3. Bestimmt, ob es sich um **abhängige** oder **unabhängig Stichproben** handelt**.**
4. **Testet** die **Stichproben** auf **Normalverteilung (Kolmogorov-Smirnov-Test).**
5. **Wählt** einen passenden **statistischen Test (siehe Tabelle)**.
6. Führt den **relevanten Test** aus und berechnet die Wahrscheinlichkeiten
7. Blickt auf das **Skalenniveau 5%**
8. Wird die **Nullhypothese** angenommen oder verworfen?

INDUKTIVE STATISTIK
STATISTISCHE TESTS

Zur Wahl eines passenden **statistischenTests** ist die **Art der Abhängigkeit**, das **Skalenniveau** und das **Vorliegen von Normalverteilung** der interessierenden Variablen zu berücksichtigen.

Anzahl der Stichproben	Art der Abhängigkeit	Skalenniveau	Normal-verteilung	Verfahren
2	unabhängig	metrisch	ja	T-Test für **unabhängige** Stichproben
2	abhängig	metrisch	ja	T-Test für **abhängige** Stichproben
2	unabhängig	ordinal	nein	U-Test nach Mann & Whitney
2	abhängig	ordinal	nein	Wilcoxon-Test
>2	unabhängig	metrisch	ja	Einfaktorielle Varianzanalyse
>2	abhängig	metrisch	ja	Einfaktorielle Varianzanalyse mit Messwiederholung
>2	abhängig	ordinal	nein	Friedman-Test

Quelle: Raab-Steiner, Benesch, 2015, S. 117.

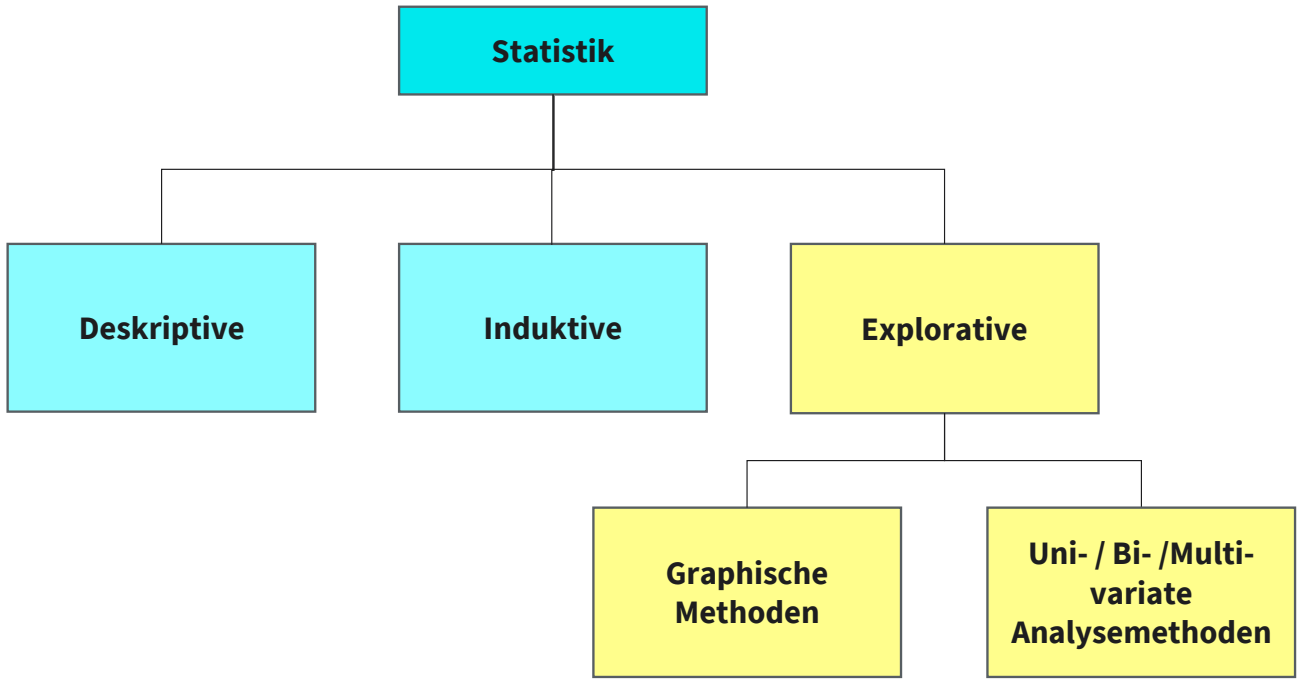
Statistische Grundlagen

Lernziele

Nach der Bearbeitung dieser Lektion werdet ihr wissen, …

- in welche **Teilbereiche** sich Statistik untergliedern lässt.
- was unter **deskripter Statistik** verstanden wird und welche **Visualisierungen**, **Maße** und **Kennzahlen** dafür verwendet werden.
- was unter **induktiver Statistik** verstanden wird, wie **Hypothesenpaare** aufgestellt werden und wie **statistische Tests** angewendet werden.
- was unter **explorativer Statistik** bzw. **explorativer Datenanalyse (EDA)** verstanden wird.
- Wie sich **univariate** , **bivariate** und **multivariate Analysemethoden** unterscheiden.

