TECHNICAL APPLICATIONS AND DATA MANAGEMENT. WS 2021/2022.

VORLESUNG 3

28.09.2021

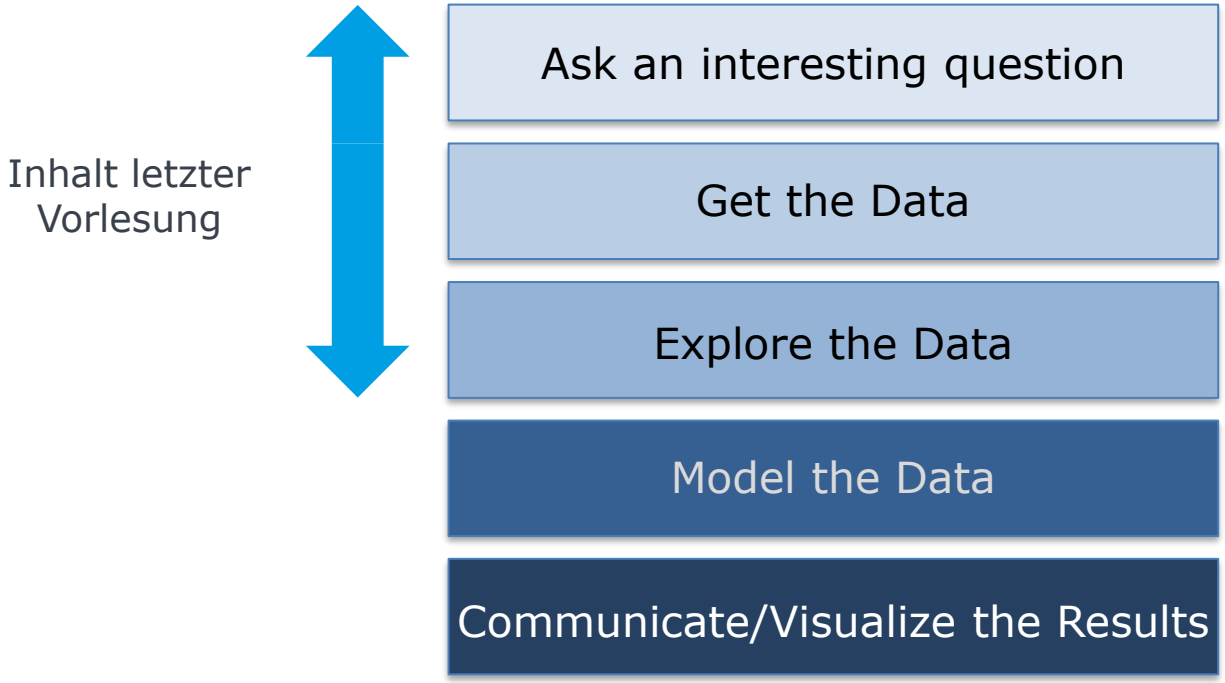
MÜNCHEN

STUDIENGANG DIGITAL MANAGEMENT.

AGENDA

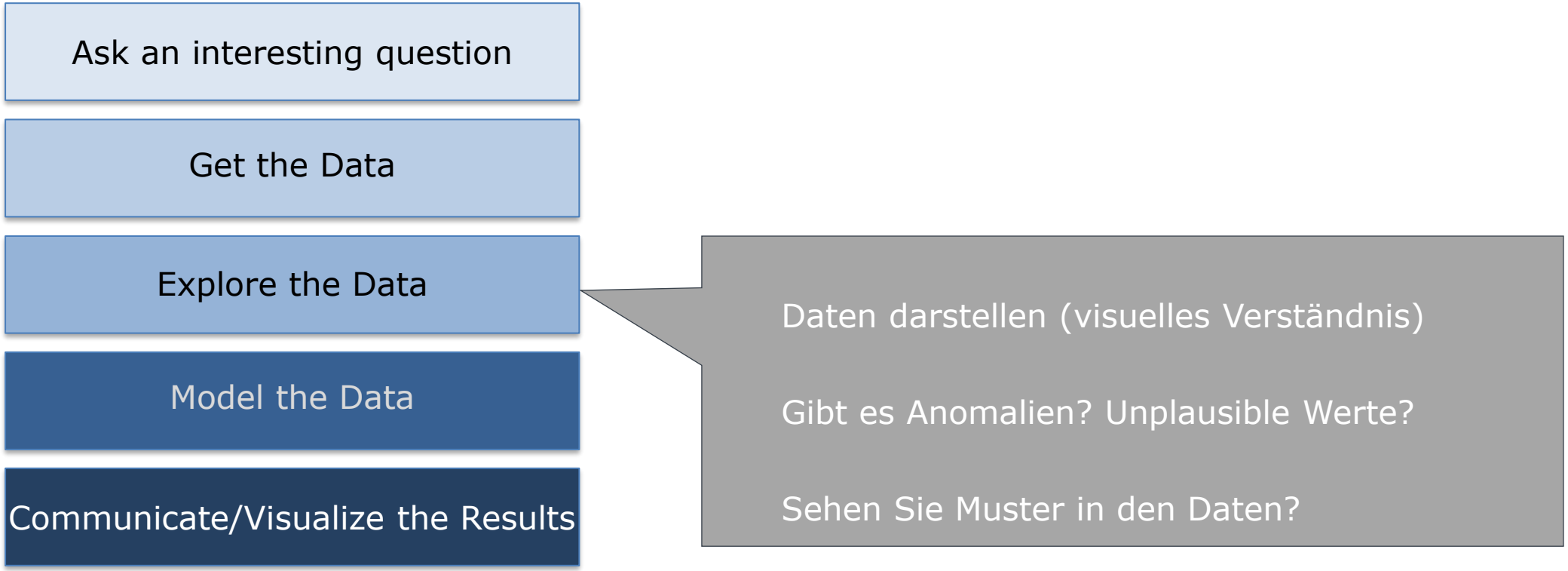
1. Grundlagen Stochastik
    1. Wahrscheinlichkeitstheorie
    2. Deskriptive Statistik
    3. Explorative Statistik
2. Ausblick: Inferenzstatistik

RÜCKBLICK AUF LETZTE WOCHE: VORGEHENSWEISE DATA SCIENCE ANGESEHEN.



Quelle: Protopapas, Rader, Tanner, CS109 Data Science, 2020, Link

FOKUS DER HEUTIGEN VORLESUNG.



Wie kann ich Daten visuell darstellen? Wie erkenne ich Anomalien oder unplausible Werte?

Quelle: Protopapas, Rader, Tanner, CS109 Data Science, 2020, Link

WAS MACHEN WIR HEUTE?

- (vereinfachte und reduzierte) Grundlagen Stochastik
- Einsatz stochastischer Methoden
- Explorative Datenanalyse inkl. Visualisierung

1.1 WAHRSCHEINLICHKEITSTHEORIE

WAS IST STOCHASTIK?

Stochastik[1] besteht aus folgenden Teilgebieten:

- Wahrscheinlichkeitstheorie: mathematische Erfassung und Analyse von zufälligen (nicht-deterministischen) Ereignissen
- Mathematische Statistik[2]:
  - Deskriptive Statistik: Daten durch Graphiken oder Tabellen visuell beschreiben.
  - Explorative Statistik[3]: Zusammenhänge/ Muster zwischen Daten finden und bewerten, Entdecken von Hypothesen
  - Inferenzstatistik: aus einzelnen Eigenschaften einer Menge Eigenschaften über Gesamtmenge ableiten, Hypothesen testen

„Lies, damned lies, and statistics“
(Mark Twain)

1 Ratekunst, von στοχαστικὴ τέχνη
2 einordnen, von στατίζω
3 Begriff wurde geprägt von John Tukey 1977 in seinem Buch “Exploratory Data Analysis”

MOTIVATION WAHRSCHEINLICHKEITSRECHNUNG.

