

TECHNICAL APPLICATIONS AND DATA MANAGEMENT. WS 2022.

VORLESUNGEN 3 UND 4

27.09.2022

MÜNCHEN

STUDIENGANG
SUSTAINABILITY
MANAGEMENT &
LEADERSHIP SOWIE
MEDIEN &
KOMMUNIKATION

GEPLANTE ROADMAP VORLESUNG.

ROADMAP	WAS HABEN WIR VOR?
Vorlesung 1	Workflow Data Management, Datentypen und Datenqualität
Vorlesung 2	Einführung Data Science und Data Science Workflow, Grundlagen Data Management
Vorlesung 3 und Vorlesung 4	Deskriptive und explorative Datenanalyse und Vertiefung anhand Case Study
	Vertiefung Datenanalyse anhand Case Study
Vorlesung 5	Aufgabenstellung Data Science, Übersicht und Einführung Machine Learning, unüberwachtes Lernen
Vorlesung 6	Überwachtes Lernen
Vorlesung 7	Schulterblick 1 und Vertiefung überwachtes Lernen anhand Case Study
Vorlesung 8	Neuronale Netze und Convolutional Neural Networks (CNN)
Vorlesung 9	Vertiefung CNN anhand Case Study, Aufgabenstellung AI
Vorlesung 10	Schulterblick 2
Vorlesung 11	Übersicht Rekurrente Neuronale Netze
Vorlesung 12	Schulterblick 3
Vorlesung 13	Ausblick zukünftige AI-Themen, „Fragestunde“
Vorlesung 14	Präsentation Ergebnisse

Zusammenlegen?

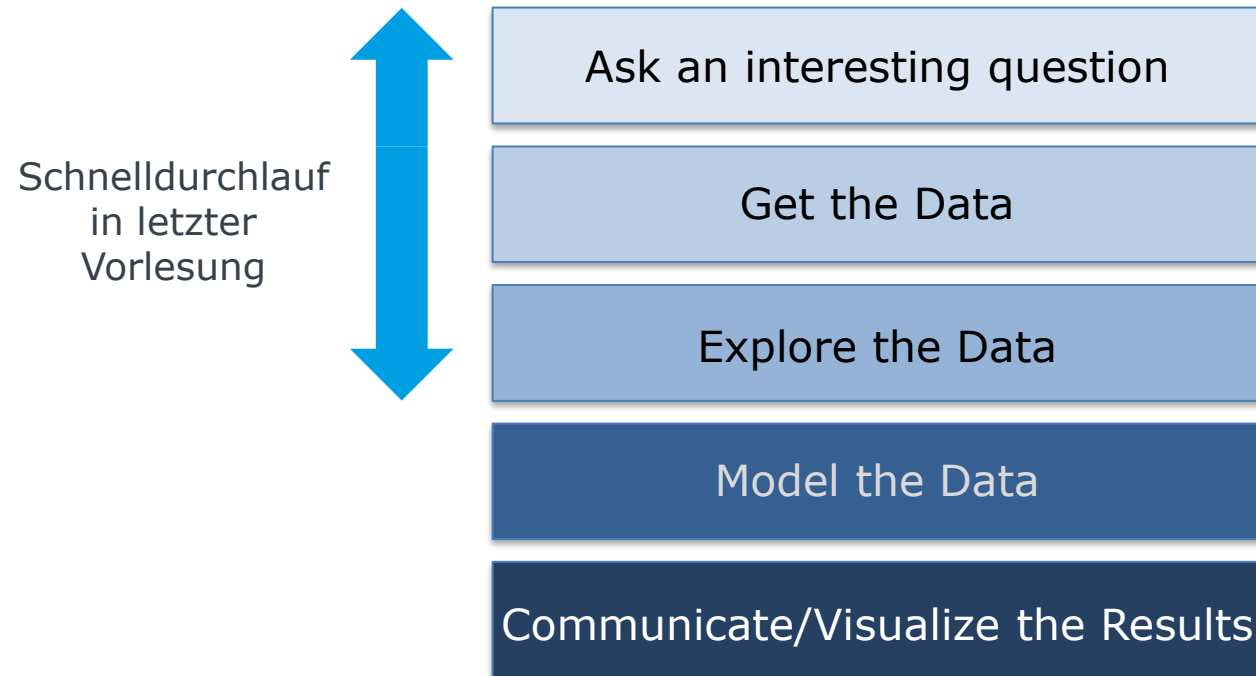
Zusammenlegen?



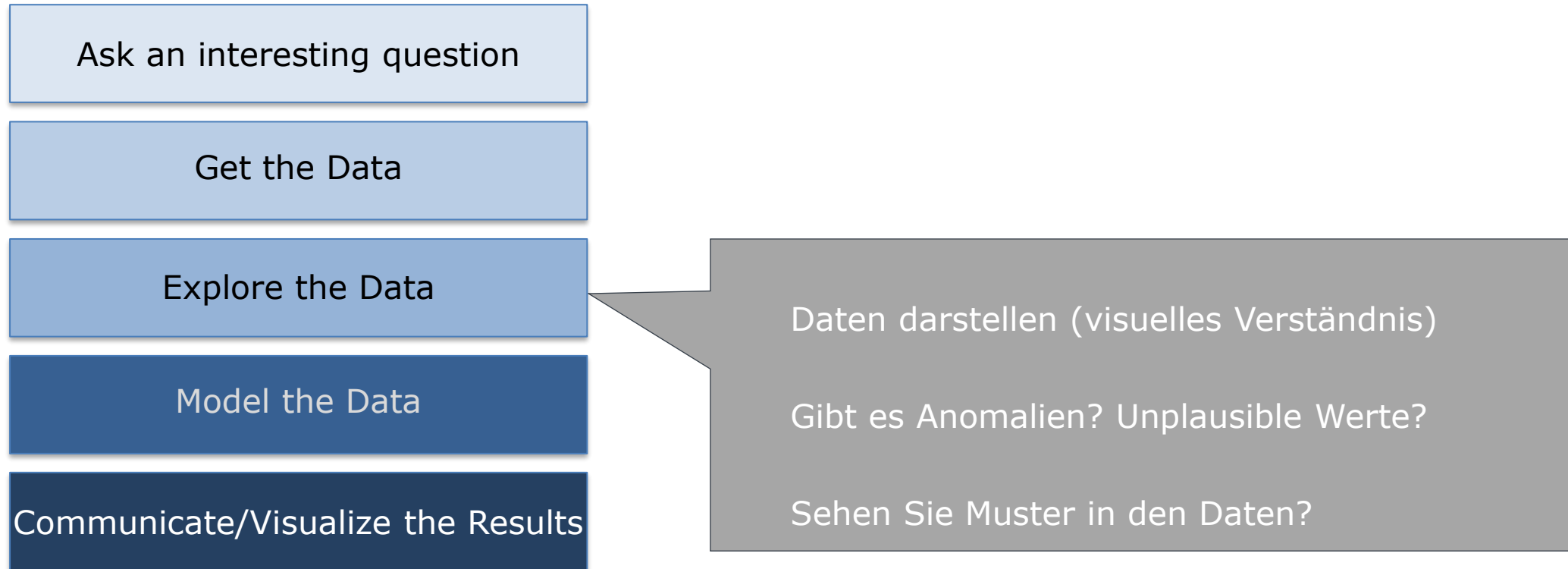
AGENDA

1. Grundlagen Stochastik
 1. Wahrscheinlichkeitstheorie (Backup)
 2. Deskriptive Statistik
 3. Explorative Statistik
2. Fallbeispiel

RÜCKBLICK AUF LETZTE WOCH: VORGEHENSWEISE DATA SCIENCE ANGESEHEN.



FOKUS DER HEUTIGEN VORLESUNG.



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

1.1 WAHRSCHEINLICHKEITSTHEORIE

WAS IST STOCHASTIK?

