Statistik für Data Scientists

Vorlesung 5: Wahrscheinlichkeit & Verteilungen

Prof. Dr. Siegfried Handschuh
DS-NLP
Universität St. Gallen

Recap & Ziele heute

- Recap V4: Korrelation & Zusammenhang.
- Heute: Grundlagen der Wahrscheinlichkeitsrechnung.
- Diskrete Zufallsvariablen: Bernoulli, Binomial, Poisson.
- Erwartungswert, Varianz, Gesetz der grossen Zahlen.

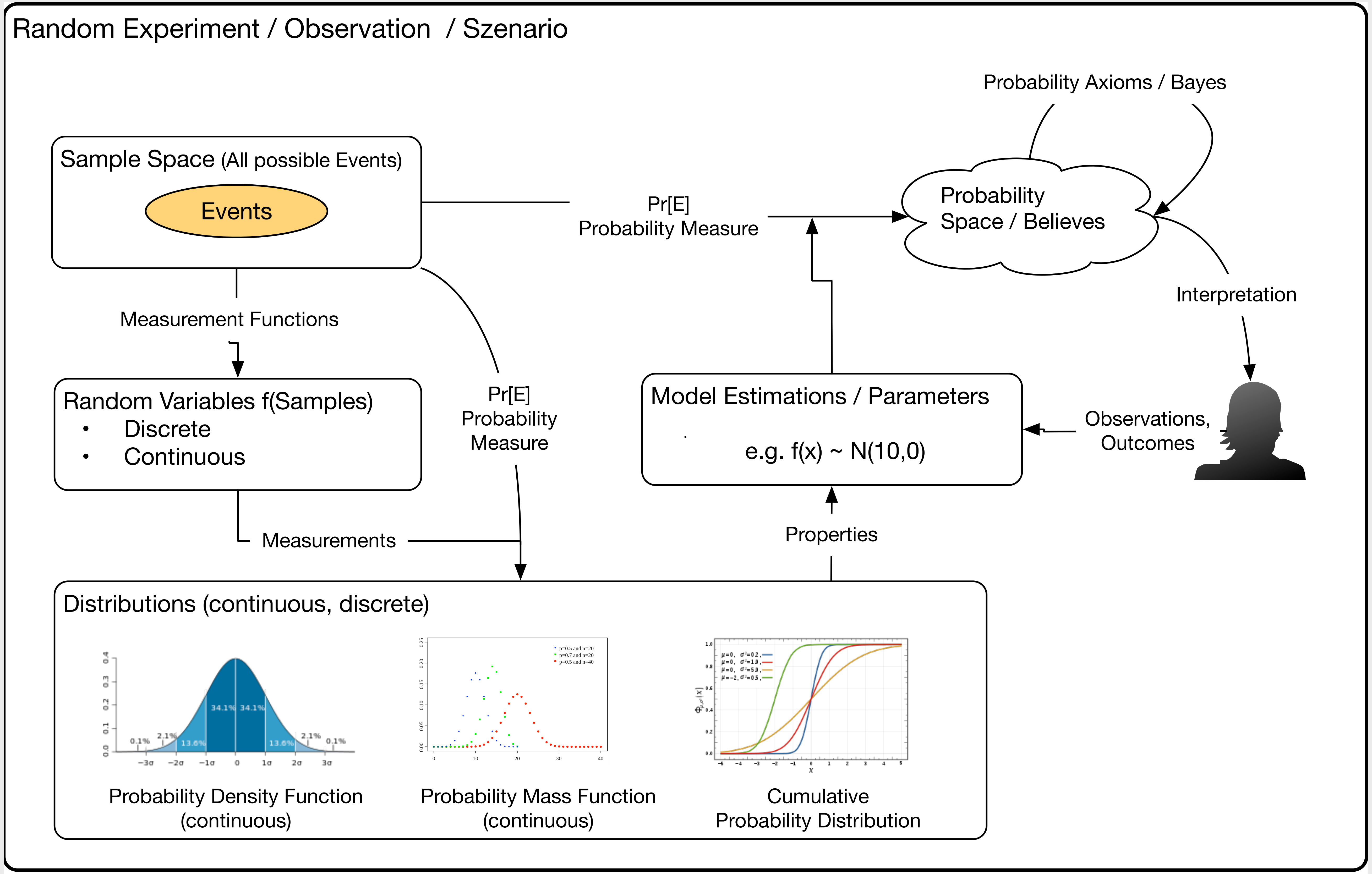


Image sources: Wikipedia

Warum Wahrscheinlichkeit?

Der Zufall als Werkzeug

Data Scientists brauchen Wahrscheinlichkeit, um Unsicherheit zu verstehen und zu steuern.

- Daten sind nie vollständig: jede Analyse beruht auf Wahrscheinlichkeiten.
- Ein Spam-Filter klassifiziert Mails nicht „sicher“, sondern mit Wahrscheinlichkeit $Pr(\text{spam} \mid \text{Text})$.
- Ohne Wahrscheinlichkeiten keine Entscheidungsmodelle, keine KI, kein Machine Learning.
- Zufall ist das Rohmaterial der Vorhersage.

Mini-Check: Warum kann ein Modell nie „sicher“ richtig liegen?

Vom Bauchgefühl zur Formel

Wahrscheinlichkeit formalisiert Intuition: sie zwingt uns, den Zufall zu quantifizieren.

- Alltag: «Es regnet wahrscheinlich.» $\rightarrow$ unpräzises Gefühl.
- Statistik: $Pr(\text{Regen morgen}) = 0.7 \rightarrow$ präzise, überprüfbar.
- Objektive und subjektive Sicht: Häufigkeit vs. Überzeugung.
- Historischer Start: Glücksspiel, Würfel, Pascal & Fermat $\rightarrow$ erstes Modell des Zufalls.

Mini-Check: Warum war Glücksspiel der perfekte Ausgangspunkt für die Wahrscheinlichkeitstheorie?

Wahrscheinlichkeit in der Praxis: Drei Szenarien

Wahrscheinlichkeiten stecken in fast allen Data-Science-Entscheidungen.

- A/B-Test: Hat Button A wirklich mehr Klicks oder ist das Zufall?
- Sensoranalyse: Wie sicher ist das „Anomalie erkannt“-Signal?
- Modellbewertung: Ist ein Modell mit $\text{Accuracy} = 0.87$ „gut genug“?
- Jedes Beispiel beruht auf dem gleichen Prinzip:
  $\text{Stichproben} \approx \text{Zufall} \rightarrow \text{Verteilungen}$

Mini-Check: Welche dieser drei Situationen kennst du aus Projekten oder Medien?

Begriffe und Ziele

Drei zentrale Fragen

1. Wie definieren wir «Zufall» mathematisch?
2. Wie berechnen wir die Wahrscheinlichkeit komplexer Ereignisse?
3. Wie entstehen Verteilungen aus vielen Zufällen?

Heute: von den Axiomen über bedingte Wahrscheinlichkeit bis zu Zufallsvariablen.

Mini-Check: Welches dieser drei Themen ist dir am wenigsten vertraut?

Kleine Denkübung: Der Zufall als Muster

Zufall wirkt chaotisch, folgt aber Regeln: genau das ist Statistik.

- Würfle zehnmal. Die Folge: 4 – 1 – 3 – 6 – 2 – 4 – 5 – 2 – 6 – 1
- Wir sehen Chaos, aber jede Zahl kommt $\approx$ gleich häufig
- Mit vielen Wiederholungen entsteht Regelmässigkeit $\rightarrow$ Gesetz der grossen Zahlen
- So verwandeln wir Zufall in Wissen

Mini-Check: Was passiert mit der relativen Häufigkeit von «6», wenn wir unendlich oft würfeln?

Take-Away: Warum Wahrscheinlichkeit zählt

Data Science lebt vom Umgang mit Unsicherheit:
Wahrscheinlichkeit ist ihre Grammatik.

- Wir quantifizieren Unsicherheit, statt sie zu ignorieren
- Vom Bauchgefühl zur Formel: Jede Aussage über Daten ist probabilistisch
- Modelle entscheiden nicht sicher, sondern wahrscheinlich
- Ohne Wahrscheinlichkeit keine Tests, keine Prognosen, kein Lernen

Ereignisse und Axiome

Vom Ergebnis zum Ereignis

Ereignisse sind Mengen von Ergebnissen: die Bausteine der Wahrscheinlichkeit.

- Ergebnis: ein mögliches Resultat eines Zufallsexperiments (z. B. eine Augenzahl 6)
- Ereignis: Menge von Ergebnissen, z. B. «gerade Zahl» = {2, 4, 6}
- Ereignisraum Ω: Menge aller möglichen Ergebnisse
- Notation: $A, B, C \subseteq \Omega$

Diese Mengen-Sprache ist die Grammatik der Wahrscheinlichkeit

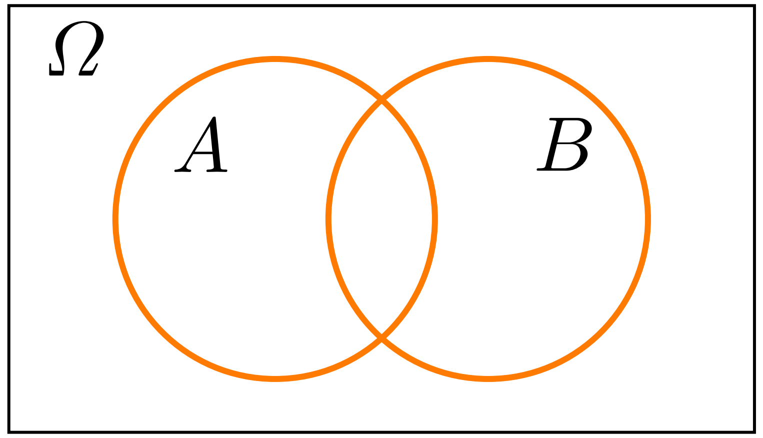
Mini-Check:
Was ist das Ereignis «Augenzahl > 4» beim Würfelwurf?