- Eine Münze wird geworfen: Welche Seite zeigt nach oben?
- Familie will Mitte August Grillen bei Sonnenschein. Kann sie die Wetterdaten der letzten Jahre nutzen, für ein gutes Datum?
- Roulette-Spielen in einer Spielbank: auf was sollte ich setzen?

Wahrscheinlichkeitsrechnung bietet uns mathematische Methoden für die Beantwortung solcher Fragestellungen.

GRUNDBEGRIFFE WAHRSCHEINLICHKEITSRECHNUNG AM FALLBEISPIEL MÜNZWURF.

Ergebnismenge: Menge aller möglichen Ergebnisse, z.B. **Ω = {Kopf, Zahl}.**
Eine **endliche Menge** wird als **diskret** bezeichnet**,** eine **nicht abzählbare Menge** als **kontinuierlich** (beispielsweise Zeit)**.**

Ereignis: auftretendes Element oder Teilmenge aus der Ergebnismenge, z.B. E:= Kopf geworfen

Definition **relative Häufigkeit von E**: $\frac{\text{relative Häufigkeit Ereignis E}}{\text{Anzahl aller Ereignisse}}$

Wir setzen die relative Häufigkeit Ereignis E gleich der Wahrscheinlichkeit E[1]. Dann können wir folgende Regeln definieren:

1. Pr[gesamte Ergebnismenge Ω] = 1
2. Pr[leere Menge ∅] = 0
3. 0 ≤ Pr[Ereignis E] ≤ 1
4. Pr[$\overline{\text{Ereignis E}}$] = 1 - Pr[Ereignis E] (Gegenwahrscheinlichkeit)
5. Pr[A∩ B] = Anzahl der gemeinsamen eingetretenen Ereignisse A und B
6. Pr[Ereignis A ∪ Ereignis B] = Pr[A] + Pr[B] – Pr[A∩ B] (A oder B trat auf)

Mit diesen 6 Regeln können wir diskrete und kontinuierliche Wahrscheinlichkeiten berechnen

1 Begründung hierfür erfolgt im späteren Verlauf dieser Vorlesung

EINFÜHRUNG WAHRSCHEINLICHKEITSRECHNUNG: BEDINGTE WAHRSCHEINLICHKEIT.

Die Wahrscheinlichkeit eines Ereignisses A kann sich ändern, wenn wir wissen, daß ein anderes Ereignis B schon geschah.

$$\mathbf{Pr}[\mathbf{A} \mid \mathbf{B}] = \frac{\mathbf{Pr}[\mathrm{A} \cap \mathrm{B}]}{\mathbf{Pr}[\boldsymbol{B}]}$$

Sprich: Wahrscheinlichkeit von A gegeben Evidenz B

Der Wert von Pr[B] „normalisiert" Pr[A|B], das heißt er passt die Wahrscheinlichkeit von A an die von B an.

Beispiele:

- Wie hoch ist die Wahrscheinlichkeit daß mindestens eine 3 gewürfelt wurde, falls eine ungerade Zahl gewürfelt wurde?
  Menge A = {3,4,5,6}, Menge B = {1,3,5}. Schnittmenge A und B = {3,5}. ➔ Pr[A|B] = $^{2/6}/_{3/6} = \frac{2}{3}$ = 66%

- Titanic: Wie hoch ist die Chance, daß ein Passagier Mann ist und überlebt?
  Anzahl überlebender Männer = 161, Anzahl männliche Passagiere = 843 ➔ Pr[A|B] = 161/843 = 19%

Die bedingte Wahrscheinlichkeit hilft bei der Untersuchung, wie stark ein Ereignis Einfluß auf ein anderes hat.

EINFÜHRUNG WAHRSCHEINLICHKEITSRECHNUNG: UNABHÄNGIGE VS. ABHÄNGIGE EREIGNISSE

Zwei (oder mehr) Ereignisse A, B sind statistisch unabhängig, falls ein Eintreten von A ein Eintreten von B nicht beeinflußt

$$\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A] * \mathbf{Pr}[B]$$
$$\mathbf{Pr}[A \mid B] = \mathbf{Pr}[A]$$

Sprich: Evidenz von B ändert nicht die Wahrscheinlichkeit von A

Beispiele:

- In einer Schublade sind 5 paar schwarze Socken und 4 Paar weiße Socken. Sie ziehen 2 Paar Socken
  a. mit Zurücklegen in die Schublade (ordentlich!). Unabhängig?
  b. Ohne Zurücklegen und auf den Boden. Unabhängig?
- Titanic
  a. Überlebensrate Mann und seine Passagierklasse.
  b. Überlebenschance eines Passagiers und die Anzahl der Musiker in der Bordkapelle.

Prüfen Sie immer, ob Ereignisse voneinander abhängig sind (Correlation does not imply causation!!)

EINFÜHRUNG WAHRSCHEINLICHKEITSRECHNUNG: ZUFALLSVARIABLEN (RANDOM VARIABLE).

Zufallsvariablen ermöglichen, Ereignisse zu quantifizieren auch ohne Kenntnisse der gesamten Verteilung.

Beispiele:

- Eine Münze wird 3 mal geworfen. Y bezeichnet die Anzahl der Würfe mit Ergebnis „Kopf".
- Wir stehen an der Autobahn A9 und machen eine Verkehrszählung der LKW.
- Wir wählen zufällige Passagiere der Titanic und zählen mit X die Anzahl der Frauen.

Zufallsvariablen ermöglichen dann die Berechnungen der Wahrscheinlichkeit, bspw. höchstens 2 mal Kopf in 3 Würfen:
$Pr[X \leq 2] = Pr[X=0 \text{ Kopf geworfen}] + Pr[X=1 \text{ Kopf geworfen}] + Pr[X=2 \text{ Kopf geworfen}] = 1/8 + 3/8 + 3/8 = 7/8$

Der **Erwartungswert** definiert das Ergebnis, das die Zufallsvariable im Mittel (nach vielen Durchführungen) annimmt.

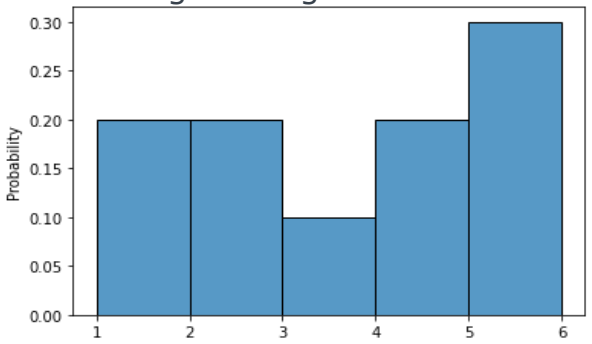
Die **Varianz** definiert die Streuung der Zufallsvariablen um den Erwartungswert (mehr dazu im nächsten Kapitel).

Wichtig ist beim Einsatz von Zufallsvariablen genügend oft zu messen („Sampling")!

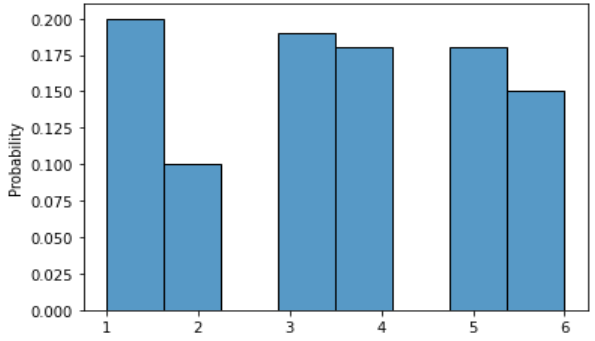
EINFÜHRUNG WAHRSCHEINLICHKEITSRECHNUNG: WAS IST GENÜGEND OFT MESSEN- ODER DAS GESETZ DER GROßEN ZAHLEN.