Stochastik¹ besteht aus folgenden Teilgebieten:

- Wahrscheinlichkeitstheorie: mathematische Erfassung und Analyse zufälliger (nicht-deterministischer) Ereignisse [Backup]
- Mathematische Statistik²:
 - Deskriptive Statistik: Daten durch Graphiken oder Tabellen visuell beschreiben.
 - Explorative Statistik³: 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 στοχαστική τέχνη

² einordnen, von στατίζω

³ Begriff wurde geprägt von John Tukey 1977 in seinem Buch "Exploratory Data Analysis"

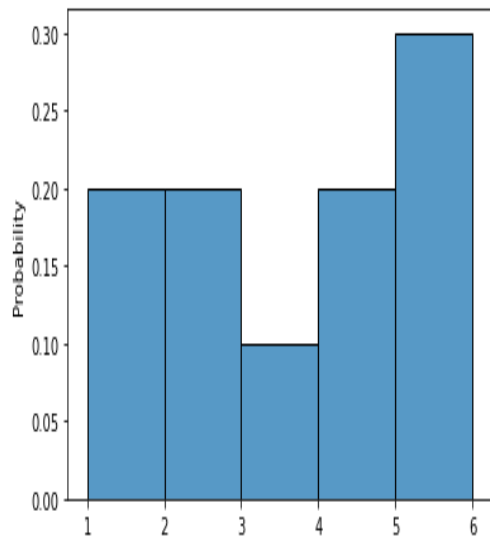


1.2 DESKRIPTIVE STATISTIK

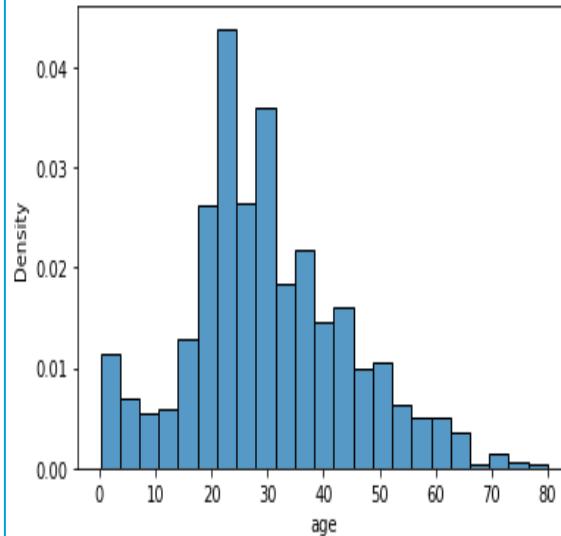
DESKRIPTIVE STATISTIK. BEISPIELE.

Diskrete Verteilung

Diskret = endlich, abzählbare Wertemenge (bspw. Ganzzahlen)



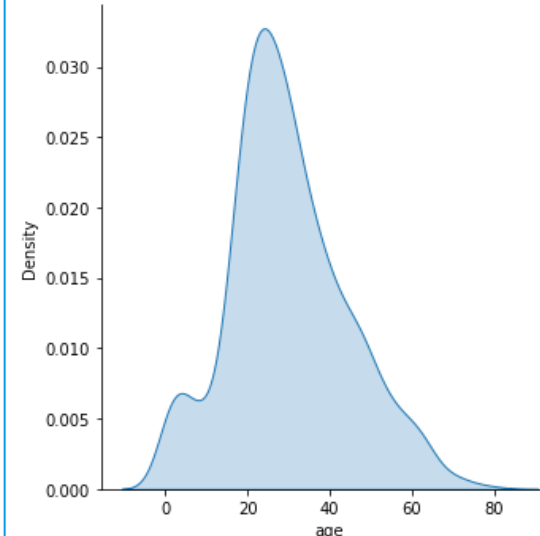
Rel. Häufigkeit Augen eines
Würfels bei 10 Würfeln.
Ergebnismenge = $\{1, \dots, 6\}$



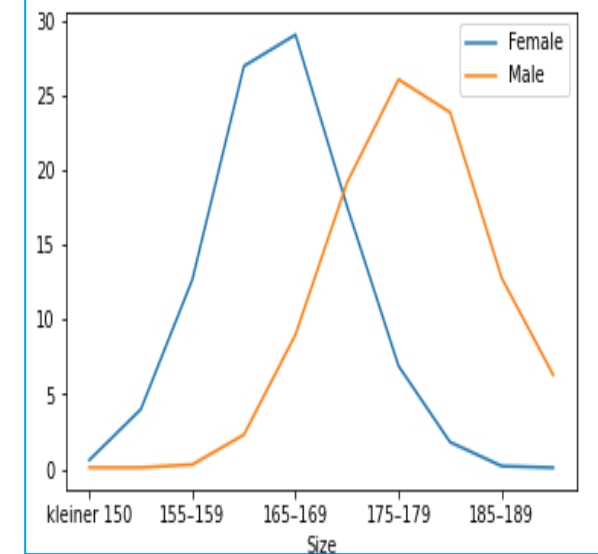
Diskretisierte Altersverteilung
der Passagiere der Titanic.
Ergebnismenge in 23 „Körbe“

Kontinuierliche Verteilung

Kontinuierlich = nicht-abzählbare Wertemenge (bspw. reelle Zahlen)



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

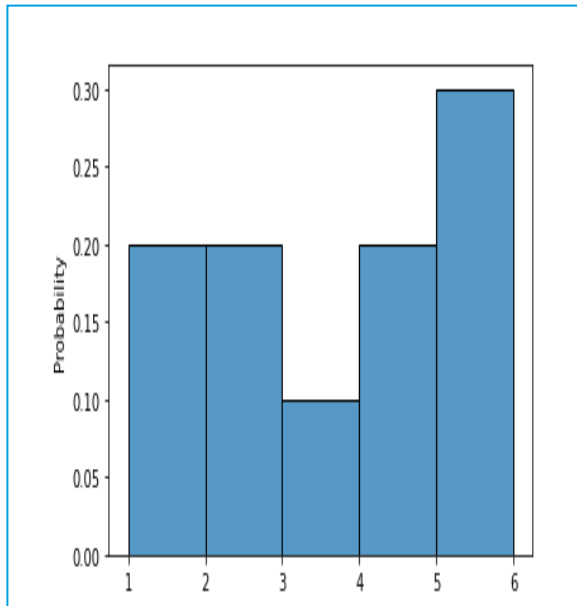


Größenverteilung Einwohner
Deutschland in 2006¹
Ergebnismenge = \mathbb{R}

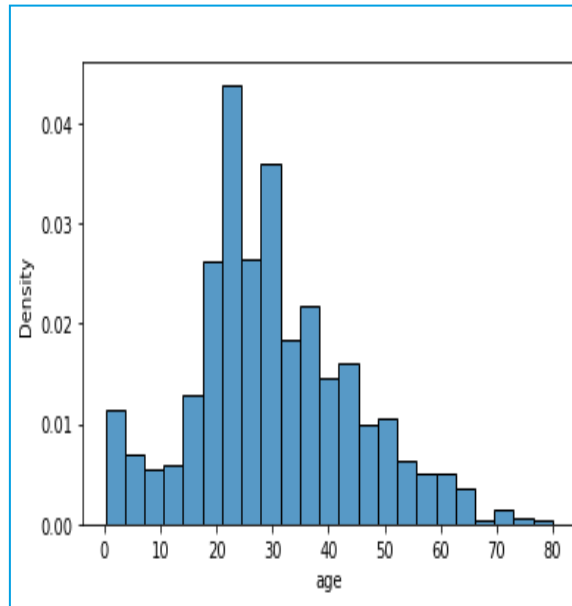
Bei diskreten, endlichen Variablen sprechen wir von einer Wahrscheinlichkeitsfunktion,
bei kontinuierlichen, „nicht-endlichen“ Variablen von einer Dichtefunktion.

DESKRIPTIVE STATISTIK. BEISPIELE.