STATISTIK
ABGRENZUNG DER ARTEN



Erkundung von Zusammenhängen in Daten (Einsatz graphischer, strukturentdeckender und strukturprüfender Methoden, häufige Verwendung im Bereich Data Science / Data Mining

EXPLORATIVE STATISTIK / EXPLORATIVE DATENANALYSE (EDA)

„Die deskriptive Datenanalyse hat den Zweck, die in einer Stichprobe gefundenen Daten mit Hilfe von Kennwerten zu beschreiben und grafisch oder tabellarisch darzustellen. Bei dieser Darstellung von Daten geht es um einzelne Variablen und ihre Ausprägungen.
In **der explorativen Datenanalyse** gehen wir nun einen Schritt weiter und versuchen, **mit Hilfe von geeigneten Darstellungen und Berechnungen** die Daten nach **Mustern** oder **Zusammenhängen** zu untersuchen. Daher auch der Begriff „explorativ" – wir forschen (explorieren) in den Daten nach interessanten Informationen, die man bei der einfachen Betrachtung in der deskriptiven Analyse nicht auf den ersten Blick sehen kann.
(Schäfer, 2010, Explorative Datenanalyse)

EDA wird häufig im Bereich „Data Science" und „Data Mining" eingesetzt („Data Understanding").

- Überblick über den Datensatz bekommen (mittels Betrachtung der Charakteristika wie Skalenniveau, statistische Verteilung und deskriptiver Statistiken)
- Attribute / Features bewerten und Visualisieren (bspw. mittels Histogramm, Boxplots, Streudiagramm)
- Datenqualität evaluieren (Anomalien, Ausreißer, Duplikate, fehlende Werte)
- Zusammenhänge zwischen den Attributen / Features erkennen (Korrelation, Heatmaps)

Schäfer, T. (2010). Explorative Datenanalyse: Muster und Zusammenhänge erkennen. In: Statistik I. VS Verlag für Sozialwissenschaften.

EXPLORATIVE DATENANALYSE (EDA)
UNI- / BI- /MULTI-VARIATE ANALYSEMETHODEN (1/2)

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

1. Univariate Analyse

Erkundung **eines Attributs** und **dessen Eigenschaften**

- Histogramm (statistische Verteilung)
- Balkendiagramm (bei kategorialen Attributen)
- Statistische Kennzahlen (Mittelwert, Median, Modus, Spannweite, Varianz, Standardabweichung)
- Box-plot (Streuung und Ausreißerkennung)

2. Bivariate Analyse

Erkundung von **zwei Attributen** und deren **Zusammenhang- und Abhängigkeitsstruktur**.

- Scatter plot (Visualisierung) oder Liniendiagramm (bspw. über die Zeitdimension)
- Korrelationskoeffizient oder Kovarianz (Stärke des Zusammenhangs, linear)
- Kreuztabelle/Kontingenztafel (kategorial)

.

EXPLORATIVE DATENANALYSE (EDA)
UNI- / BI- /MULTI-VARIATE ANALYSEMETHODEN (2/2)

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

3. Multivariate Analyse

Erkundung von **mehreren Attributen zugleich** und Aufdeckung von **Zusammenhangs- und Abhängigkeitsstrukturen**

„**Strukturentdeckende Verfahren** sind solche multivariaten Verfahren, deren Ziel in der Entdeckung von Zusammenhängen zwischen Variablen oder zwischen Objekten liegt.“(Backhaus, et. al., 2011, S. 14)

- Zu Beginn der Analyse noch keine Vorstellung darüber, welche Zusammenhänge existieren
- Verfahren die dem Bereich zugeordnet werden:
    - Faktoranalyse, Clusteranalyse, Multidimensionale Skalierung, Korrespondenzanalyse, Künstliche Neuronale Netze

„**Strukturprüfende Verfahren** sind solche multivariaten Verfahren, deren primäres Ziel in der Überprüfung von Zusammenhängen zwischen Variablen liegt.“(Backhaus, et. al., 2011, S. 13)

- Kausale Abhängigkeit einer interessierenden Variablen von einer oder mehreren „unabhängigen“ Variablen
- Bereits Vorstellungen über Zusammenhänge vorhanden, welche mithilfe von Verfahren überprüft werden sollen
- Verfahren die dem Bereich zugeordnet werden:
    - Lineare und nichtlineare Regressionsanalyse, Zeitreihenanalyse, Varianzanalyse, Diskriminanzanalyse, Kontingenzanalyse, Logistische Regression, Strukturgleichungsmodelle, Conjoint-Analyse

Quellen: Backhaus et. al, 2011, S- 11, Runkler 2015, S12-22.

.

Explorative Statistik

Übunsgfragen

1. Was wird mit dem Begriff **explorative Statistik** bzw. **explorative Datenanalyse (EDA)** verbunden?
2. Was versteht man unter einer **univariaten Analyse**? Nenne ein **Beispiel**!
3. Was versteht man unter einer **bivariaten Analyse**? Nenne ein **Beispiel**!
4. Was versteht man unter einer **multivariaten Analyse**? Nenne ein **Beispiel**!
5. Wie unterscheiden sich **strukturentdeckende Verfahren** von **strukturprüfenden Verfahren**?

Explorative Statistik

20 Minuten Zeit

Aufgabe:

Erstellt einfache **explorative Statistiken** für Gebrauchtwagen.

1. Verschafft euch einen Überblick mittels Betrachtung der Charakteristika wie **Skalenniveau**, **statistische Verteilung** und **deskriptiver Statistiken**
2. Schaut euch ausgewählte Attribute / Features anhand geeigneter Visualisierungen an (bspw. mittels **Histogramm**, **Boxplots**, **Streudiagramm**)
3. Evaluiert die **Datenqualität** auf **Anomalien**, **Ausreißer**, **Duplikate**, **fehlende Werte**.
4. Ermittelt erste Zusammenhänge zwischen den Attributen / Features anhand einer **Heatmap.**