Wir messen mit den Zufallsvariablen X_1, …, X_6 wie oft bei einem Würfel Auge 1,…,6 gewürfelt wird.
Dabei interessiert uns, wie sich die relative Häufigkeit über die Anzahl der Würfe ändert.

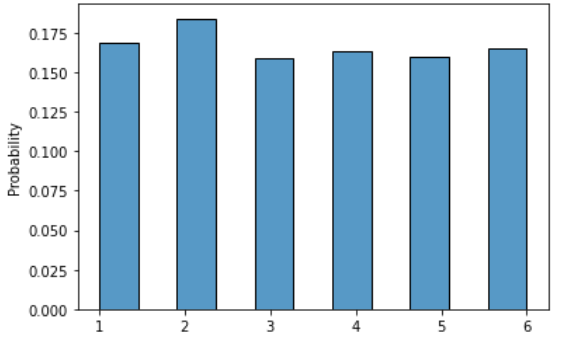
Rel. Häufigkeit Augen bei 10 Würfen



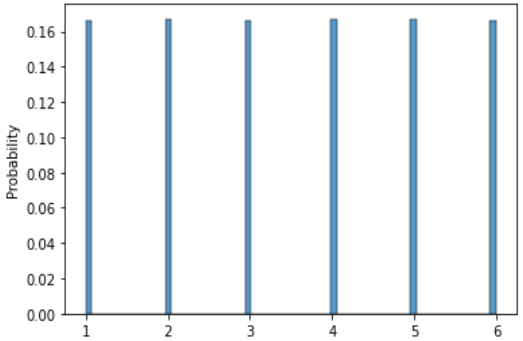
100 Würfen



1000 Würfen



500‘000 Würfen



Gesetz der großen Zahlen: die relative Häufigkeit eines Ereignisses E nähert sich für hinreichend viele Wiederholungen seiner Wahrscheinlichkeit an.

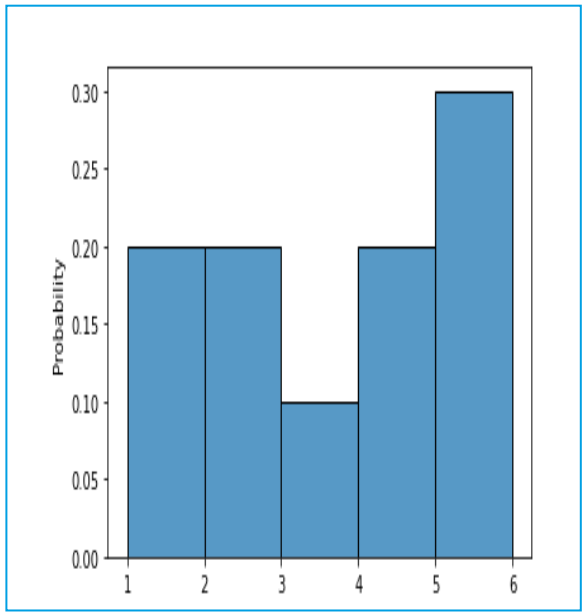
Die Ergebnisse von Zufallsvariablen sind **nur dann** belastbar,
falls sie einer genügend großen Menge an Versuchen zugrunde liegen!!!

Quelle Bilder: Visualisierung Ergebnisse eines eigenen Python-Skripts

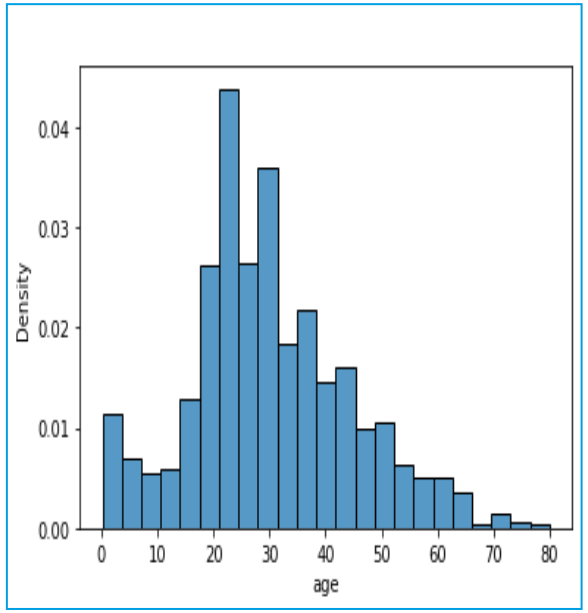
1.2 DESKRIPTIVE STATISTIK

DESKRIPTIVE STATISTIK. BEISPIELE.

Diskrete Verteilung

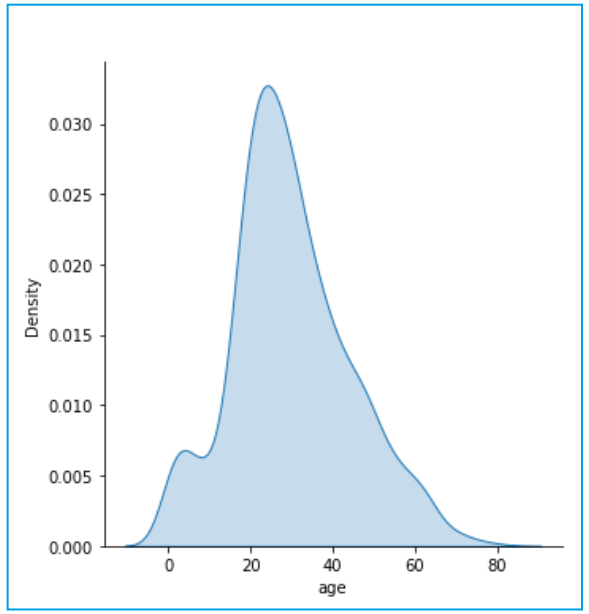


Rel. Häufigkeit Augen eines Würfels bei 10 Würfen. Ergebnismenge = {1,…,6}

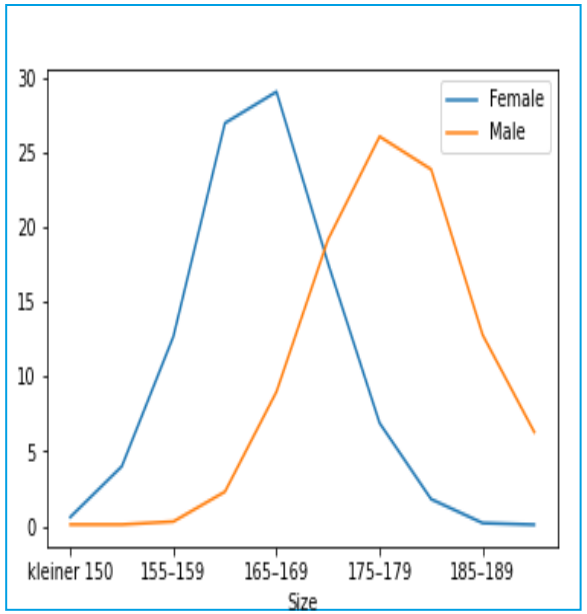


Diskretisierte Altersverteilung der Passagiere der Titanic. Ergebnismenge in 23 „Körbe"

Kontinuierliche Verteilung



Kontinuierliche Altersverteilung der Passagiere der Titanic. Ergebnismenge = $\mathbb{R}$



Größenverteilung Einwohner Deutschland in 2006[1] Ergebnismenge = $\mathbb{R}$

Bei diskreten, endlichen Variablen sprechen wir von einer Wahrscheinlichkeitsfunktion, bei kontinuierlichen, „nicht-endlichen" Variablen von einer Dichtefunktion.