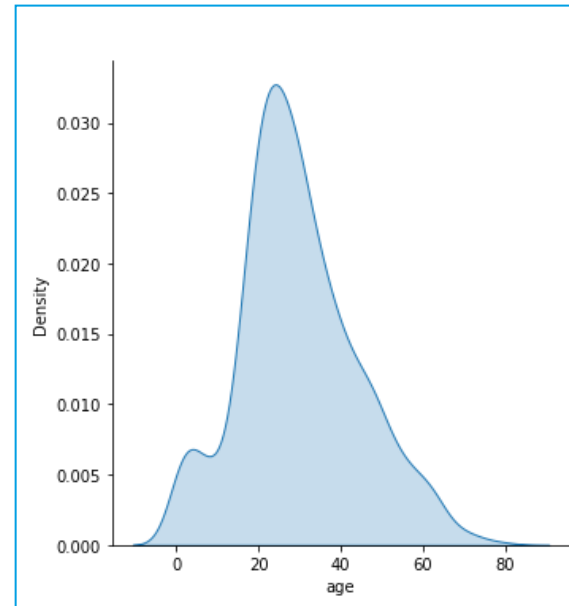
Diskrete Verteilung



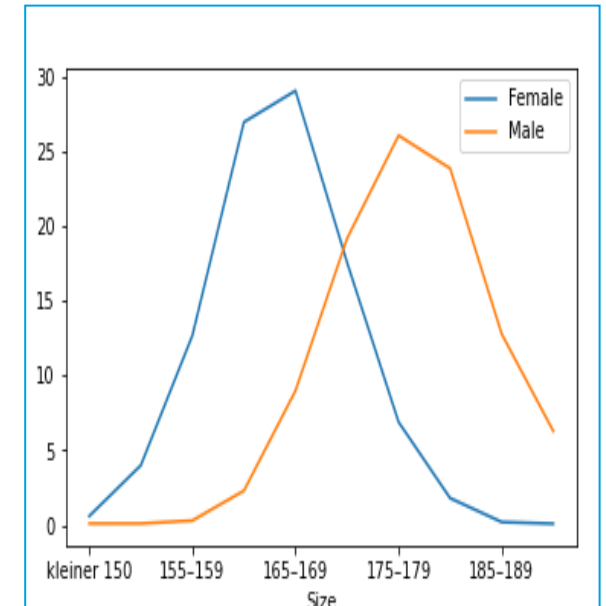
Rel. Häufigkeit Augen eines
Würfels bei 10 Würfeln.
Ergebnismenge = $\{1, \dots, 6\}$



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



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



Größenverteilung Einwohner
Deutschland in 2006¹
Ergebnismenge = \mathbb{R}

Diskreten, endlichen Variablen sind abzählbare, **ganzzahlige Werte** mit einer Wahrscheinlichkeitsfunktion.
Kontinuierliche Variablen sind nicht abzählbare, **reelle Werte** mit einer Dichtefunktion.

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

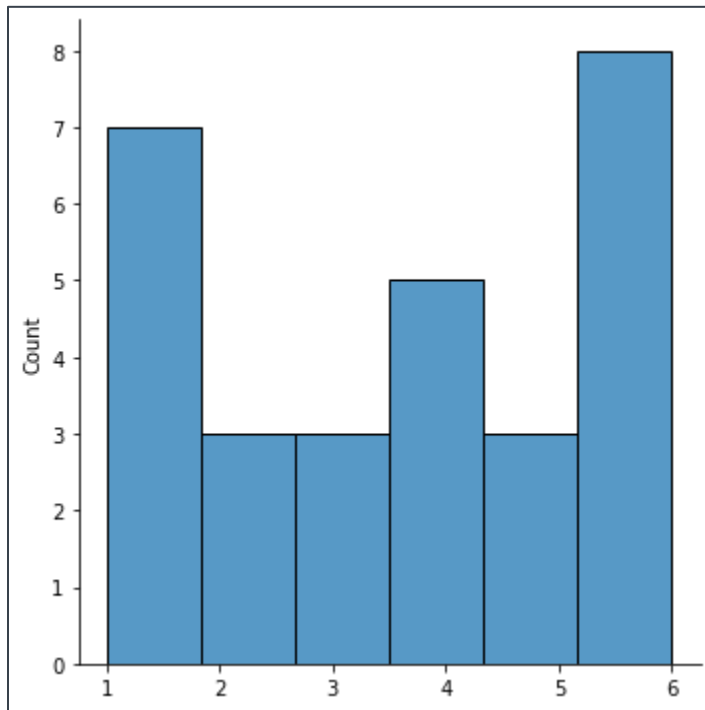
- **Spannweite:** Abstand Min und Max-Wert
- **Varianz:** (quadratische) Abweichung Werte vom Mittelwert. Basis für Standardabweich.
- **Standardabweichung:** durchschnittliche Abweichung/Streuung Werte um Mittelwert.
- **Schiefe:** beschreibt Assymetrie Verteilung. Bei Rechtsschief sind häufiger Werte 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

Wird oft verwechselt!!!

Mean := Durchschnitt

Median := Wert, der Menge in genau 2 gleiche Hälften teilt

Modus: 6 ist häufigstes Ergebnis

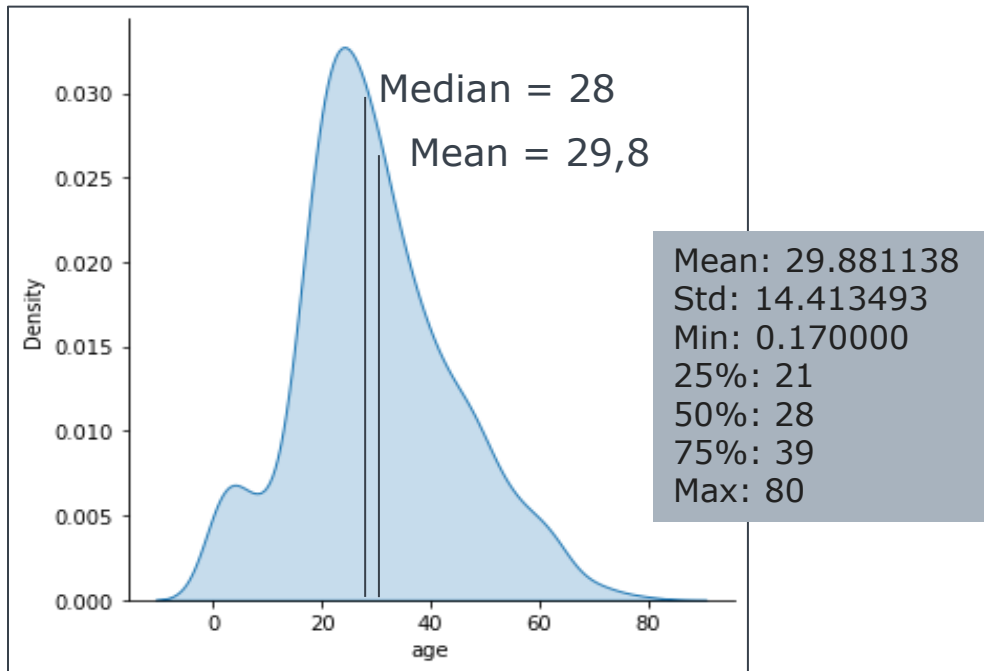
Min: 1 ist niedrigster Ergebniswert

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.



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.



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.

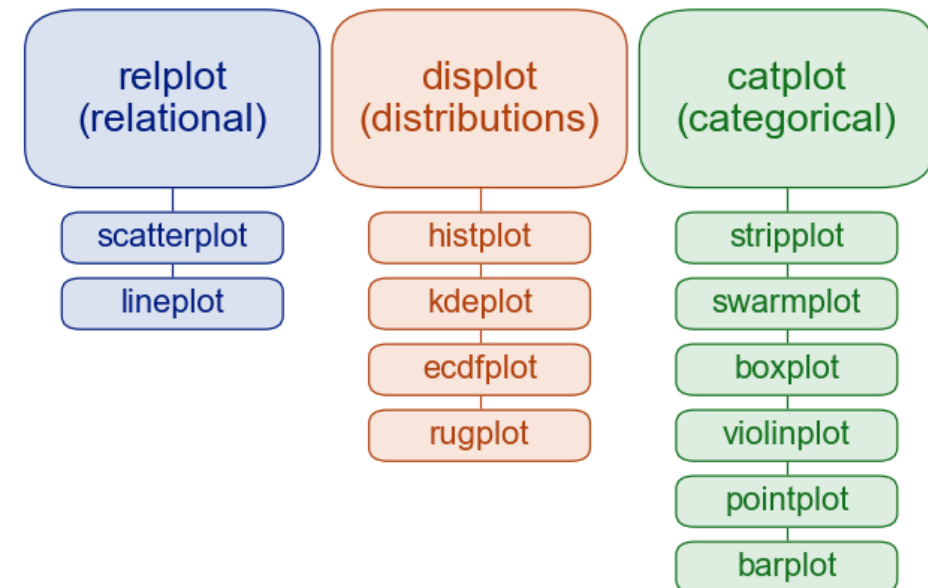
VISUALISIERUNG DATEN: ÜBERSICHT.