Venn-Diagramme: Grafische Darstellung von Ereignissen.

- Das Rechteck stellt den Stichprobenraum aller Ereignisse dar: Ω
- Die Kreise zeigen Teilmengen von Ereignissen $A \subseteq \Omega$, $A \in \mathcal{P}(\Omega)$

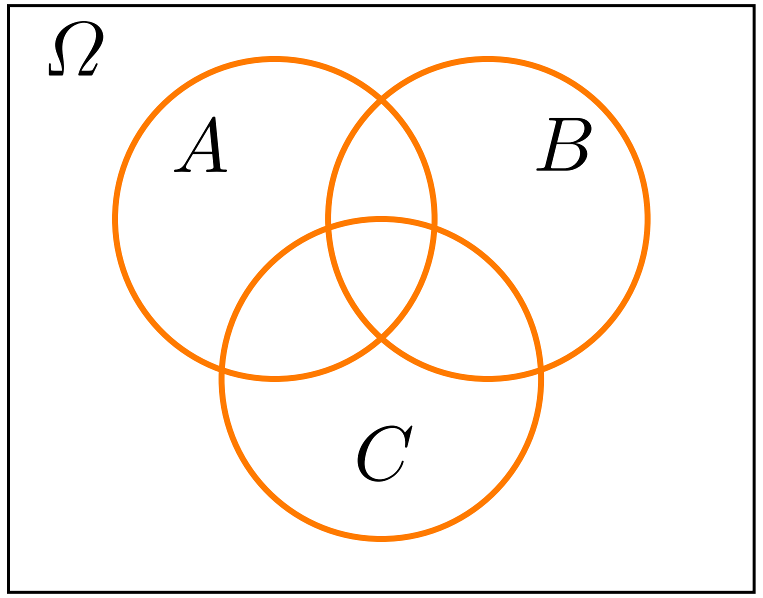
Mit zwei Ereignismengen

$$A \subset \Omega, B \subset \Omega.$$



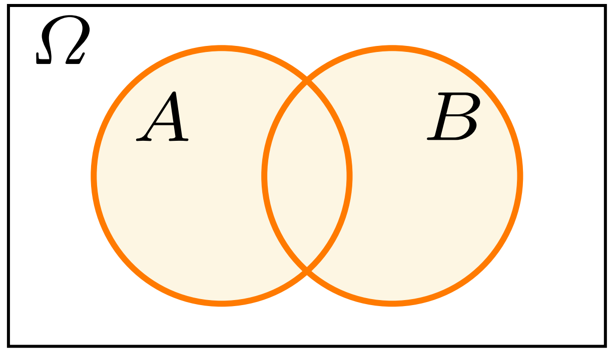
Mit drei Ereignismengen

$$A \subset \Omega, B \subset \Omega, C \subset \Omega.$$



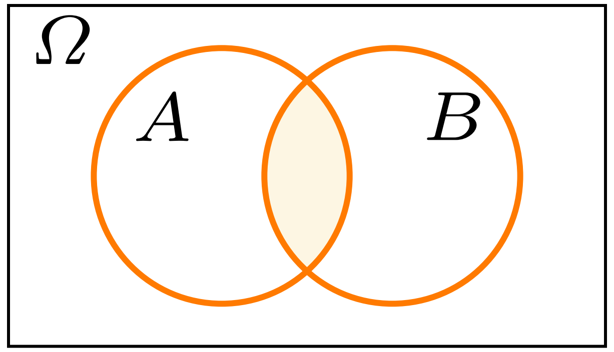
Kommutativgesetz	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Assoziativgesetz	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributivgesetz	$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$	$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

$A \cup B$



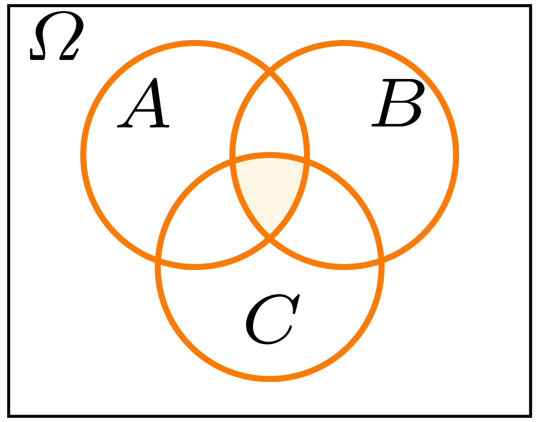
Vereinigung: Mindestens eines der Ereignisse tritt ein.

$A \cap B$



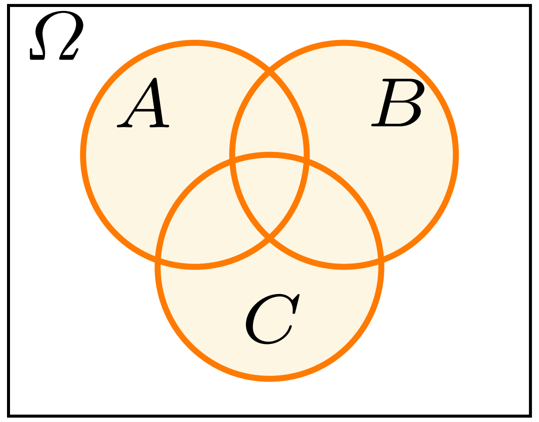
Schnitt: Beide Ereignisse treten gleichzeitig ein.

$A \cap B \cap C$



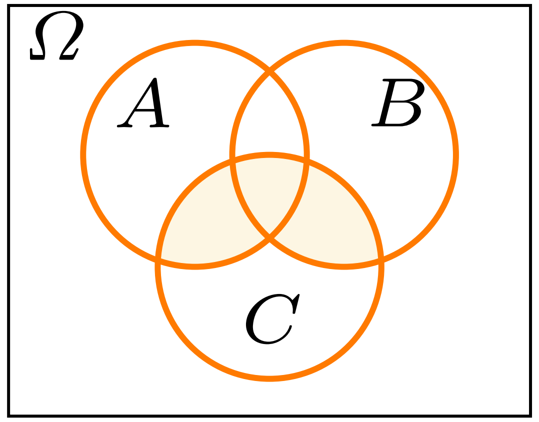
Dreifacher Schnitt: Alle drei Ereignisse treten gemeinsam ein.

$A \cup B \cup C$



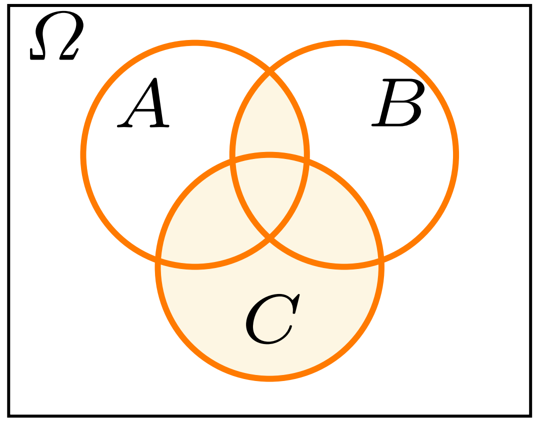
Dreifache Vereinigung: Mindestens eines der Ereignisse tritt ein.

$(A \cup B) \cap C$



Distributivbeispiel 1: Nur Teilmengen, in denen C auftritt.

$A \cap B \cup C$



Distributivbeispiel 2: A und B zusammen oder C allein.

Die drei Axiome nach Kolmogorov

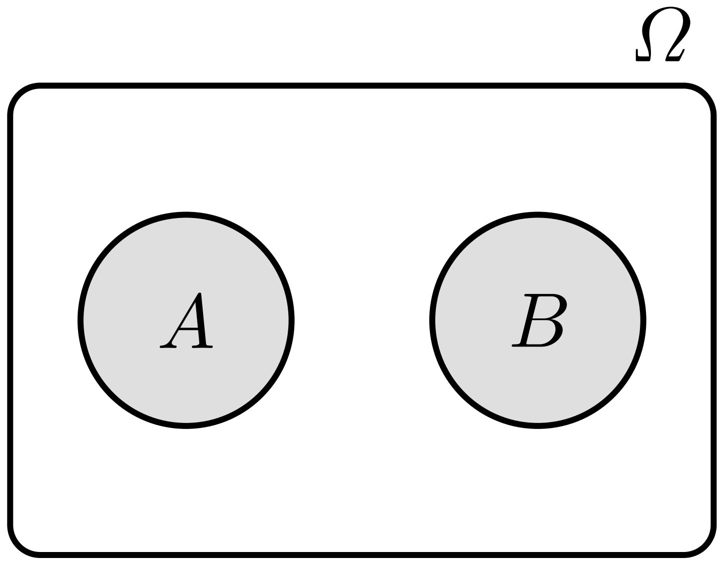
Alle Wahrscheinlichkeiten folgen drei einfachen Gesetzen.