Quellen: Bilder 1-3 eigene Darstellung, Bild 4 abgerufen unter Link

DESKRIPTIVE STATISTIK: ÜBERSICHT WICHTIGSTE PARAMETER.

Lageparameter

- **Mean**: Mittelwert.
- **Median**: teilt Verteilung in 2 genau gleich große Hälften. Stabiler gegenüber Extremwerten als Mean.
- **Modus**: häufigster Wert der Verteilung.
- **Min**: kleinster Wert der Verteilung
- **Max**: größter Wert der Verteilung
- **P-Quantil**: Schwellenwert, der größer als p in % Elemente der Verteilung ist.

Streuungsparameter

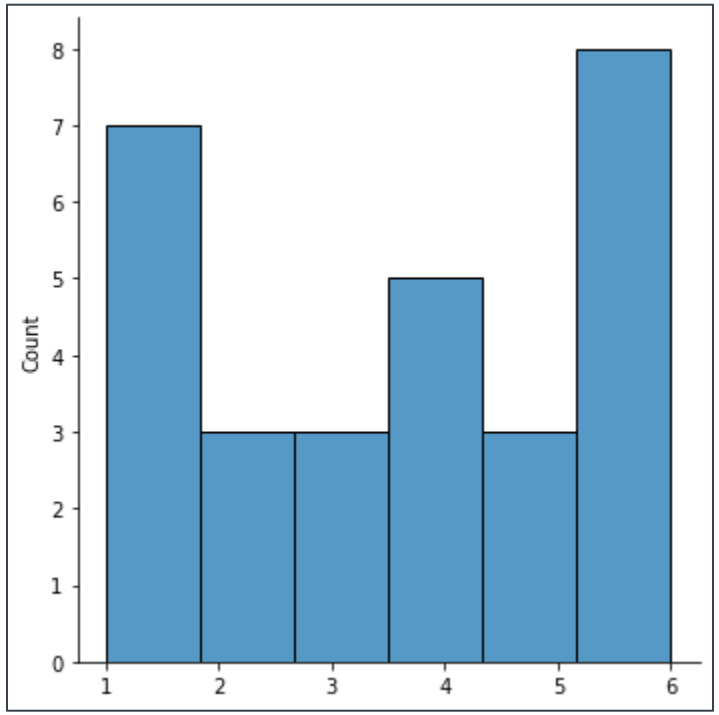
- **Spannweite**: Abstand Min und Max-Wert
- **Varianz**: (quadratische) Abweichung Werte vom Mittelwert. Basis für Standardabweich.
- **Standardabweichung**: durchschnittliche Abweichung/Streuung Werte um Mittelwert
- **Schiefe**: beschreibt Assymetrie Verteilung. Bei Rechtsschief sind häufiger Werte kleiner als Mittelwert, bei linksschief größer.
- **Wölbung**: Verteilungen mit geringer Wölbung streuen gleichmäßig; hohe W. bedeutet extremere, seltenere Ergebnisse.

Zusammenhangsparam.

Spätere Vorlesung

Parameter ermöglichen eine komprimierte Erfassung einer Verteilung.

DETAILLIERUNG LAGEPARAMETER.



Ergebnisse Würfeln

Mean = 3.62

Median: 4

Modus: 6 ist häufigstes Ergebnis

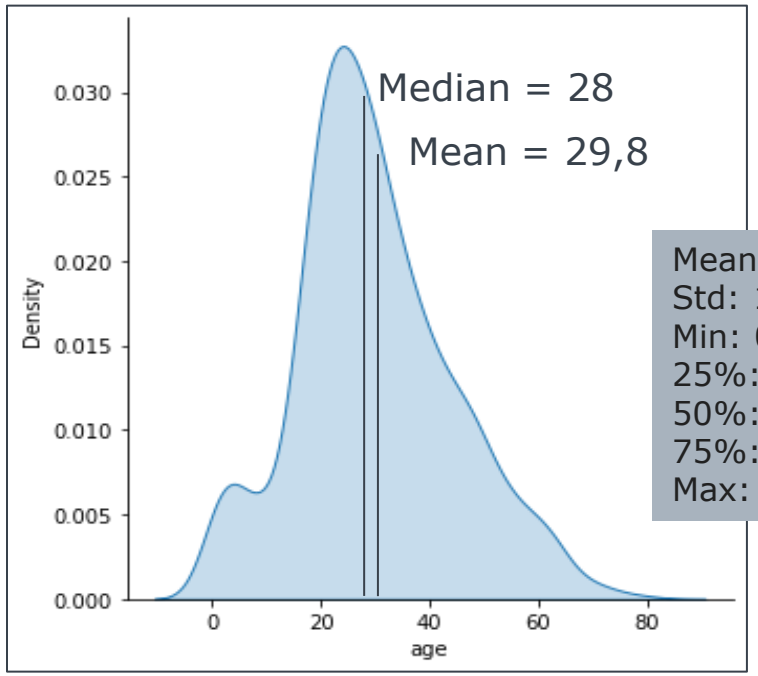
Min: 1 ist niedrigster Ergebniswert

Max: 6 ist höchster Ergebniswert

P-Quantil:

- 25% = 2 (7 von 30 Ergebnissen kleiner als 2)
- 50% = Median
- 75% = 6 (21 von 30 Ergebnissen kleiner als 6)

DETAILLIERUNG STREUUNGSPARAMETER.



Mean: 29.881138
Std: 14.413493
Min: 0.170000
25%: 21.000000
50%: 28.000000
75%: 39.000000
Max: 80.000000

Altersverteilung Titanic-Passagiere

Spannweite: 80 Jahre – 0,29 Jahre = 79,71 Jahre

Varianz: 207.55

Standardabweichung:14,41 → weite Streuung Alter

Schiefe: rechtsschief, da Median kleiner als Mean.

Mehr als 50% der Passagiere jünger als Durchschnittsalter.

Wölbung: geringe Wölbung, gleichmäßige Streuung.

Quelle Bild: Visualisierung Ergebnisse Titanic-Notebook

1.3 EXPLORATIVE STATISTIK

EXPLORATIVE STATISTIK: WAS MACHEN WIR DA?

- Daten aufbereiten und säubern:
  - Ersetzen von Nullwerten oder fehlende Werte (Data Imputation).
  - Entfernen von Duplikaten.
- Prüfen, ob Features relevant für die Hypothesen sind und ggf. Entfernen Features (Dimensionsreduktion).
- Entdecken von Ausreißern/ Anomalien in Features (Beispiel: Menschen mit Größe von 2,40 Meter oder mehr).
- Entdecken von Mustern in den Daten (Beispiel: gegenseitige Abhängigkeiten von Features wie Einkommen und Wohnort).
- Bilden von Hypothesen (Beispiel: „In der 1. Klasse auf der Titanic war die Überlebenschance am höchsten“).

Ziel der Explorativen Statistik ist das Visualisieren von Daten um daraus Hypothesen oder Annahmen abzuleiten.