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

Was für Features werden geplottet?

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

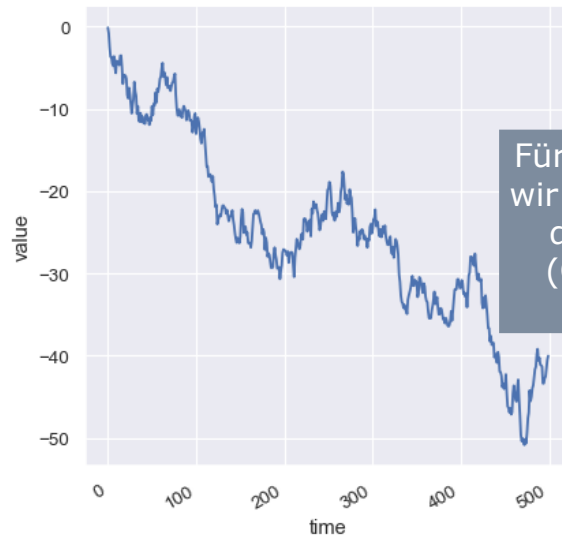
Was für Plots gibt es?



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

VISUALISIERUNG DATEN: RELATIONAL PLOTS.

Line Plots

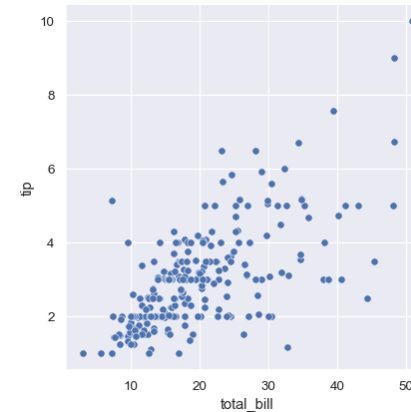


Für die x- und y-Achse nehmen wir ein Feature des Datensatzes der bei Data angegeben ist (Groß- und Kleinschreibung Feature beachten!)

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

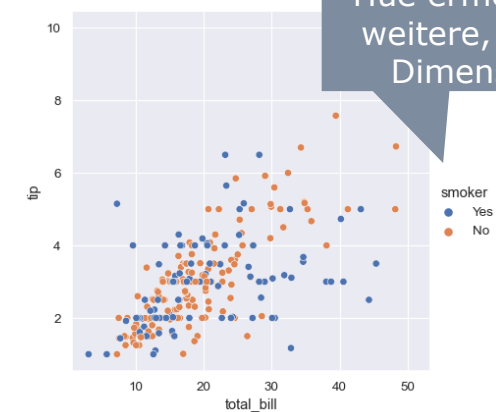
Ziel: Visualisierung von Änderungen über Zeit

Scatter-Plots



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

Ziel: Entdecken von Beziehungen zwischen 2 Features

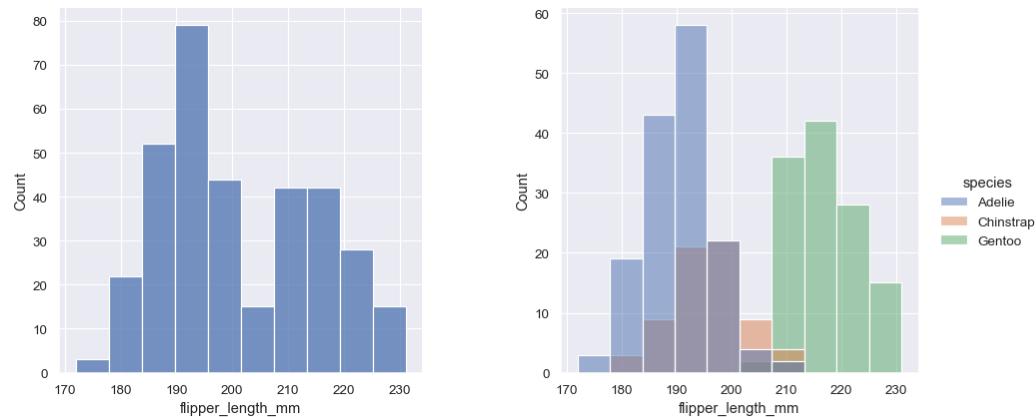


Hue ermöglicht weitere, dritte Dimension

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

VISUALISIERUNG DATEN: VERTEILUNGEN.

Histogram (Histplot)

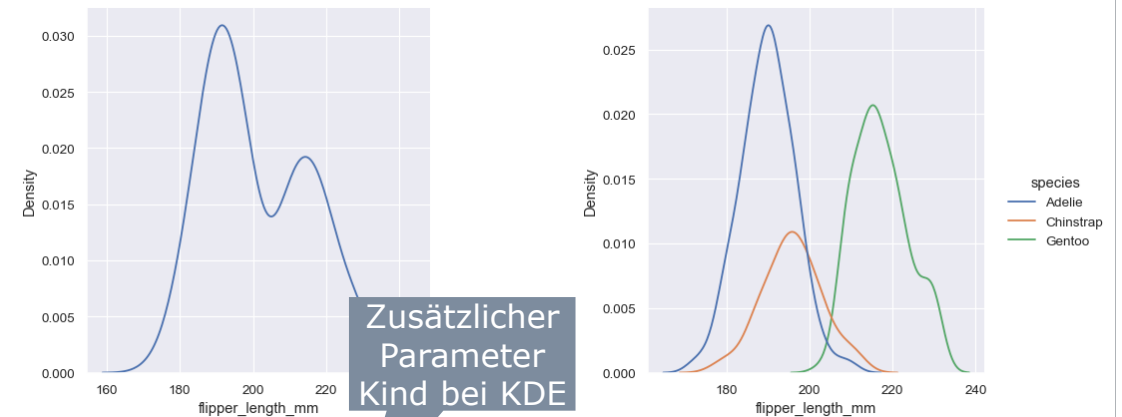


```
sns.displot(penguins,
x="flipper_length_mm")
```

```
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

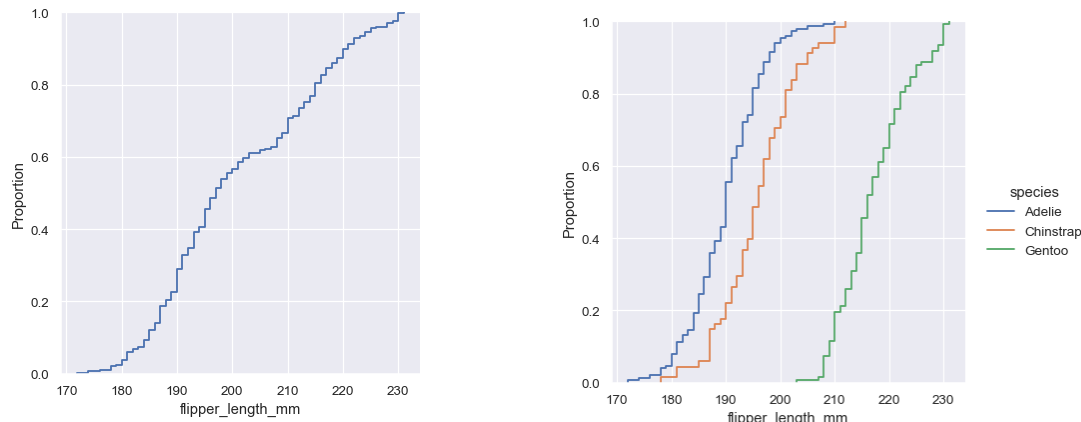
```
sns.displot(penguins,
x="flipper_length_mm",
kind="kde")
```

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

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

VISUALISIERUNG DATEN: VERTEILUNGEN.

Empirical cumulative distributions (ECDF)

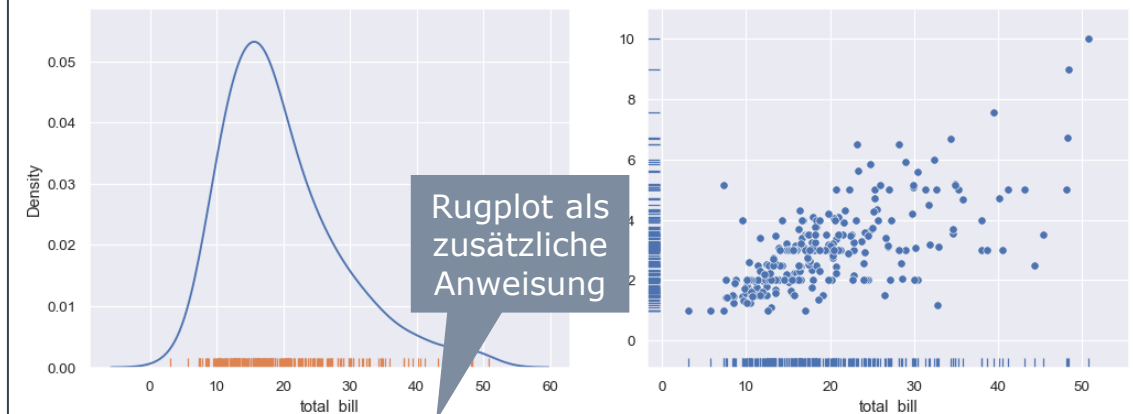