1. Nichtnegativität: $Pr(A) \geq 0$
2. Normierung: $Pr(\Omega) = 1$
3. Additivität: Wenn $A \cap B = \varnothing$, dann $Pr(A \cup B) = Pr(A) + Pr(B)$

Aus diesen Regeln folgt alles: von Bedingungen bis Bayes

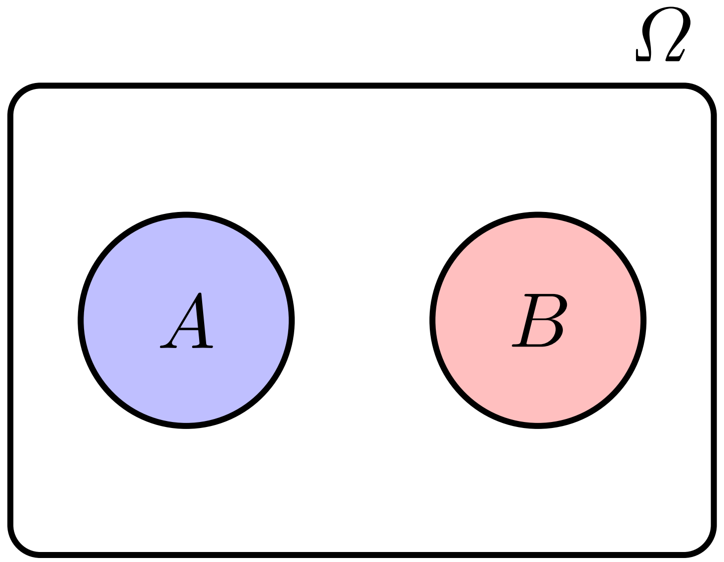
Mini-Check:
Wenn A und B sich nicht überlappen und $Pr(A) = 0.3$, $Pr(B) = 0.5$,
was gilt für $Pr(A \cup B)$?

$$A \cap B = \emptyset$$



Unvereinbare Ereignisse: A und B haben keinen gemeinsamen Teil.

$$Pr(A \cup B) = Pr(A) + Pr(B)$$



Additivität: Die Wahrscheinlichkeit beider Teilereignisse addiert sich.

$$A \cap B = \emptyset \quad \Longrightarrow \quad Pr(A \cup B) = Pr(A) + Pr(B)$$

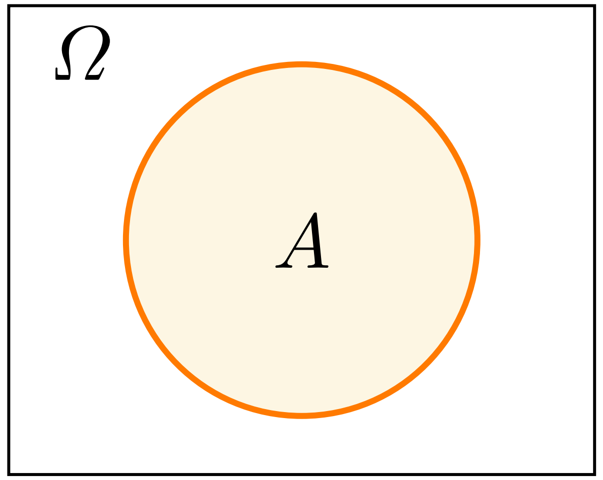
Das Gesetz vom komplementären Ereignis

Jedes Ereignis hat ein Gegenteil: zusammen ergeben sie Sicherheit.

- Komplement: A^c = «A tritt nicht ein»
- Formel: $Pr(A^c) = 1 - Pr(A)$
- Beispiel: $Pr(\text{Regen}) = 0.3 \rightarrow Pr(\text{kein Regen}) = 0.7$
- Praktisch: Machine-Learning-Modelle arbeiten immer mit p und $(1 - p)$

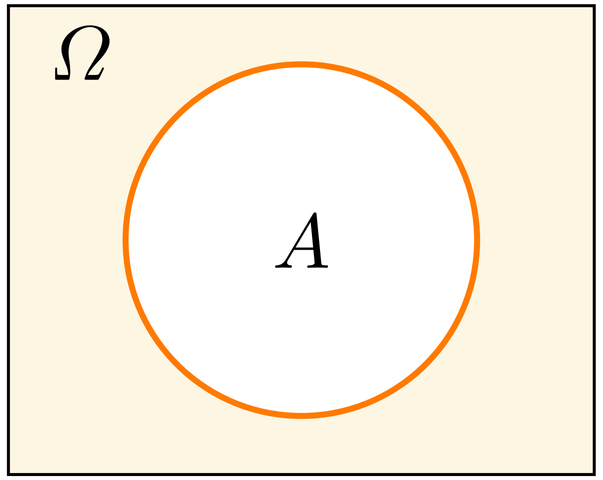
Mini-Check: Wenn ein Klassifikator $p(\text{spam}) = 0.92$, wie gross ist $p(\text{nicht spam})$?

A



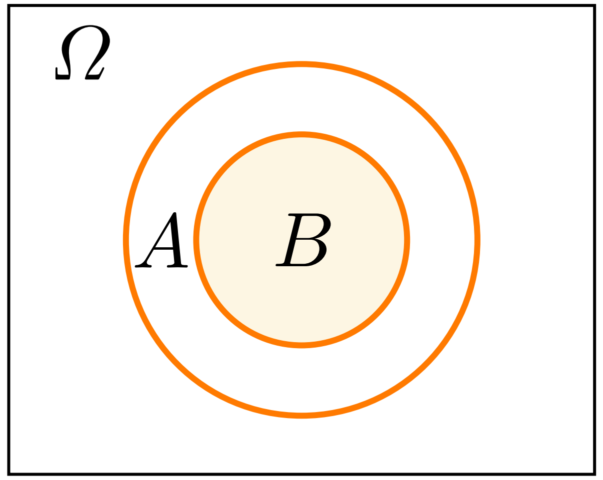
Einfaches Ereignis: Ereignis A innerhalb des Ergebnisraums Ω.

$A^c := \Omega - P(A)$



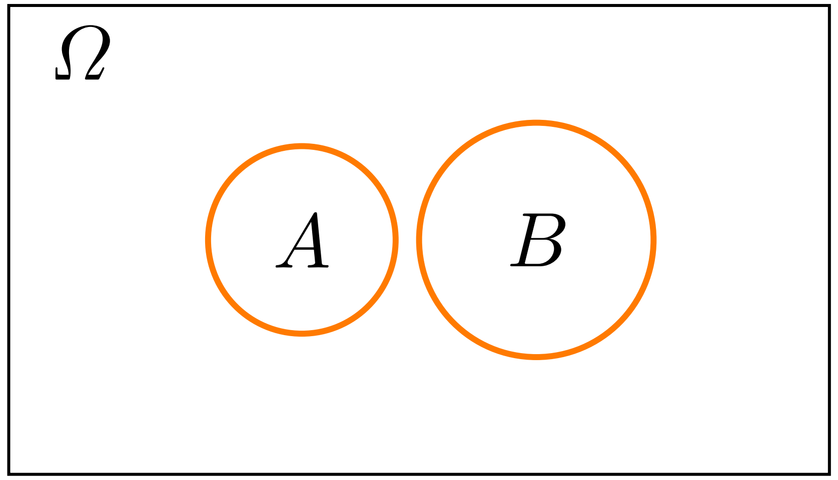
Gegenteil (Komplement): Alles, was *nicht* zu A gehört.

$B \subset A$



Teilereignis: B tritt nur auf, wenn auch A auftritt.

$A \cap B = \emptyset$



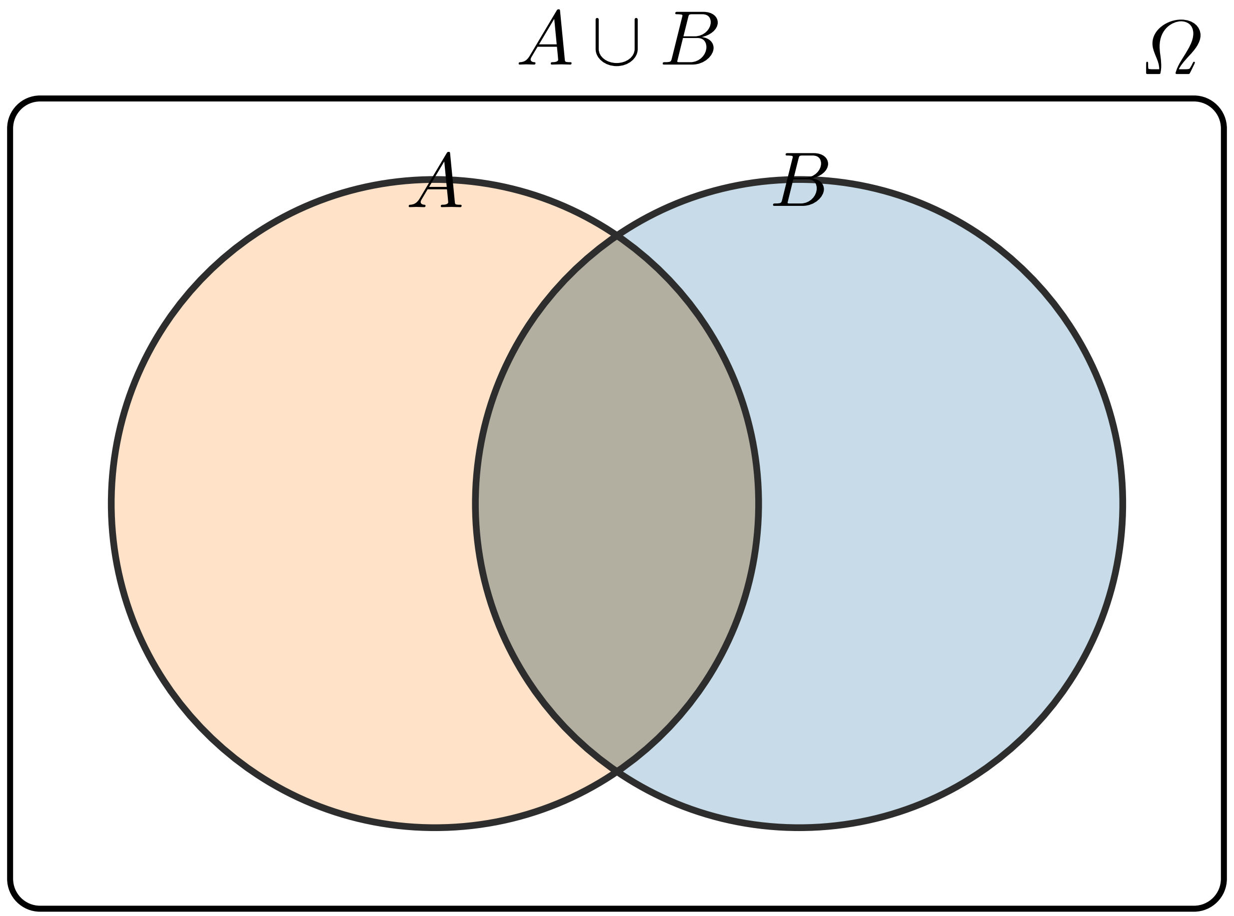
Unvereinbare Ereignisse: A und B können nicht gleichzeitig eintreten.