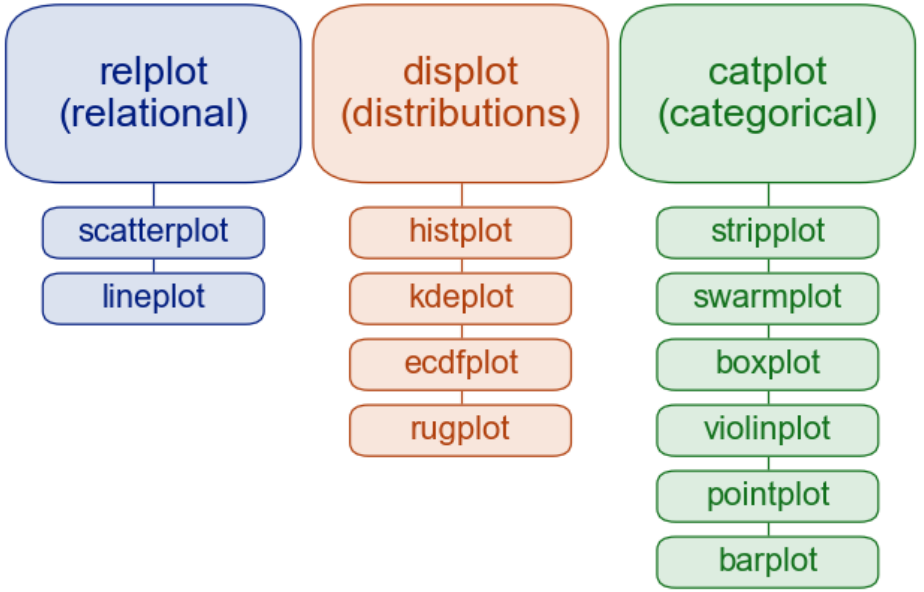
VISUALISIERUNG DATEN: ÜBERSICHT.

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

Was für Features werden geplottet?

- Zahlen
  - diskrete Werte: abzählbare Werte wie Ganzzahlen.
  - kontinuierliche Werte: nicht abzählbare Werte wie reelle Zahlen.
- kategorische Variablen: Variablen mit einem Wert aus einer definierten Menge (bspw. Farben: rot, grün, ...).

Was für Plots gibt es?

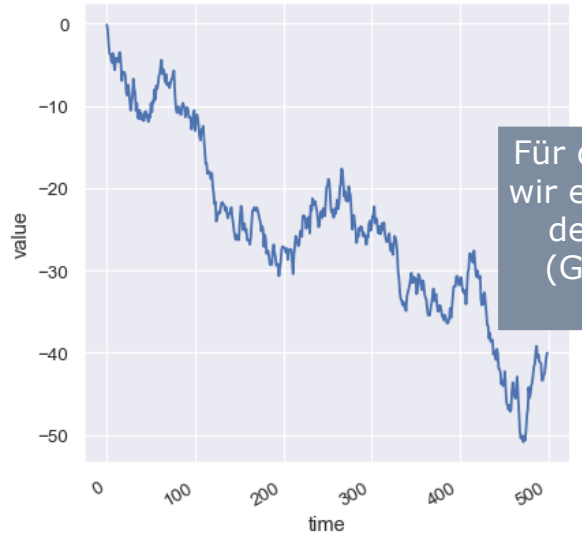


Die verschiedenen Plots unterscheiden sich, der Programmieraufruf ist aber prinzipiell gleich.

Quelle: Seaborn Dokumentation, verfügbar unter: Link

VISUALISIERUNG DATEN: RELATIONAL PLOTS.

Line Plots

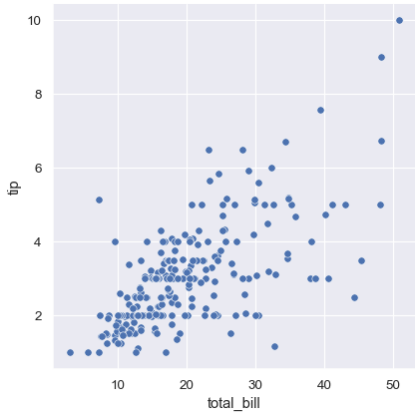


Für die x- und y-Achse nehmen wir ein Feature des Datensatzes der bei Data angegeben ist (Groß- und Kleinschreibung Feature beachten!)

```
sns.relplot(x="time",
            y="value",
            kind="line",
            data=df)
```

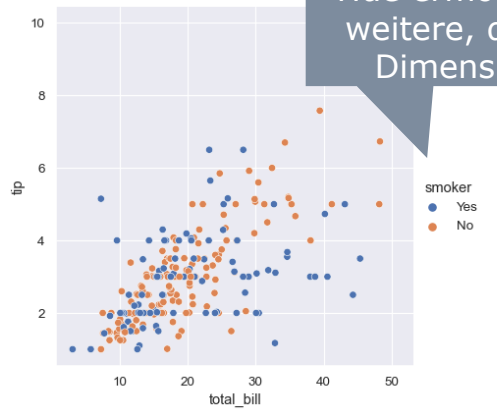
Ziel: Visualisierung von Änderungen über Zeit

Scatter-Plots




```
sns.relplot(x="total_bill",
            y="tip",
            data=tips)
```

Hue ermöglicht weitere, dritte Dimension



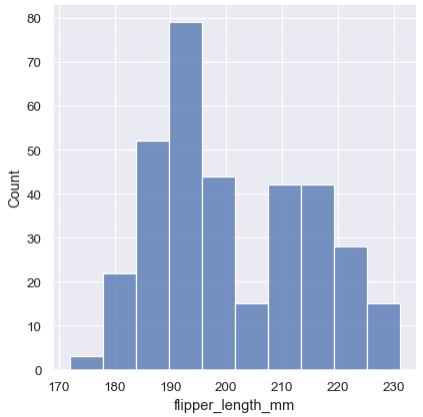

```
sns.relplot(x="total_bill",
            y="tip",
            hue="smoker",
            data=tips)
```

Ziel: Entdecken von Beziehungen zwischen 2 Features

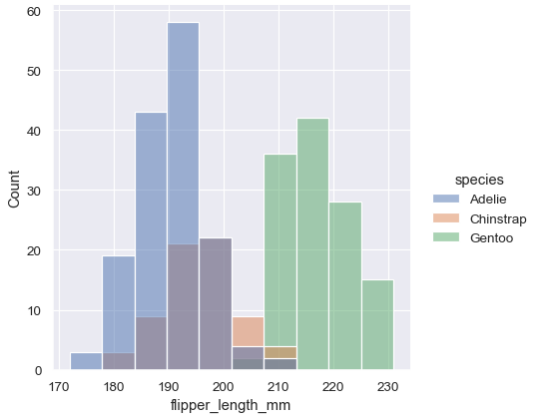
Quelle: Seaborn Dokumentation, verfügbar unter: Link

VISUALISIERUNG DATEN: VERTEILUNGEN.

Histogram (Histplot)




```
sns.displot(penguins,
x="flipper_length_mm")
```



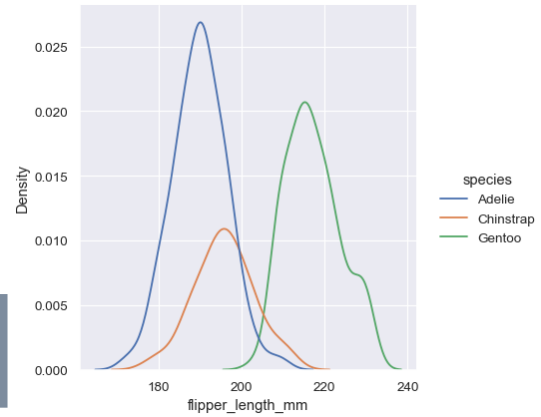

```
sns.displot(penguins,
            x="flipper_length_mm",
            hue="species")
```

Ziel: Entdecken Anomalien, Tendenzen, Verteilungen.
Aber: keine Visualisierung für kontinuierliche Features!