```
sns.displot(penguins,  
x="flipper_length_mm",  
kind="ecdf")
```

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

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

Rugplots



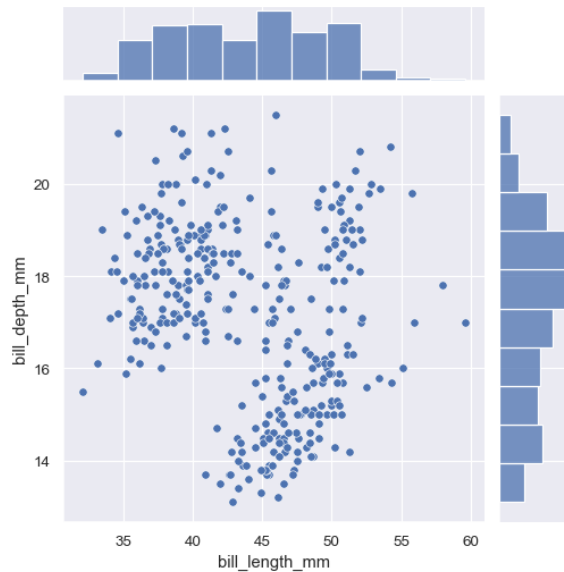
```
sns.kdeplot(data=tips,  
x="total_bill")  
sns.rugplot(data=tips,  
x="total_bill")
```

```
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

VISUALISIERUNG DATEN: WEITERE DISTRIBUTION PLOTS.

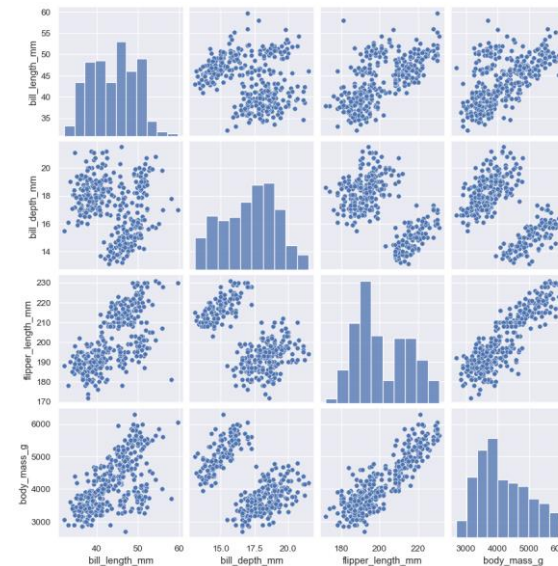
Joint-Plot



```
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

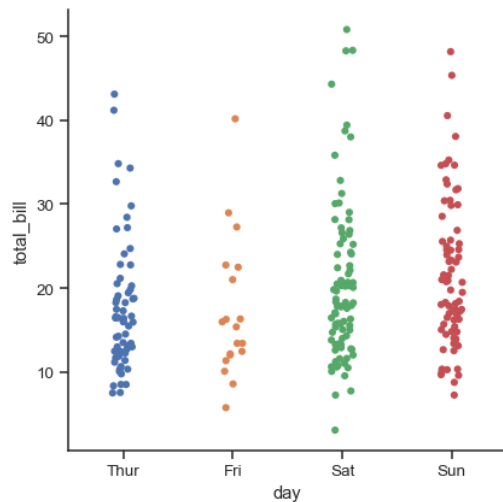