Das Gesetz der Vereinigung

Die Wahrscheinlichkeit von «A oder B» bezieht Überschneidungen ein.

- Formel: $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$
- Wenn A und B unabhängig sind: $Pr(A \cap B) = Pr(A) \times Pr(B)$
- Beispiel: Wahrscheinlichkeit, dass eine Person Kaffee oder Tee mag
- Die Überlappung sind Menschen, die beides mögen

Mini-Check:
Wenn $Pr(A) = 0.4$, $Pr(B) = 0.5$, $Pr(A \cap B) = 0.1 \rightarrow$
Wie gross ist $Pr(A \cup B)$?



$$\boxed{Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)}$$

Das Gesetz der Vereinigung berücksichtigt den Schnittbereich: «Oder» ist in der Wahrscheinlichkeit *inklusiv*.

Unabhängigkeit und Intuition

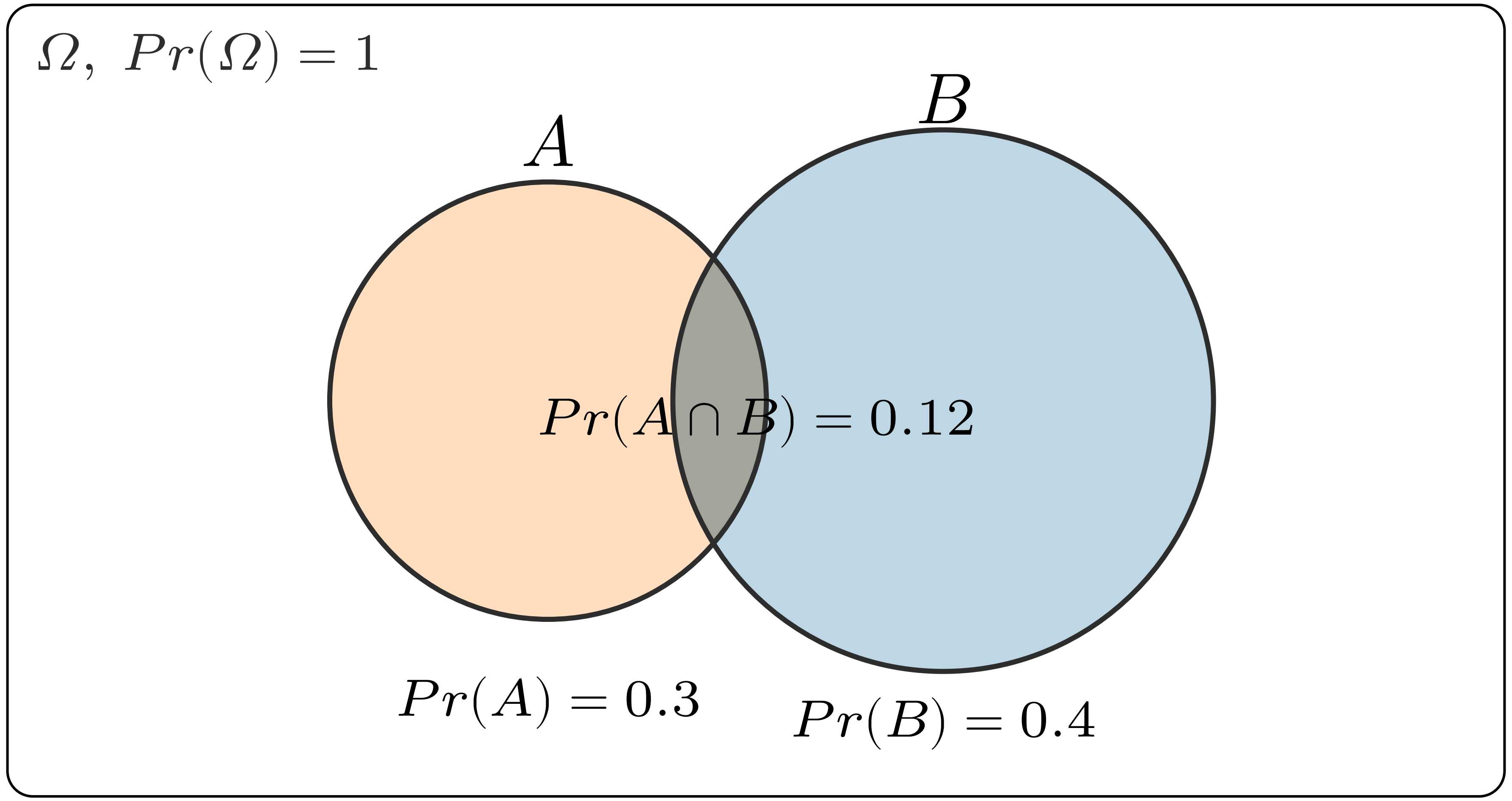
Unabhängigkeit bedeutet: Das eine Ereignis verändert nicht die Wahrscheinlichkeit des anderen.

- Formel: $A \perp B \Leftrightarrow Pr(A \cap B) = Pr(A) \times Pr(B)$
- Beispiel: Zwei Würfe eines Würfels
- Data-Science-Beispiel: $p(\text{Klick} \mid \text{Alter}) \neq p(\text{Klick}) \rightarrow \text{Abhängigkeit}$
- Merke: Unabhängigkeit ist selten, aber ein nützliches Modell

Mini-Check: Sind «Würfelergebnis des ersten Wurfs» und «Würfelergebnis des zweiten Wurfs» unabhängig?

Unabhängigkeit im Venn-Diagramm bedeutet *nicht*, dass sich die Ereignisse nicht überlappen[1], also dass $Pr(A \cap B) = 0$. Entscheidend ist das **Verhältnis der Flächen**.

Angenommen: $Pr(A) = 0.3$, $Pr(B) = 0.4$ und $Pr(A \cap B) = 0.12$.



Da $Pr(A \mid B) = Pr(A)$ gilt, folgt:

$$\frac{\text{Fläche von } A}{\text{Gesamtfläche}} = \frac{\text{Fläche von } A \cap B}{\text{Fläche von } B}.$$

[1] `https://www.youtube.com/watch?v=pV3nZAsJxl0`

Take-Away: Die Regeln des Zufalls

Hinter jeder komplexen Statistik stehen drei einfache Gesetze.

1. Nichtnegativität
2. Normierung
3. Additivität

→ Fundament der ganzen Theorie

Ergänzt um: Komplement, Vereinigung, Unabhängigkeit

Alle Formeln – von Bayes bis zum LLN, bauen genau darauf auf

Wer die Axiome versteht, kann jede Wahrscheinlichkeit begründen

Bedingte Wahrscheinlichkeit &

Bayes

Wenn neue Information dazukommt

Bedingte Wahrscheinlichkeit beschreibt, wie sich unser Wissen ändert, wenn etwas bekannt ist.

- $Pr(A \mid B)$ = Wahrscheinlichkeit von A, unter der Bedingung, dass B eingetreten ist
- Beispiel: «Was ist die Wahrscheinlichkeit, dass es regnet, wenn dunkle Wolken am Himmel sind?»
- $A = \text{Regen},\ B = \text{Wolken} \to Pr(A \mid B) > Pr(A)$
- Bedingung = zusätzliche Information, die den Ereignisraum verkleinert

Mini-Check: Wenn $Pr(A) = 0.2$, $Pr(B) = 0.5$, $Pr(A \cap B) = 0.15 \to$
Wie gross ist $Pr(A \mid B)$?