KDEPlot (Kernel density estimation)

Zusätzlicher Parameter Kind bei KDE

```
sns.displot(penguins,
            x="flipper_length_mm",
            kind="kde")
```



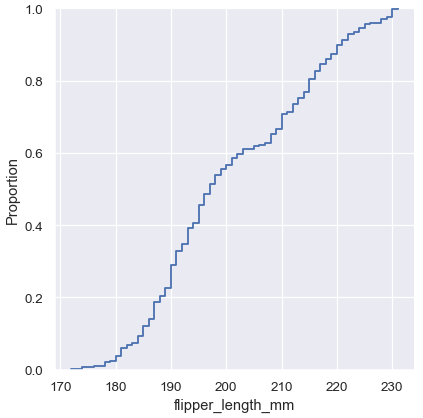

```
sns.displot(penguins,
            x="flipper_length",
            hue="species",
            kind="kde")
```

Ziel: Histogram für kontinuierliche Features.
Aber: Interpolation Zwischenwerte, kann falsch sein!

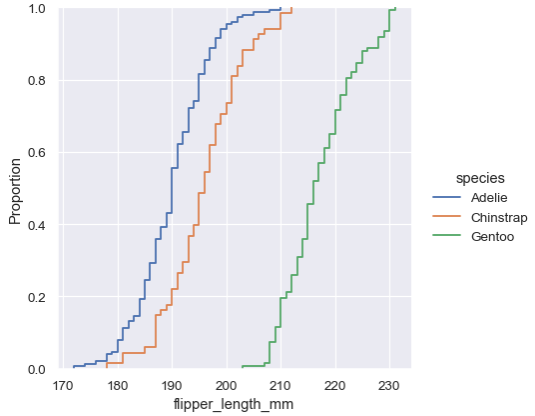
Quelle: Seaborn Dokumentation, verfügbar unter: Link

VISUALISIERUNG DATEN: VERTEILUNGEN.

Empirical cumulative distributions (ECDF)



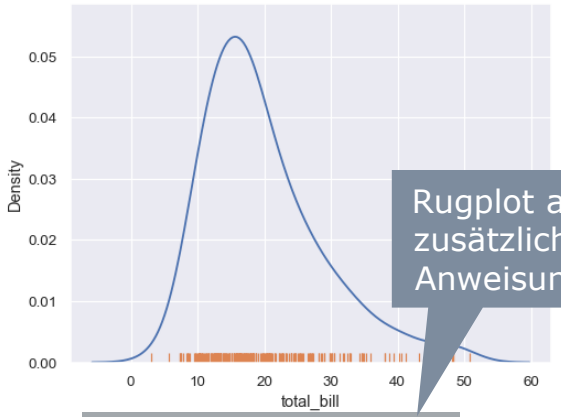

```
sns.displot(penguins,
x="flipper_length_mm",
kind="ecdf")
```




```
sns.displot(penguins,
x="flipper_length_mm",
hue="species",
kind="ecdf")
```

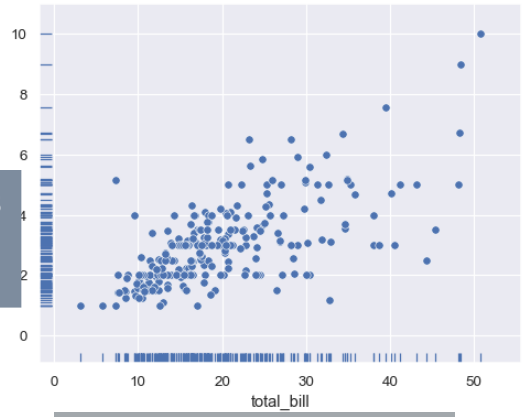
Abbilden jedes Wertes in Plot (Treppenfunktion).
Aber: weniger intuitiv.

Rugplots



Rugplot als zusätzliche Anweisung

```
sns.kdeplot(data=tips,
x="total_bill")
sns.rugplot(data=tips,
x="total_bill")
```



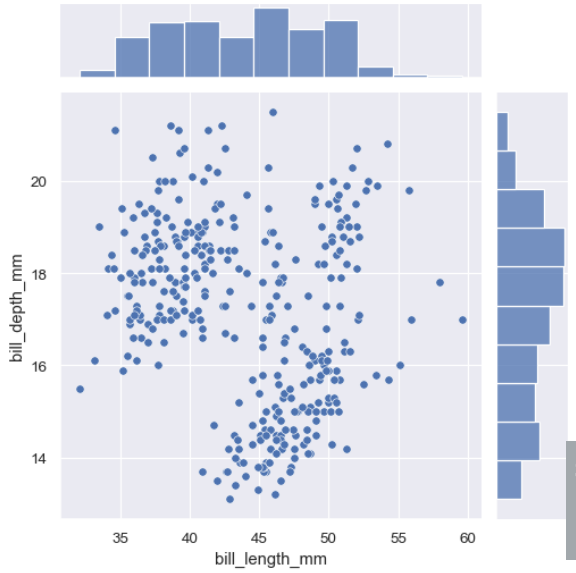

```
sns.scatterplot(data=tips,
x="total_bill", y="tip")
sns.rugplot(data=tips,
x="total_bill", y="tip")
```

Ziel: Aufzeigen Verteilung einer Variablen als zusätzliches Element in einem Plot. Aber: wird wenig genutzt

Quelle: Seaborn Dokumentation, verfügbar unter: Link

VISUALISIERUNG DATEN: WEITERE DISTRIBUTION PLOTS.

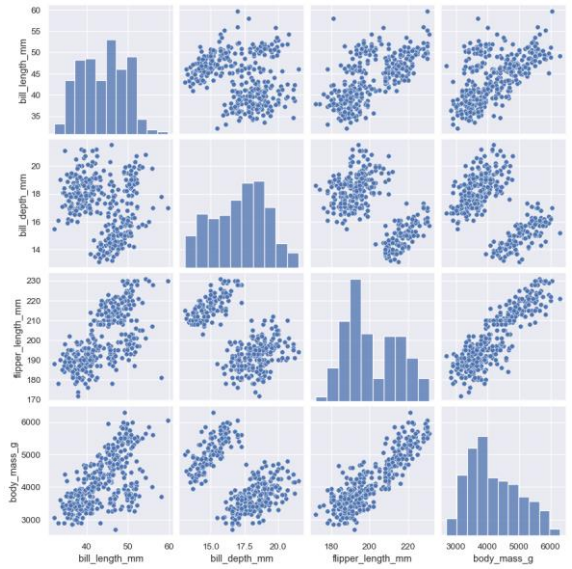
Joint-Plot




```
sns.jointplot(data=penguins,
              x="bill_length_mm",
              y="bill_depth_mm")
```

Ziel: Kombination von 2 verschiedenen Plots für Erkennen der Verteilung von Variablen und Beziehungen

Pairplot



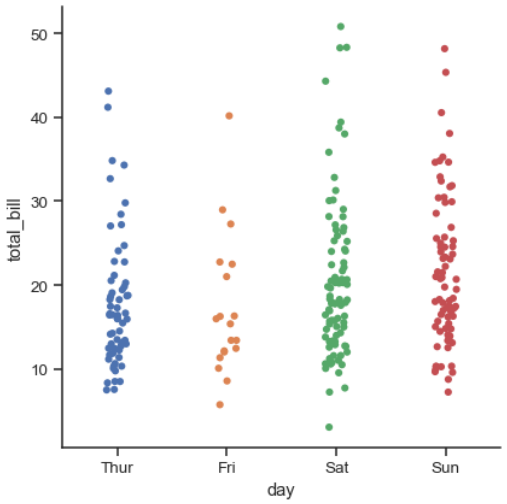

```
sns.pairplot(penguins)
```

Ziel: Entdecken von Beziehungen der Features zueinander

Quelle: Seaborn Dokumentation, verfügbar unter: Link

VISUALISIERUNG DATEN: KATEGORISCHE PLOTS.