```
sns.pairplot(penguins)
```

Ziel: Entdecken von Beziehungen der Features zueinander

VISUALISIERUNG DATEN: KATEGORISCHE PLOTS.

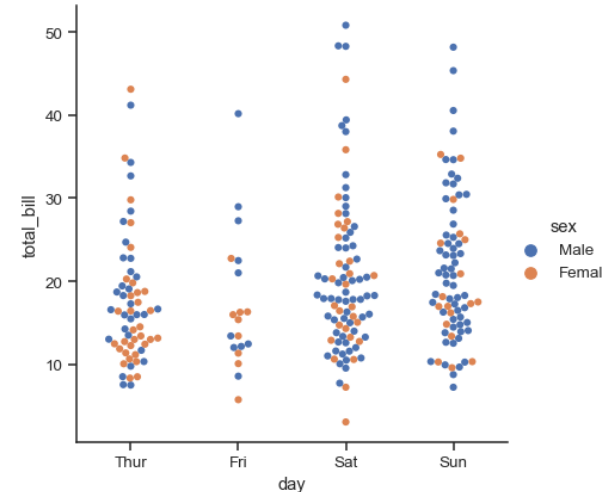
Kategorischer Scatterplot (Stripplot)



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

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

Kategorischer Scatterplot (Swarmplot)



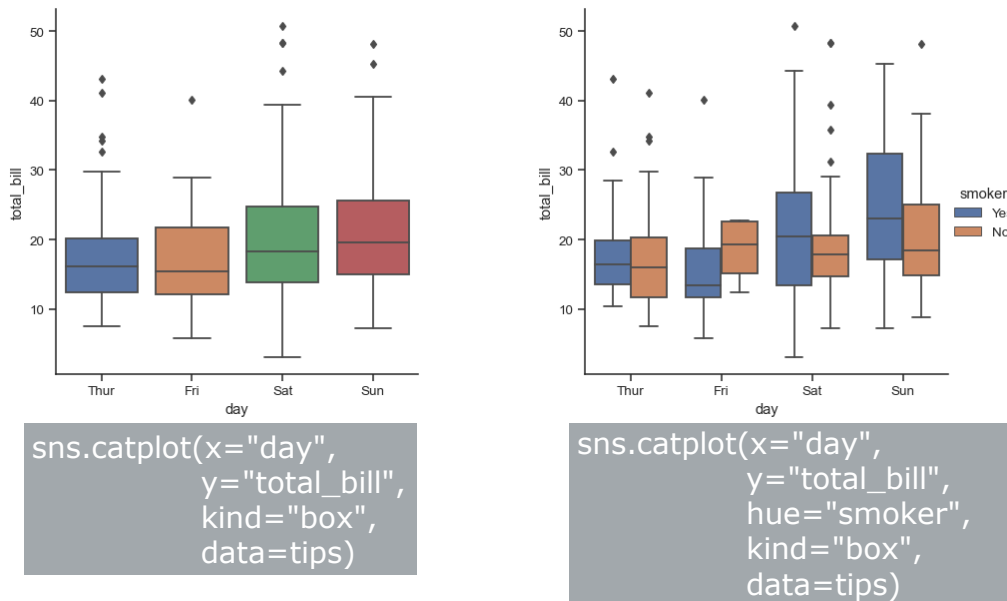
```
sns.catplot(x="day",
            y="total_bill",
            hue="sex",
            kind="swarm",
            data=tips)
```

Zusätzlicher
Parameter kind
für Swarmplot

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

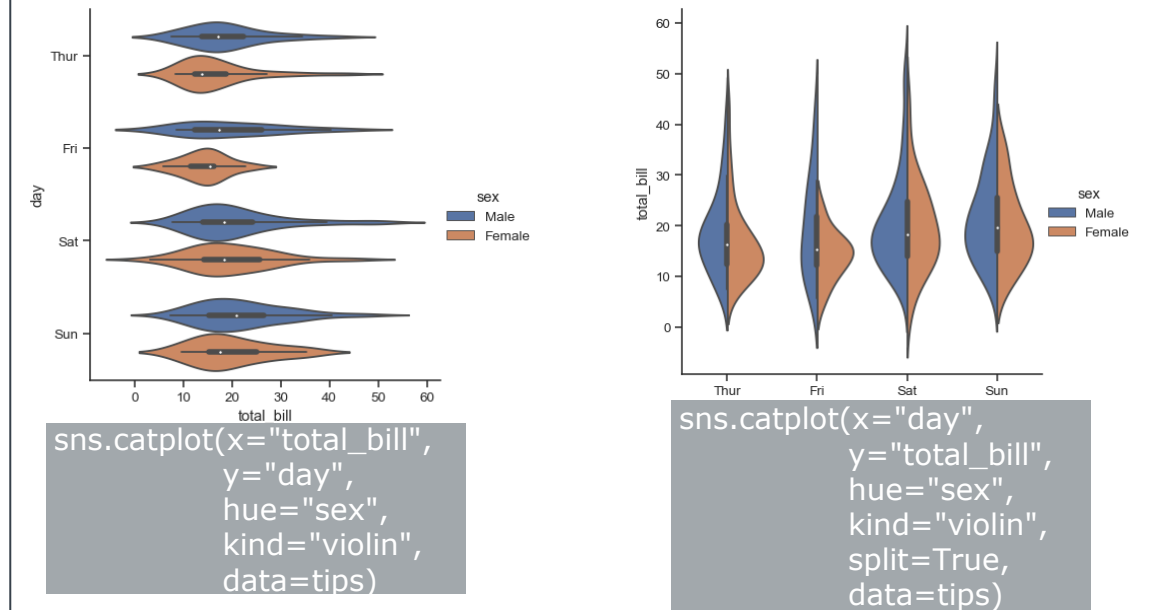
VISUALISIERUNG DATEN: KATEGORISCHE PLOTS.

Kategorischer Verteilungsplot (Boxplot)



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

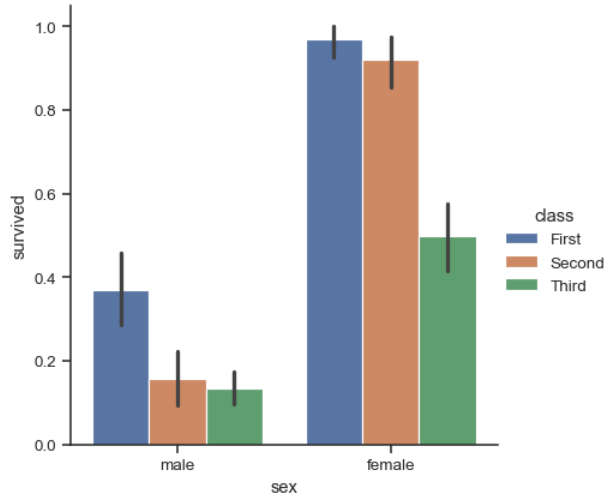
Kategorischer Verteilungsplot (Violinplot)



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

VISUALISIERUNG DATEN: KATEGORISCHE PLOTS.

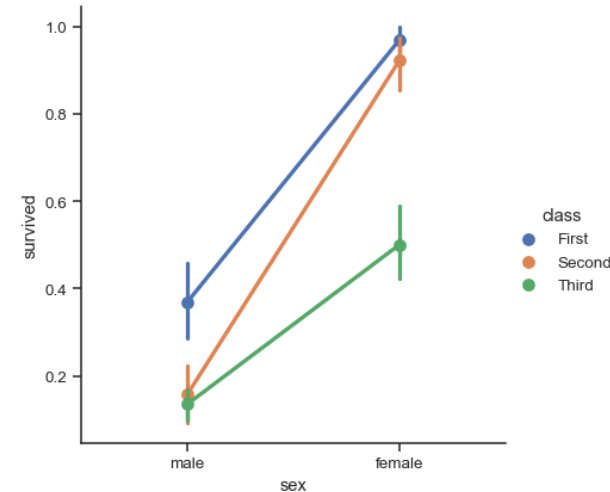
Statistische Abschätzung (Barplots)



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

Ziel: Aufzeigen von Tendenzen.

Statistische Abschätzung (Pointplot)

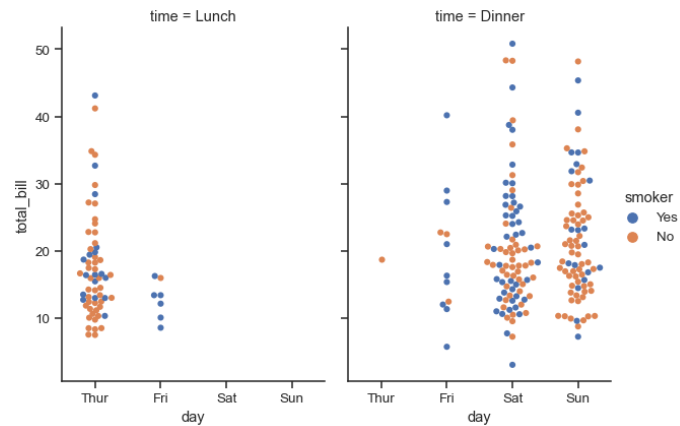