Formel und Intuition

Die Formel für bedingte Wahrscheinlichkeit spiegelt unsere Alltagserfahrung wider.

$$Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)}$$

- Nur der Anteil von A innerhalb von B zählt
- Beispiel: Wahrscheinlichkeit, dass jemand „Python kann“, gegeben „Data-Scientist“
- Wenn 80 % aller Data-Scientists Python nutzen $\rightarrow Pr(\text{Python} \mid \text{Data-Scientist}) = 0.8$
- Vertauschung: $Pr(B \mid A) \neq Pr(A \mid B)$

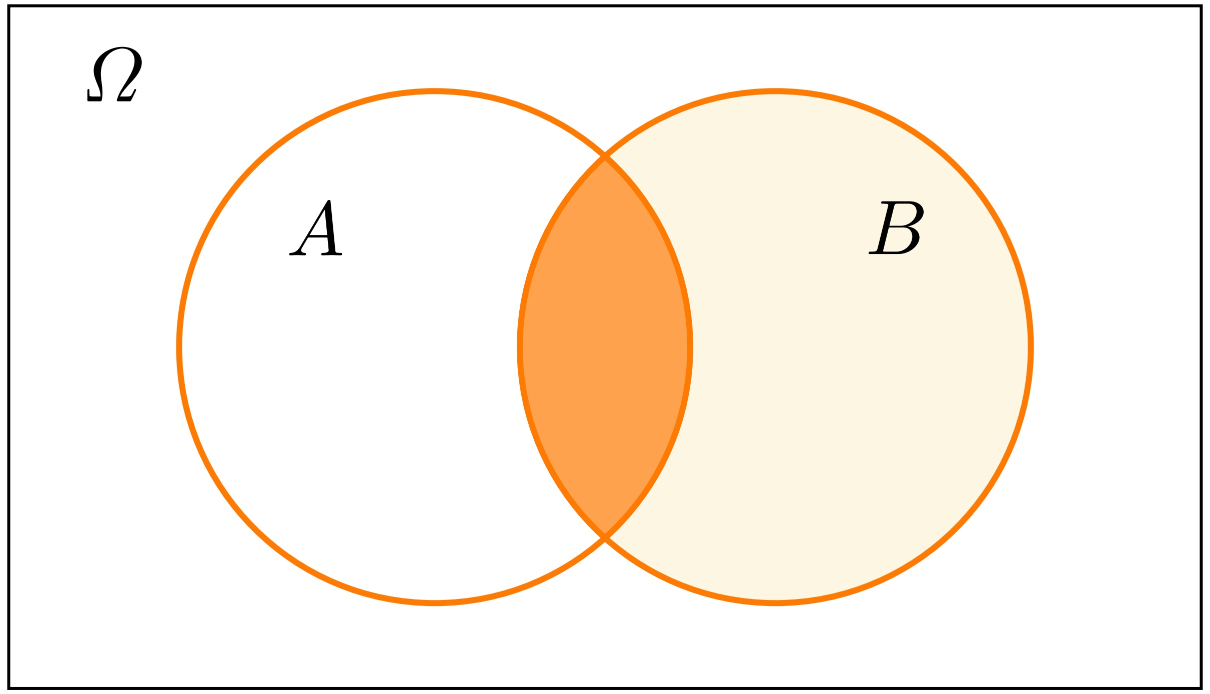
Mini-Check: Warum kann $Pr(A \mid B) \neq Pr(B \mid A)$ sein?

$\text{Pr}(A \mid B)$ heisst die **Wahrscheinlichkeit von A unter der Bedingung B**.

Formel:

$$\text{Pr}(A \mid B) = \frac{\text{Pr}(A \cap B)}{\text{Pr}(B)}, \quad \text{falls } \text{Pr}(B) > 0$$

Visualisierung:



Gedankenexperiment: Angenommen, wir wissen, dass unser Stein im Bereich B gelandet ist – wie gross ist dann die Wahrscheinlichkeit, dass er auch in A liegt?

Das Theorem von Bayes

Bayes zeigt, wie man Vorwissen mit Beobachtung kombiniert.

$$Pr(B \mid A) = \frac{Pr(A \mid B) \cdot Pr(B)}{Pr(A)}$$

Zerlegt Denken in drei Komponenten:
- Likelihood: $Pr(A \mid B) \rightarrow$ Wie plausibel sind die Daten, wenn B wahr ist?
- Prior: $Pr(B) \rightarrow$ Vorwissen
- Evidence: $Pr(A) \rightarrow$ Gesamtwahrscheinlichkeit der Daten

Ergebnis: Posterior = aktualisierte Überzeugung nach Beobachtung der Daten

Beispiel: Medizinischer Test: Krankheit (K) und positives Ergebnis (T+)

Mini-Check: Wie nennt man $Pr(K \mid T^{+})$?

LIKELIHOOD

The probability of "B" being True, given "A" is True

PRIOR

The probability "A" being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

POSTERIOR

The probability of "A" being True, given "B" is True

MARGINALIZATION

The probability "B" being True.

Beispiel: Medizinischer Test

Bayes entscheidet, wie wir Testergebnisse richtig interpretieren.

Gegeben:
- Sensitivität (Likelihood): $Pr(T^+ \mid K) = 0.95$
- Spezifität (negative Case): $Pr(T^- \mid \neg K) = 0.9$
- Prävalenz (Prior): $Pr(K) = 0.007$

Gesucht: (Posterior) $Pr(K \mid T^+)$

Berechnung:

$$\frac{0.95 \times 0.007}{0.95 \times 0.007 + 0.1 \times 0.993} \approx 0.062$$

Interpretation: Nur 6 % der positiven Tests sind tatsächlich krank

Intuition: Bei seltenen Ereignissen sind falsche Positive dominant

Mini-Check: Warum bedeutet «95 % Sensitivität» nicht «95 % sicher krank»?

Denken wie ein Bayesianer

Bayesisches Denken heisst: Wissen ständig anpassen, nicht dogmatisch bleiben.

- Wir starten mit Priors (Vorwissen): und lassen Daten sprechen.
- Machine Learning macht genau das: Modelle lernen Posterior Parameter aus Daten.
- Bayesische Modelle schätzen Unsicherheit statt sie zu verdrängen.
- Motto: «Neue Evidenz = neues Wissen.»

Mini-Check: Nenne ein Beispiel, wo du deine Meinung nach neuer Evidenz geändert hast.

Take-Away: Lernen aus Information (Bayes)

Bayes formt Denken: Wissen ist nie fix, sondern wird mit Daten aktualisiert.

- Bedingte Wahrscheinlichkeit: Wie ändert sich $Pr(A)$, wenn B bekannt ist?
- Bayes-Theorem: $\text{Vorwissen} \times \text{Daten} = \text{neue Überzeugung}$
- Praxis: Medizinische Tests, ML-Modelle, Risikoabschätzung

Risikomasse

Warum wir Risiken vergleichen

Statistik untersucht auch Unterschiede zwischen Wahrscheinlichkeiten.

- Motivation: «Erhöht Rauchen das Risiko für Lungenkrebs?» $\rightarrow$ Vergleich zweier Gruppen
- Ereignisnotation: $D = \text{Krankheit},\ E = \text{Risikofaktor (Rauchen)}$
- Wir vergleichen: $Pr(D \mid E)$ vs. $Pr(D \mid E^c)$
- Wenn beide gleich sind $\rightarrow$ Unabhängigkeit
- Wenn unterschiedlich $\rightarrow$ Assoziation

Mini-Check:
Wenn $Pr(D \mid E) = 0.20$ und $Pr(D \mid E^c) = 0.10$, was sagt das über E und D?

Risikodifferenz (absolute Risikoänderung)

Die Risikodifferenz zeigt den Zusatznutzen oder Schaden durch den Faktor E.

$$ER = Pr(D \mid E) - Pr(D \mid E^c)$$

- ER: Excess Risk, Risikodifferenz oder zusätzliches Risiko
- Beispiel: $0.20 - 0.10 = 0.10 \rightarrow$ 10 Prozentpunkte mehr Erkrankungen bei Rauchern
- Interpretation: «Absolute Wirkung»
- In Data Science: z. B. Conversion-Rate-Differenz in A/B-Tests

Mini-Check: Was bedeutet $ER < 0$?

Relatives Risiko und Odds Ratio

Relative Masse zeigen den Faktor, um den ein Risiko steigt.

- Relatives Risiko:
  $$RR = \frac{Pr(D \mid E)}{Pr(D \mid E^c)}$$
- Beispiel: $0.20/0.10 = 2 \to$ zweifach erhöhtes Risiko
- Odds Ratio:
  $$OR = \frac{Pr(D \mid E)/(1 - Pr(D \mid E))}{Pr(D \mid E^c)/(1 - Pr(D \mid E^c))}$$

Näher an der ML-Welt: Logistic Regression modelliert $\log(OR)$

Bei seltenen Ereignissen: $OR \approx RR$

Mini-Check: Wenn $RR = 1$, was heisst das?

Exposition	**Dupuytren (D)**	**Kein Dupuytren (D^c)**	**Gesamt**
Männer (E)	$92/763 = 0.1206$	$256/763 = 0.3355$	$348/763 = 0.4561$
Frauen (E^c)	$77/763 = 0.1009$	$338/763 = 0.4430$	$415/763 = 0.5439$
Gesamt	$169/763 = 0.2215$	$594/763 = 0.7785$	1

Berechnung der Risikomasse:

$$\text{Exzessrisiko (ER)} = Pr(D \mid E) - Pr(D \mid E^c) = 0.1206 - 0.1009 = 0.0197$$

$$\text{Relatives Risiko (RR)} = \frac{Pr(D \mid E)}{Pr(D \mid E^c)} = \frac{0.1206}{0.1009} = 1.195$$

$$\text{Odds Ratio (OR)} = \frac{Pr(D \mid E)/Pr(D^c \mid E)}{Pr(D \mid E^c)/Pr(D^c \mid E^c)} = \frac{0.1206/0.3355}{0.1009/0.4430} = 1.59$$

Interpretation: Männer haben ein leicht erhöhtes Risiko für Dupuytren (RR 1.2); die Chance auf Erkrankung ist etwa 1.6-mal höher (OR 1.6); das absolute Mehr an Risiko beträgt rund 0.02 (ER 0.02).