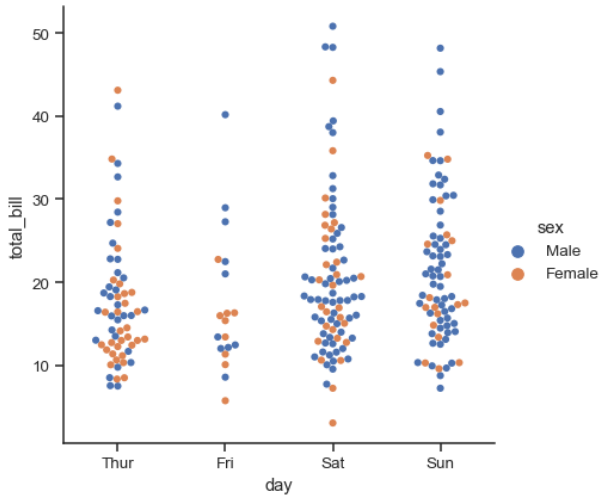
Kategorischer Scatterplot (Stripplot)




```
sns.catplot(x="day",
            y="total_bill",
            data=tips)
```

Ziel: Scatterplot für kategorische Variablen
Aber: eingeschränkte Sicht, da Punkte überlappen.

Kategorischer Scatterplot (Swarmplot)



Zusätzlicher Parameter Kind für Swarmplot

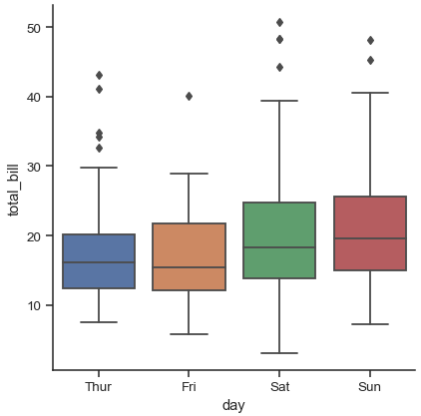
```
sns.catplot(x="day",
            y="total_bill",
            hue="sex",
            kind="swarm"
            data=tips)
```

Ziel: Verbessern Sichtbarkeit bei überlappenden Werten.
Aber: nur für kleine Datensätze.

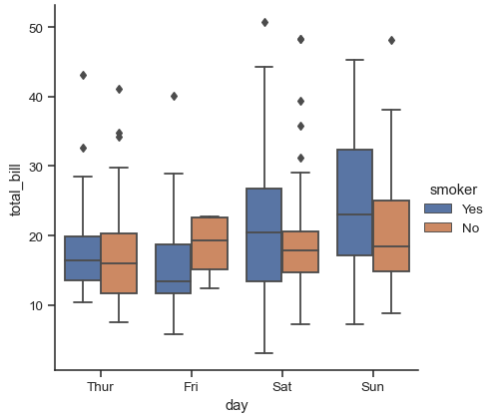
Quelle: Seaborn Dokumentation, verfügbar unter: Link

VISUALISIERUNG DATEN: KATEGORISCHE PLOTS.

Kategorischer Verteilungsplot (Boxplot)



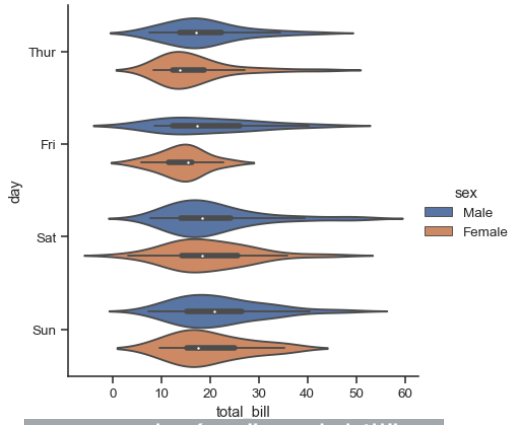

```
sns.catplot(x="day",
            y="total_bill",
            kind="box",
            data=tips)
```



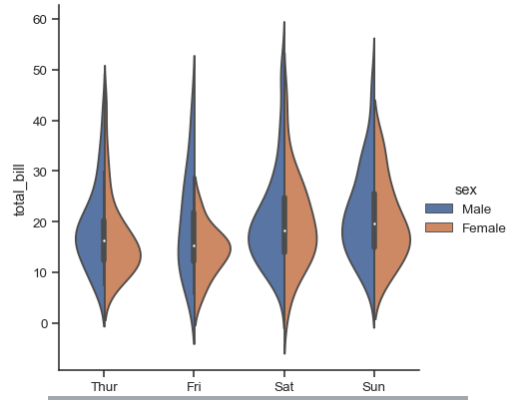

```
sns.catplot(x="day",
            y="total_bill",
            hue="smoker",
            kind="box",
            data=tips)
```

Ziel: Entdecken Anomalien, Tendenzen, Verteilungen.
Aber: keine Visualisierung für kontinuierliche Features!

Kategorischer Verteilungsplot (Violinplot)




```
sns.catplot(x="total_bill",
            y="day",
            hue="sex",
            kind="violin",
            data=tips)
```



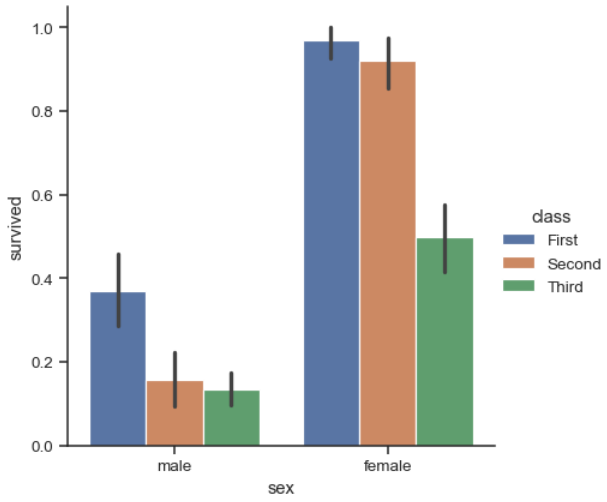

```
sns.catplot(x="day",
            y="total_bill",
            hue="sex",
            kind="violin",
            split=True,
            data=tips)
```

Ziel: Histogram für kontinuierliche Features.
Aber: Interpolation Zwischenwerte, kann falsch sein!

Quelle: Seaborn Dokumentation, verfügbar unter: Link

VISUALISIERUNG DATEN: KATEGORISCHE PLOTS.

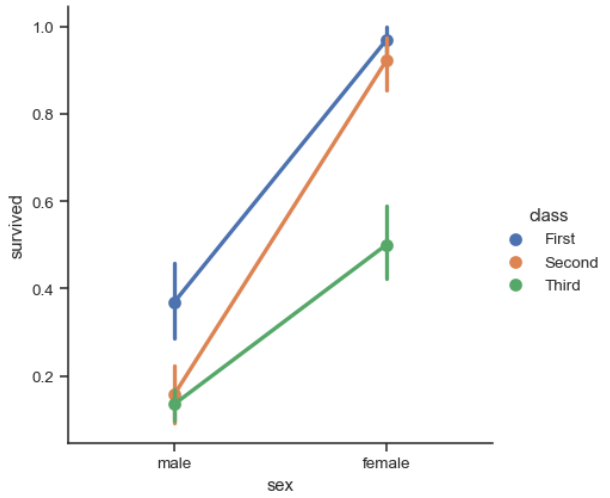
Statistische Abschätzung (Barplots)




```
sns.catplot(x="sex",
            y="survived",
            hue="class",
            kind="bar",
            data=titanic)
```

Ziel: Aufzeigen von Tendenzen.