```
sns.catplot(x="sex",
            y="survived",
            hue="class",
            kind="point",
            data=titanic)
```

Ziel: Aufzeigen von Tendenzen

VISUALISIERUNG DATEN: WEITERE KATEGORISCHE PLOTS.

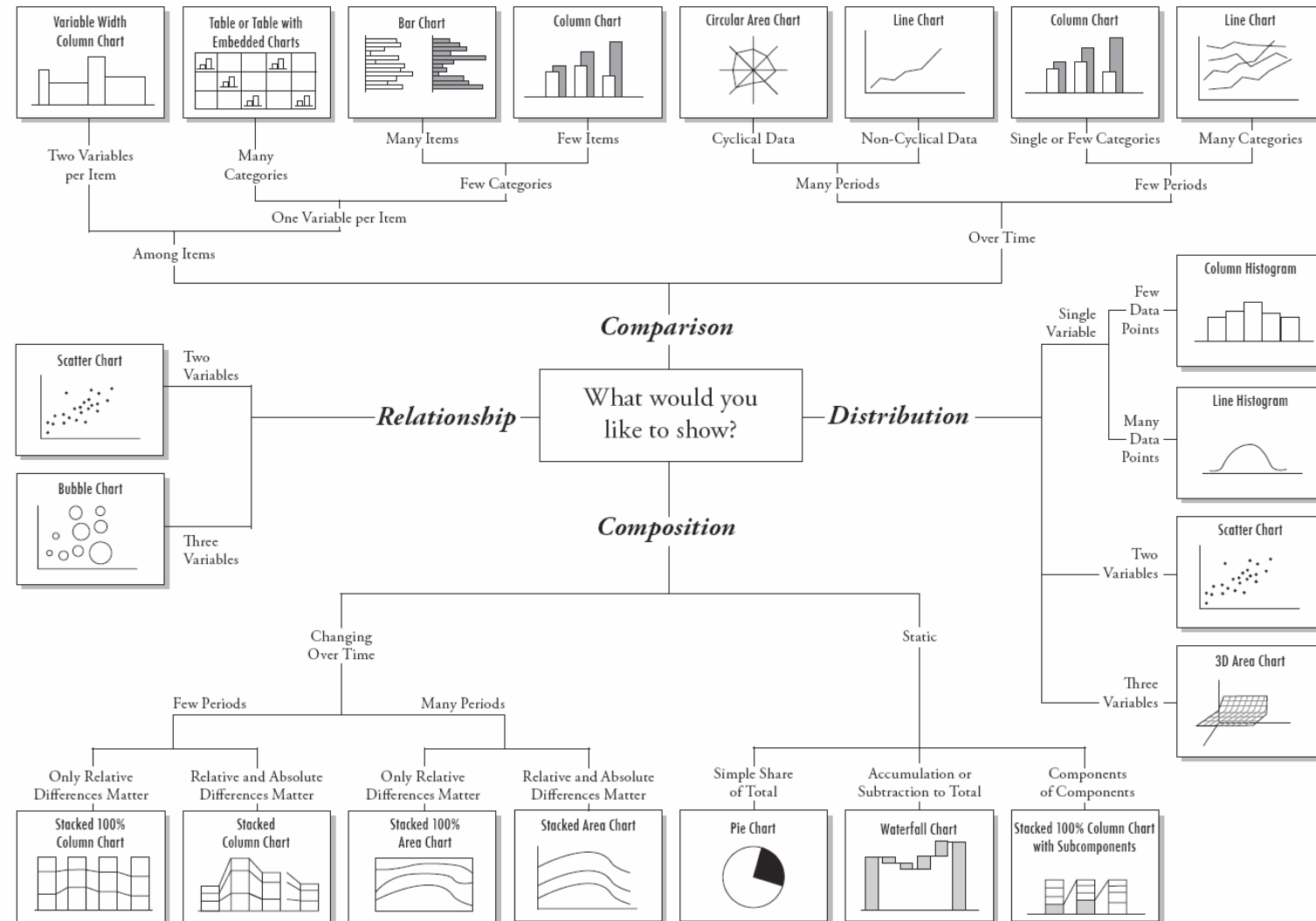
Visualisierung verschiedener Features.



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

Ziel: Aufzeigen von Tendenzen.

ÜBERSICHT PLOTS



FALLBEISPIEL IN GRUPPENARBEIT: DESKRIPTIVE UND EXPLORATIVE STATISTIK.

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

Datensatz verfügbar unter: [Link](#)

IMDB: Top 1000 Filme auf IMDB

Datensatz verfügbar unter: [Link](#)

Empfehlung zum Einstieg

Loan_Data: Daten von Privatkrediten.

Datensatz verfügbar unter: [Link](#).

Erklärung Datensatz: [Link](#)

Bisschen schwieriger, aber
probieren Sie es aus!

Sustainability of Companies: NSC Rating von Firmen

Datensatz verfügbar unter: [Link](#).

EXPLORATIVE STATISTIK: FALLBEISPIEL IN GRUPPENARBEIT

1. Gehen Sie auf die Seite <https://colab.research.google.com/>
2. Loggen Sie sich dort mittels Ihres Google-Accounts ein.
3. Laden Sie eines der folgenden Notebooks in Colab hoch: [Link](#) für IMDB und [Link](#) für Sustainability. **Empfehlung: IMDB ist leichter**
4. Führen Sie die vorhandenen grauen Codezeilen aus (Links-Klick auf den Kasten oder Tastenkombination SHIFT-ENTER).
5. Wenden Sie die gelernten deskriptiven Statistik-Methoden auf den Datensatz an (Tip: Pandas-Describe Funktion) und beschreiben Sie die Ergebnisse.
6. Plotten Sie für jede der vorgestellten Plot-Kategorien je ein Beispiel (Übersicht auf Folie 25).
7. Leiten Sie aus den Plots Hypothesen oder Ergebnisse ab.
8. Können Sie die Hypothesen durch weitere Analysen bestätigen oder widerlegen?
9. Sind die Ergebnisse statistisch belastbar?
10. Stellen Sie Ihre Ergebnisse und Hypothesen vor.

BEISPIELHAFTE PLOTS FÜR DAS IMDB-DATASET.

Allgemein:

- Was ist die häufigste Länge eines Filmes? (Histogram/ KDEPlot: duration)
- Was ist das häufigste Rating eines Filmes? (Histogram/ KDEPlot: star_rating)
- Gibt es einen Zusammenhang zwischen Filmlänge und Rating? (Violinplot/ Scatterplot: x = star_rating und y = duration)
- Gibt es einen Zusammenhang zwischen Filmgenre und Rating? (Stripplot/ Boxplot: genre vs. star_rating)
- Welches Genre hat die meisten Filme unter den Top 1000?
- Gibt es Zusammenhänge zwischen Länge, Rating und Genre? (Violinplot/Scatterplot: star_rating vs. Duration mit hue=content_rating)
- Plotten Sie einen Pairplot. Was kann man für Auffälligkeiten sehen?

Detailanalysen:


- Genre: Schauen Sie sich ein beliebiges Genre an, z.B. Crime. Machen Sie die gleichen Auswertungen: Gibt es Unterschiede?
ACHTUNG: hierfür müssen Sie das Dataset filtern. Dafür müssen Sie Sie statt **data=IMDB_df** folgendes einsetzen:
data=IMDB_df[IMDB_df['genre']=='Crime']. Das ist ein sogenannter Filter in Pandas.
- Rating: Schauen Sie sich ein beliebiges Rating an. Wie heißt der Filter? Was sehen Sie für Erkenntnisse?
ACHTUNG: hierfür müssen Sie wie bei der obigen Frage die Menge nach dem gewählten Rating filtern.....

BEISPIELHAFTE PLOTS FÜR DAS SUSTAINABILITY-DATASET.

Allgemein:

- Welcher Sektor hat die meisten Firmen? (Histogram: Sector).
- Geben Sie das Histogramm in absoluten und Prozentwerten aus. Integrieren Sie das Rating (Hue="Overall ESG RATING").
- Was ist die häufigste ESG-Rating? (Histogram/ KDEPlot: Overall ESG RATING)
- Was ist das häufigste Governance Rating? In absoluten oder Prozentzahlen (Histogram/ KDEPlot: Governance SCORE)
- Gibt es einen Zusammenhang zwischen Subsector und Rating? (Stripplot/ Boxplot: Subsector vs. Overall ESG RATING)
- Was ist das durchschnittliche Overall Rating je Branche (Barplot: Sector vs. Overall ESG Score). Erweitern Sie den Plot mit einer Unterteilung in Rating-Kategorie (hue="Overall ESG RATING").
- Gibt es Zusammenhänge zwischen den einzelnen Scores (replot mit x und y je einen der SCORE-Werte). Integrieren Sie eine Unterscheidung nach Rating (hue="Overall ESG RATING").

Detailanalysen:

- Schauen Sie sich ein beliebigen Sektor an, z.B. Banks. Machen Sie die gleichen Auswertungen: Gibt es Unterschiede?