Das Simpson-Paradoxon

Ein Gesamtergebnis kann täuschen, wenn eine dritte Variable mitspielt.

- Beispiel: Erfolg von zwei Behandlungen bei Nierensteinen (≙ SDS Tabelle 3.5)
- Gesamt: Minimal-invasiv besser (83 % vs. 78 %)
- Nach Grösse stratifiziert: Offene OP besser in beiden Gruppen!
- Erklärung: Verzerrung durch ungleiche Verteilung des Schwierigkeitsgrads
- Moral: Immer prüfen, ob eine verdeckte Variable (Effekt C) die Beziehung dreht

Mini-Check: Wie nennt man eine Variable, die eine scheinbare Abhängigkeit erzeugt?

Nierensteine 2 cm (C)

Behandlungsart	**Erfolg (D)**	**Kein Erfolg (D^c)**	**Gesamt**
Minimal-invasiv (E)	234	36	270
Offene Operation (E^c)	81	6	87
Total	315	42	357

Nierensteine $>$ 2 cm (C^c)

Behandlungsart	**Erfolg (D)**	**Kein Erfolg (D^c)**	**Gesamt**
Minimal-invasiv (E)	55	25	80
Offene Operation (E^c)	192	71	263
Total	247	96	343

Interpretation:

Innerhalb beider Gruppen höhere Erfolgsrate für offene Operation (93.1 % vs. 86.7 % und 73.0 % vs. 68.8 %).
Gesamt zeigt Minimal-invasiv scheinbar höhere Erfolgsrate (82.6 % vs. 78.0 %):
das klassische **Simpson-Paradoxon**.

Was wir daraus lernen

Risikoanalyse ist nicht nur Rechnen, sondern Kausalitäts-Denken.

- Prüfe immer: «Vergleiche ich Äpfel mit Äpfeln?»
- Kontrolliere Kontext- oder Confounder-Variablen
- In Data Science: Segmentierung oder Stratifizierung vor Interpretation
- Statistische Frage: Gibt es einen echten Effekt oder nur eine ungleiche Stichprobe?

Mini-Check:

Wie würdest du verhindern, dass Simpson's Paradoxon deine Analyse täuscht?

Take-Away: Risiko verstehen, Verzerrung erkennen

Wahrscheinlichkeiten werden erst sinnvoll im Vergleich und im richtigen Kontext.

- Risk Diff / Rel Risk / Odds Ratio zeigen unterschiedliche Blickwinkel
- Simpson-Paradoxon mahnt: Prüfe immer auf versteckte Variablen
- $\text{Statistik} \neq \text{Wahrheit}$, sondern Abwägen zwischen Erklärungen
- Kontext entscheidet, ob ein Risiko real oder scheinbar ist

Zufallsvariablen & Verteilungen

Vom Ereignis zur Zufallsvariablen

Eine Zufallsvariable ordnet jedem Zufallsexperiment einen Zahlenwert zu: das macht Wahrscheinlichkeit rechenbar.

- Beispiel: Würfeln $\rightarrow$ Ergebnis {1, 2, 3, 4, 5, 6} $\rightarrow$ Zufallsvariable $X = \text{Augenzahl}$
- Ereignis «gerade Zahl» = {2, 4, 6} entspricht $X \bmod 2 = 0$
- Idee: Zufall $\rightarrow$ Zahl $\rightarrow$ Statistik
- So werden Daten Realisationen von X

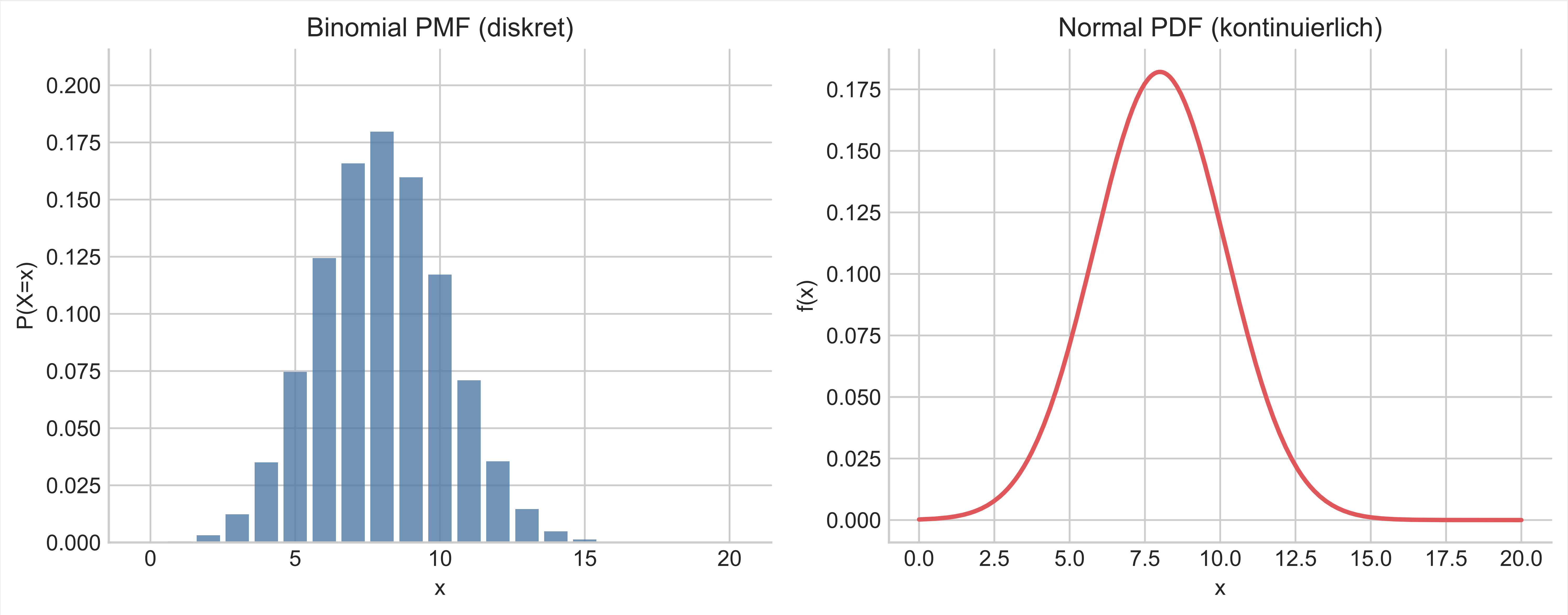
Mini-Check: Was ist bei einem Münzwurf eine Zufallsvariable X?

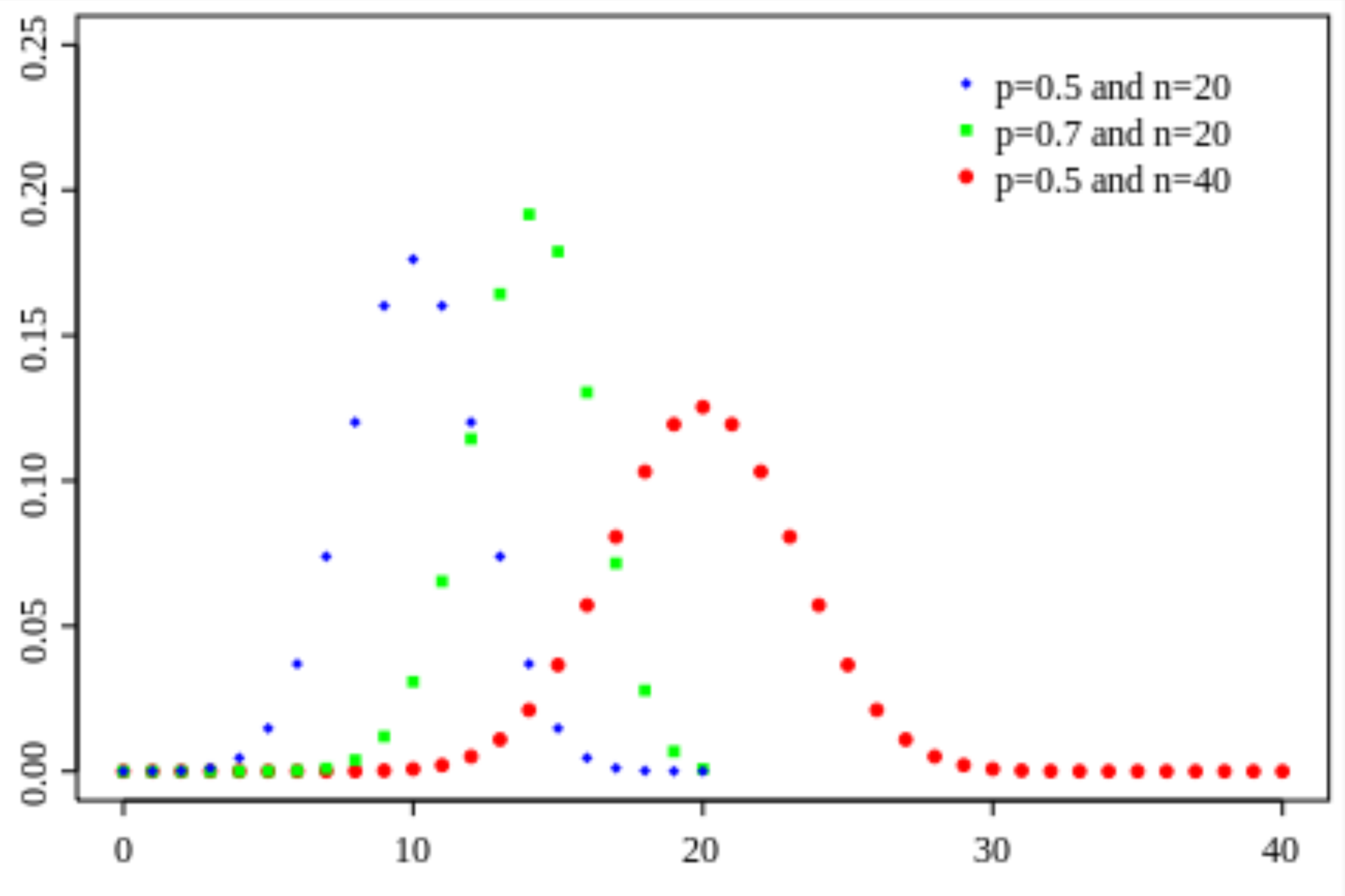
Diskrete und kontinuierliche Zufallsvariablen

Zufall kann zählen oder messen: das ändert die Art der Verteilung.