Statistische Abschätzung (Pointplot)



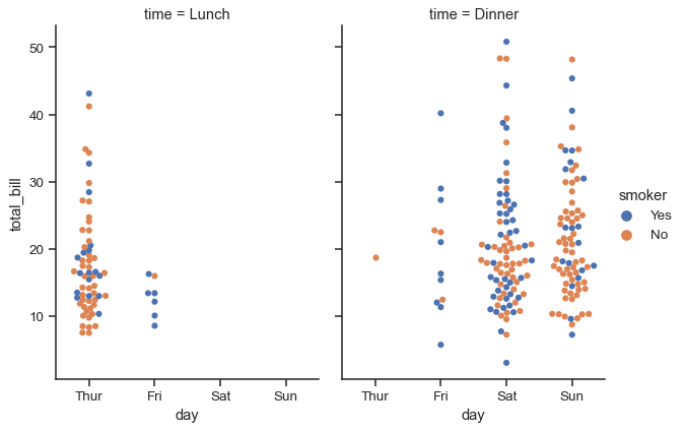

```
sns.catplot(x="sex",
            y="survived",
            hue="class",
            kind="point",
            data=titanic)
```

Ziel: Aufzeigen von Tendenzen

Quelle: Seaborn Dokumentation, verfügbar unter: Link

VISUALISIERUNG DATEN: WEITERE KATEGORISCHE PLOTS.

Visualisierung verschiedener Features.




```
sns.catplot(x="day",
            y="total_bill",
            hue="smoker",
            col="time",
            kind="swarm",
            data=tips)
```

Ziel: Aufzeigen von Tendenzen.

Quelle: Seaborn Dokumentation, verfügbar unter: Link

EXPLORATIVE STATISTIK: FALLBEISPIEL IN GRUPPENARBEIT.

Amazon: 50 bestselling novels on Amazon each year from 2009 to 2020.

Datensatz verfügbar unter: Link

IMDB: Top 1000 Filme auf IMBDB

Datensatz verfügbar unter: Link

EXPLORATIVE STATISTIK: FALLBEISPIEL IN GRUPPENARBEIT

1. Erstellen Sie ein Notebook in Google Colab.
2. Laden Sie den Datensatz in das Notebook mit Pandas.read_csv().
3. Wenden Sie die gelernten deskriptiven Statistik-Methoden auf den Datensatz an (Tip: Pandas-Describe Funktion) und beschreiben Sie die Ergebnisse.
4. Plotten Sie für jede der vorgestellten Plot-Kategorien je ein Beispiel.
5. Leiten Sie aus den Plots Hypothesen oder Ergebnisse ab.
6. Können Sie die Hypothesen durch weitere Analysen bestätigen oder widerlegen?
7. Sind die Ergebnisse statistisch belastbar?
8. Stellen Sie Ihre Ergebnisse und Hypothesen vor.

ZUSAMMENFASSUNG DER HEUTIGEN VORLESUNG.

- Grundlagen der Wahrscheinlichkeitsrechnung
- deskriptive Statistik zur visuellen Beschreibung und Analyse Daten
- Explorative Statistik zur Identifikation Muster und Zusammenhänge
- Anwendung deskriptive und explorative Statistik anhand eines Fallbeispiels.

Damit können wir schon viele Data Science Fragen beantworten

LITERATUR UND WEITERE QUELLEN (AUSZUG).

Statistik:

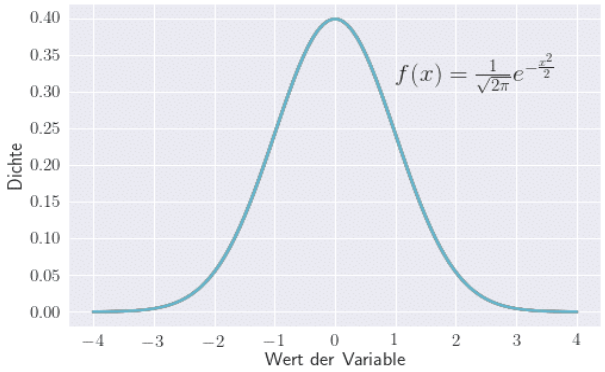
- Schickinger, Steger: Diskrete Strukturen 2 – Wahrscheinlichkeitstheorie und Statistik.
- James, Witten, Hastie and Tibshirani: An Introduction to Statistical Learning (freies Ebuch unter: Link)
- Spiegelhalter: The Art of Statistics: Learning from Data
- Witte: Statistics (10th Edition)
- Silver: The Signal and the noise
- Taleb: Black Swan
- Huff: How to Lie with Statistics
- Wheelan: Naked statistics

Kostenfreie Online-Kurse (bei Interesse):

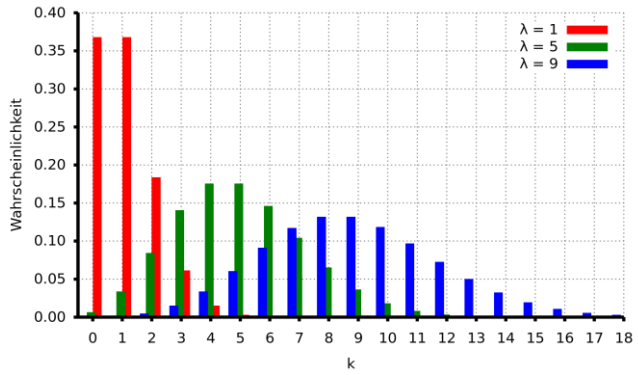
- Khan Academy für Statistics (Link oder Link)
- Data Science mit Excel (Link)
- Python-Kurse
  - Python for Everybody (Link)
  - Udacity Python Course (Link)
  - Kaggle Courses:
    - Python (Link)
    - Python Library Pandas (Link)
    - Python Data Visualization (Link)

BACKUP

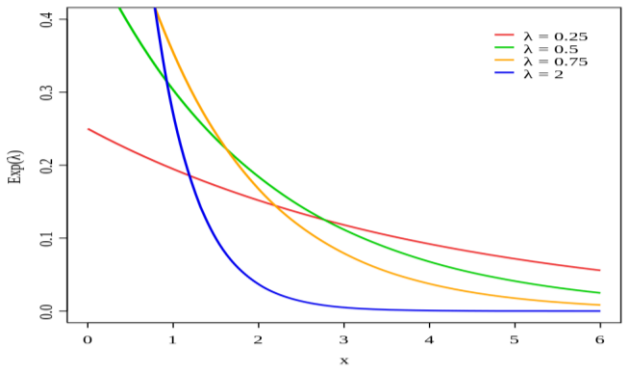
ÜBERSICHT WICHTIGER VERTEILUNGEN.



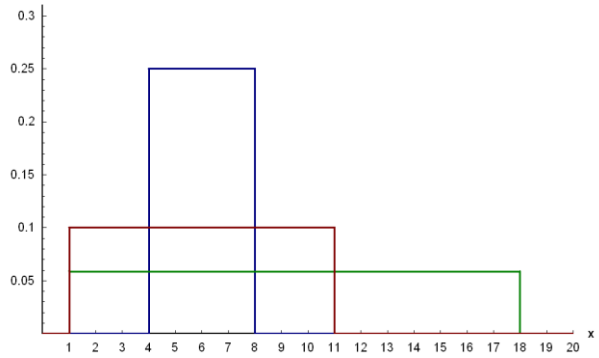
Normalverteilung:
Modellierung vieler natürlicher und statistischer Prozesse.



Poisson-Verteilung:
Modellierung Ereignisse, die bei konstanter mittlerer Rate unabhängig voneinander in einem festen Zeitintervall oder räumlichen Gebiet eintritt.



Expontentialverteilung:
Modellierung von Zeitintervallen.



Gleichverteilung:
jeder Wert ist gleich wahrscheinlich (konstanter y-Wert).