ACHTUNG: hierfür müssen Sie das Dataset filtern. Dafür müssen Sie statt **data= Sustainability_df** folgendes einsetzen:
data=Sustainability_df[Sustainability_df['Sector']=="Banks"].  Das ist ein sogenannter Filter in Pandas.
- Rating: Schauen Sie sich ein beliebiges Rating an. Wie heißt der Filter? Was sehen Sie für Erkenntnisse?
ACHTUNG: hierfür müssen Sie wie bei der obigen Frage die Menge nach dem gewählten Rating filtern.....

ZUSAMMENFASSUNG DER HEUTIGEN VORLESUNG.

- 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

1.1 WAHRSCHEINLICHKEITSTHEORIE

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 WAHRSCHEINLICKEITSRECHNUNG AM FALLBEISPIEL MÜNZWURF.

Ergebnismenge: Menge aller möglichen Ergebnisse, z.B. $\Omega = \{\text{Kopf, Zahl}\}$.

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. $E := \text{Kopf geworfen}$

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. $\text{Pr}[\text{gesamte Ergebnismenge } \Omega] = 1$
2. $\text{Pr}[\text{leere Menge } \emptyset] = 0$
3. $0 \leq \text{Pr}[\text{Ereignis E}] \leq 1$
4. $\text{Pr}[\overline{\text{Ereignis E}}] = 1 - \text{Pr}[\text{Ereignis E}]$ (Gegenwahrscheinlichkeit)
5. $\text{Pr}[A \cap B] = \text{Anzahl der gemeinsamen eingetretenen Ereignisse A und B}$
6. $\text{Pr}[\text{Ereignis A} \cup \text{Ereignis B}] = \text{Pr}[A] + \text{Pr}[B] - \text{Pr}[A \cap B]$ (A oder B trat auf)

Mit diesen 6 Regeln können wir diskrete und kontinuierliche Wahrscheinlichkeiten berechnen

EINFÜHRUNG WAHRSCHEINLICHKEITSRECHNUNG: BEDINGTE WAHRSCHEINLICHKEIT.

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

$$\Pr[A | B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

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}. $\rightarrow \Pr[A|B] = \frac{2/6}{3/6} = \frac{2}{3} = 66\%$

- Titanic: Wie hoch ist die Chance, daß ein Passagier Mann ist und überlebt?

Anzahl überlebender Männer = 161, Anzahl männliche Passagiere = 843 $\rightarrow \Pr[A|B] = 161/843 = 19\%$

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

$$\begin{aligned}\Pr[A \cap B] &= \Pr[A] * \Pr[B] \\ \Pr[A | B] &= \Pr[A]\end{aligned}$$

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 Würfeln:

$$\Pr[X \leq 2] = \Pr[X=0 \text{ Kopf geworfen}] + \Pr[X=1 \text{ Kopf geworfen}] + \Pr[X=2 \text{ Kopf geworfen}] = 1/8 + 3/8 + 3/8 = 7/8$$

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

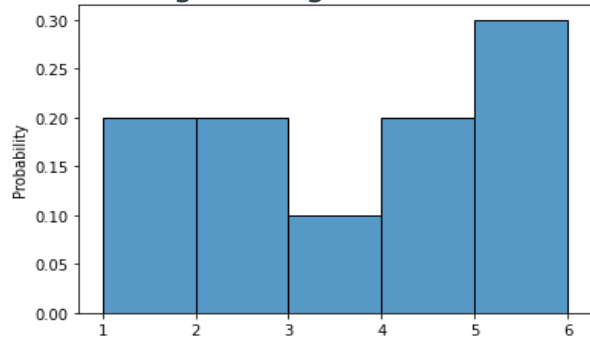
Die **Varianz** definiert die Streuung der Zufallsvariablen um den Erwartungswert (mehr dazu im nächsten Kapitel).

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

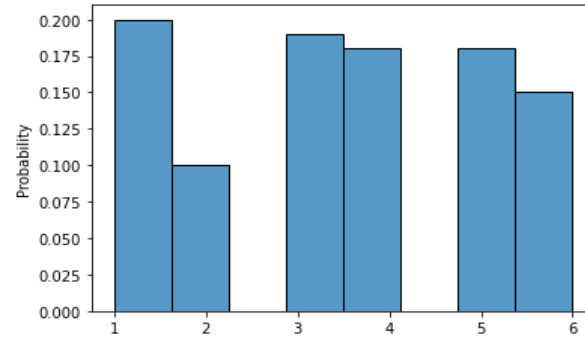
EINFÜHRUNG WAHRSCHEINLICHKEITSRECHNUNG: WAS IST GENÜGENDE OFT MESSEN- ODER DAS GESETZ DER GROßEN ZAHLEN.

Wir messen mit den Zufallsvariablen X_1, \dots, 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.

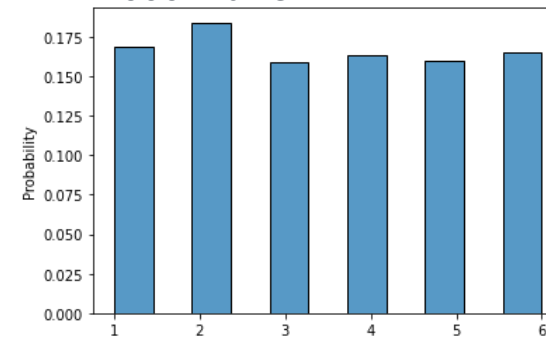
Rel. Häufigkeit Augen bei 10 Würfeln



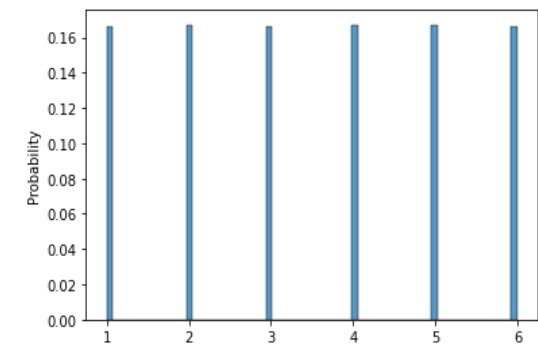
100 Würfeln



1000 Würfeln



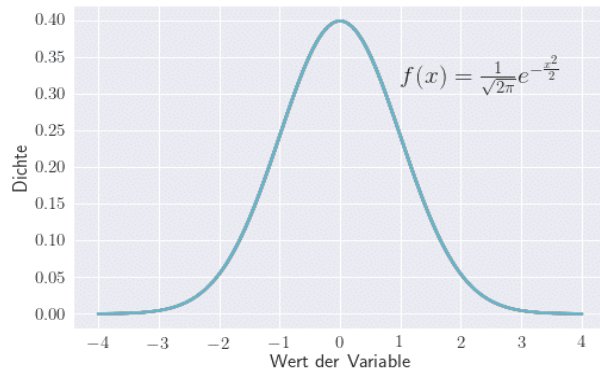
500'000 Würfeln



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

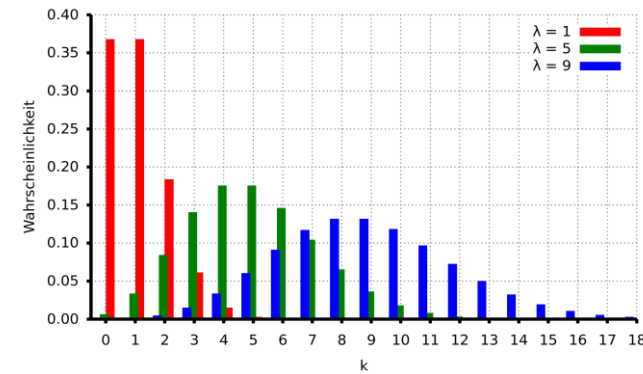
Die Ergebnisse von Zufallsvariablen sind **nur dann** belastbar,
falls sie einer genügend großen Menge an Versuchen zugrunde liegen!!!

ÜBERSICHT WICHTIGER VERTEILUNGEN.



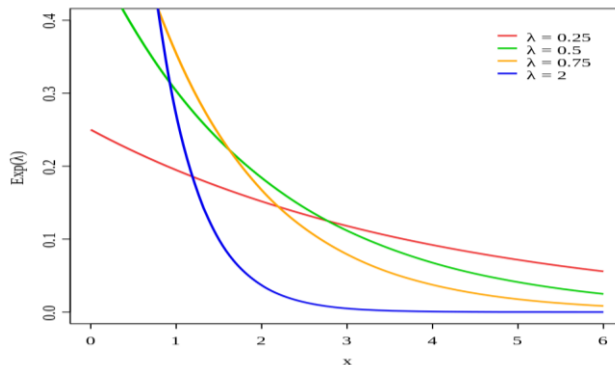
Normalverteilung:
Modellierung vieler natürlicher und statistischer Prozesse.

- Beispiele:
- Größe Bevölkerung
 - Prüfungsergebnisse
 - Prozessqualität in einer Fabrik.



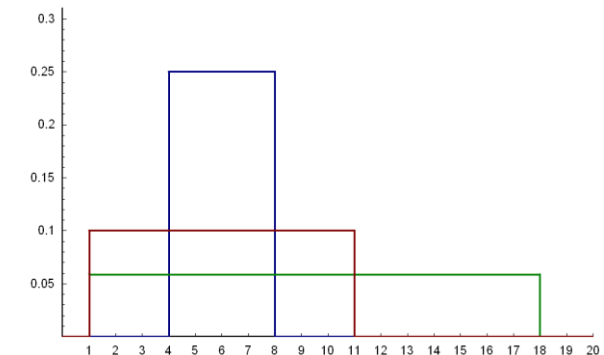
Poisson-Verteilung:
Modellierung Ereignisse, die bei konstanter mittlerer Rate unabhängig voneinander in einem festen Zeitintervall oder räumlichen Gebiet eintritt.

- Beispiele:
- Hotline-Anrufe je Stunde
 - Website-Ausfälle je Stunde



Exponentialverteilung:
Modellierung von Zeitintervallen.

- Beispiele:
- Zeit bis Ausfall eines Geräts
 - Wartezeit in Hotline



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

- Beispiele:
- Wurf einer idealen Münze oder Würfel