- Diskret (Nominal, Ordinal): endliche oder zählbare Werte (z. B. Augenzahl, Klicks, Fehler)
- Kontinuierlich (Intervall, Ratio): beliebige reelle Werte (z. B. Messungen, Zeit, Temperatur)
- Notation: PMF (für diskrete X) und PDF (für kontinuierliche X)
- PMF: Probability Mass Function (Wahrscheinlichkeitsfunktion)
- PDF: Probability Density Function (Wahrscheinlichkeitsdichtefunktion)
- Unterschied: Summen vs. Integrale

Mini-Check: Ist «Messfehler in mm» diskret oder kontinuierlich?


Binomial PMF (diskret)
P(X=x)
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
0
5
10
15
20
x
Normal PDF (kontinuierlich)
f(x)
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
x



p=0.5 and n=20
p=0.7 and n=20
p=0.5 and n=40
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40

Wahrscheinlichkeitsdichte f(x) und Verteilungsfunktion F(x)

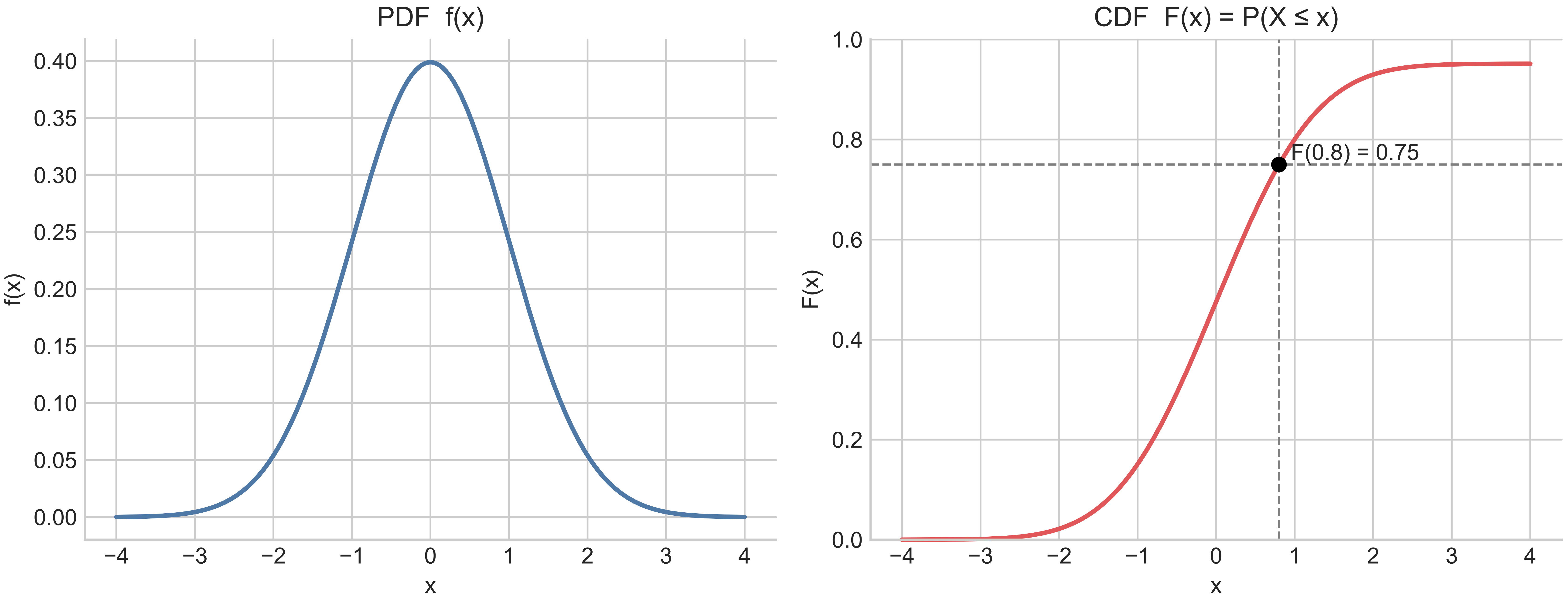
Die Dichte zeigt, wo Werte wahrscheinlicher sind: die Verteilungsfunktion summiert sie auf.

- PDF: Probability Density Function (Wahrscheinlichkeitsdichtefunktion)
  $f(x) \geq 0$ und $\int f(x)\,dx = 1$
- CDF: Cumulative Distribution Function (Kumulative Verteilungsfunktion)
  $F(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(z)\,dz$

Beispiel: Normalverteilung $\rightarrow$ glatte S-Kurve in $F(x)$

Wichtig: $f(x) \neq Pr(X = x)$
Bei kontinuierlichen X gilt:
$Pr(X = x) = 0$

Mini-Check:
Was bedeutet $F(0.8) = 0.75$?

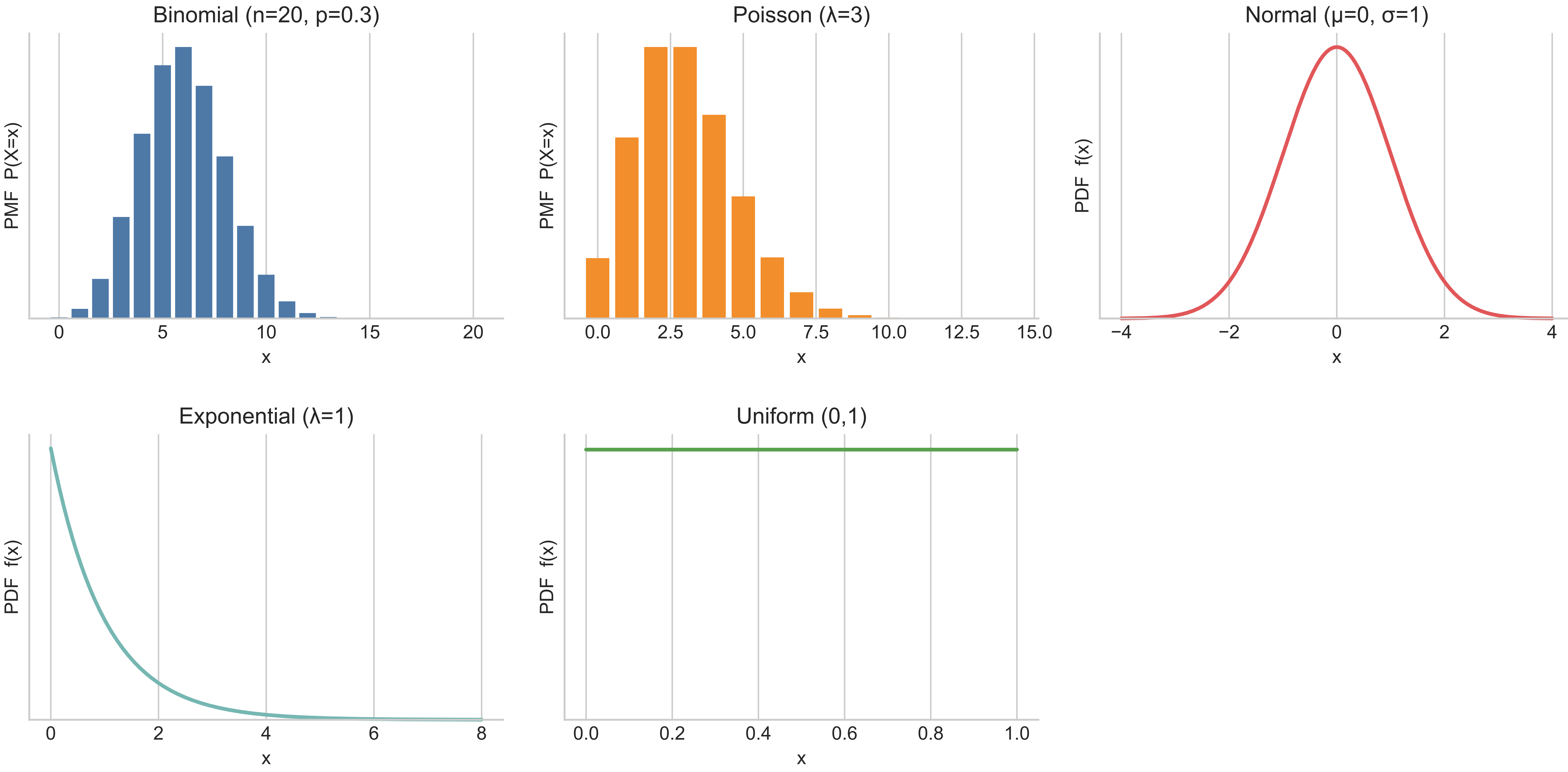

PDF f(x)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
f(x)
−4
−3
−2
−1
0
1
2
3
4
x
CDF F(x) = P(X ≤ x)
1.0
0.8
0.6
0.4
0.2
0.0
F(x)
F(0.8) = 0.75
x

Bekannte Verteilungen im Überblick

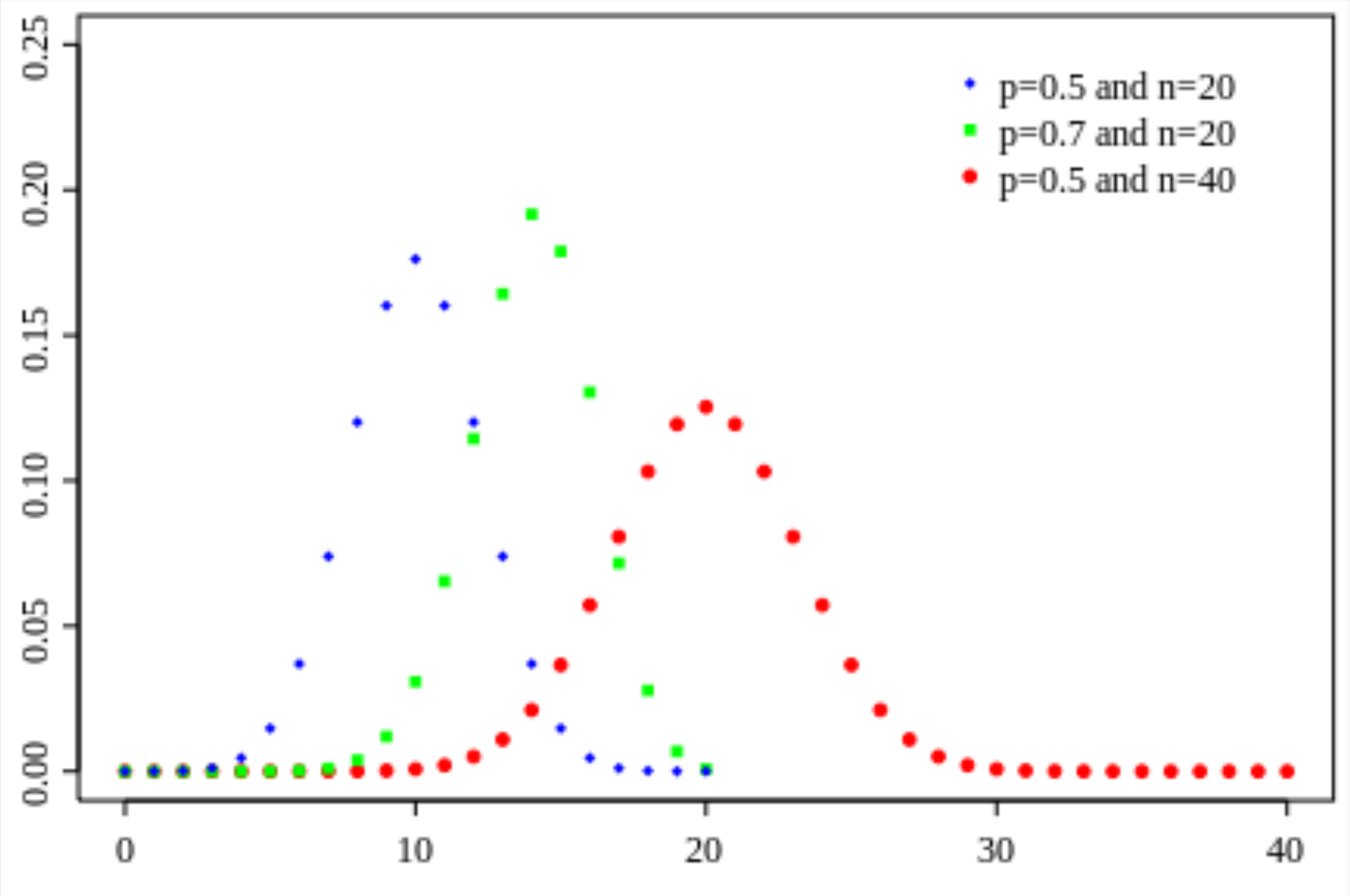
Hinter fast jedem Data-Science-Problem steckt eine klassische Verteilung.

- Binomial: Zählprozesse, n Versuche, p Erfolg $\rightarrow$ z. B. Klicks pro 100 User
- Poisson: seltene Ereignisse im Intervall $\rightarrow$ z. B. Server-Fehler pro Minute
- Normal: Summeneffekte, Mittelwerte $\rightarrow$ Messungen und Fehler
- Exponential: Zeit bis zum nächsten Ereignis
- Uniform: vollständige Unkenntnis: alle gleich wahrscheinlich

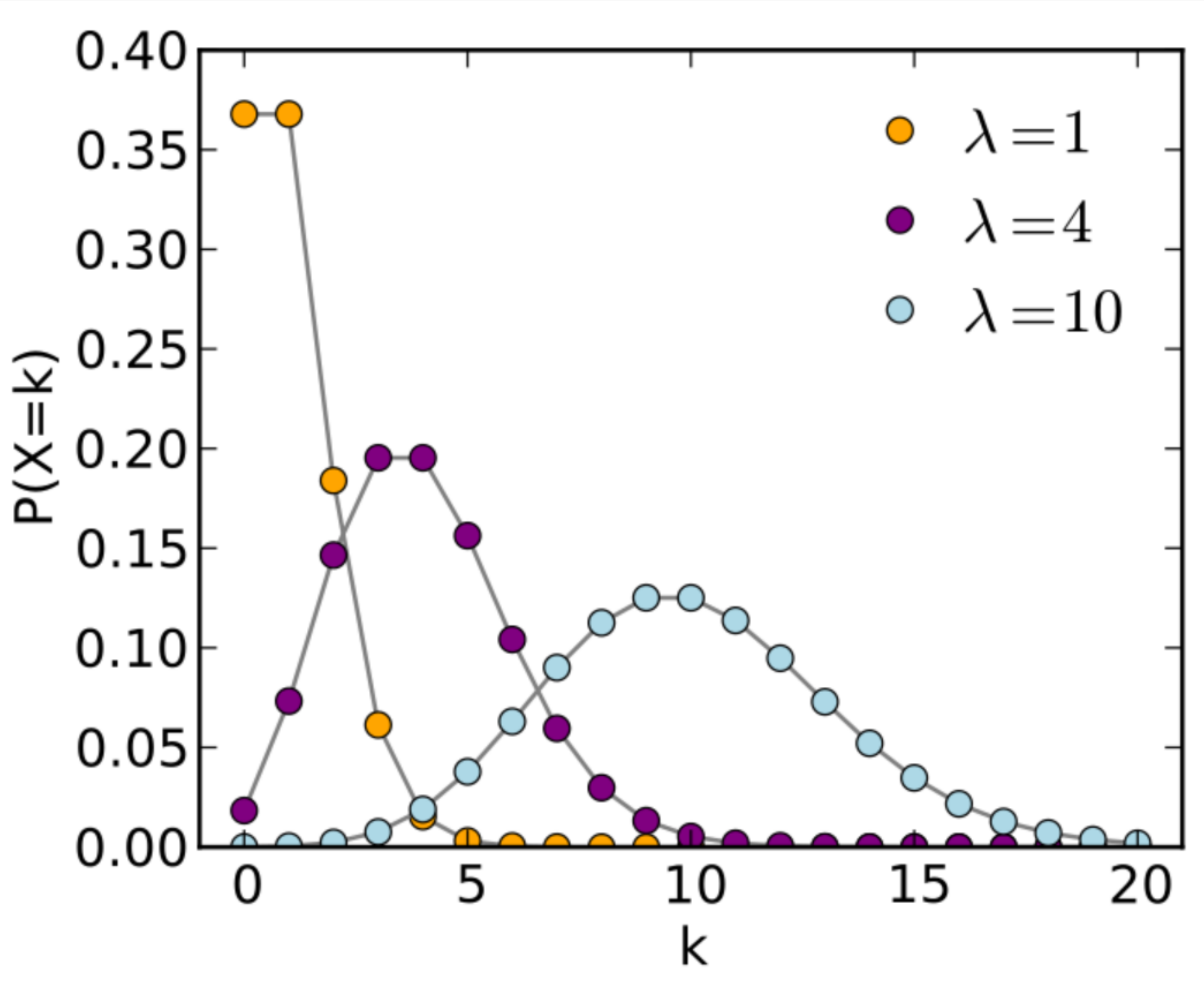
Mini-Check:
Welche Verteilung würdest du für «Anzahl Support-Tickets pro Tag» verwenden?


Binomial (n=20, p=0.3)
PMF P(X=x)
0
5
10
15
20
x
Poisson (λ=3)
PMF P(X=x)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
x
Normal (μ=0, σ=1)
PDF f(x)
−4
−2
0
2
4
x
Exponential (λ=1)
PDF f(x)
0
2
4
6
8
x
Uniform (0,1)
PDF f(x)
0.0
0.2
0.4
0.6
0.8
1.0
x

Binominal


p=0.5 and n=20
p=0.7 and n=20
p=0.5 and n=40
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40

Poisson



$\lambda=1$
$\lambda=4$
$\lambda=10$
P(X=k)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
k

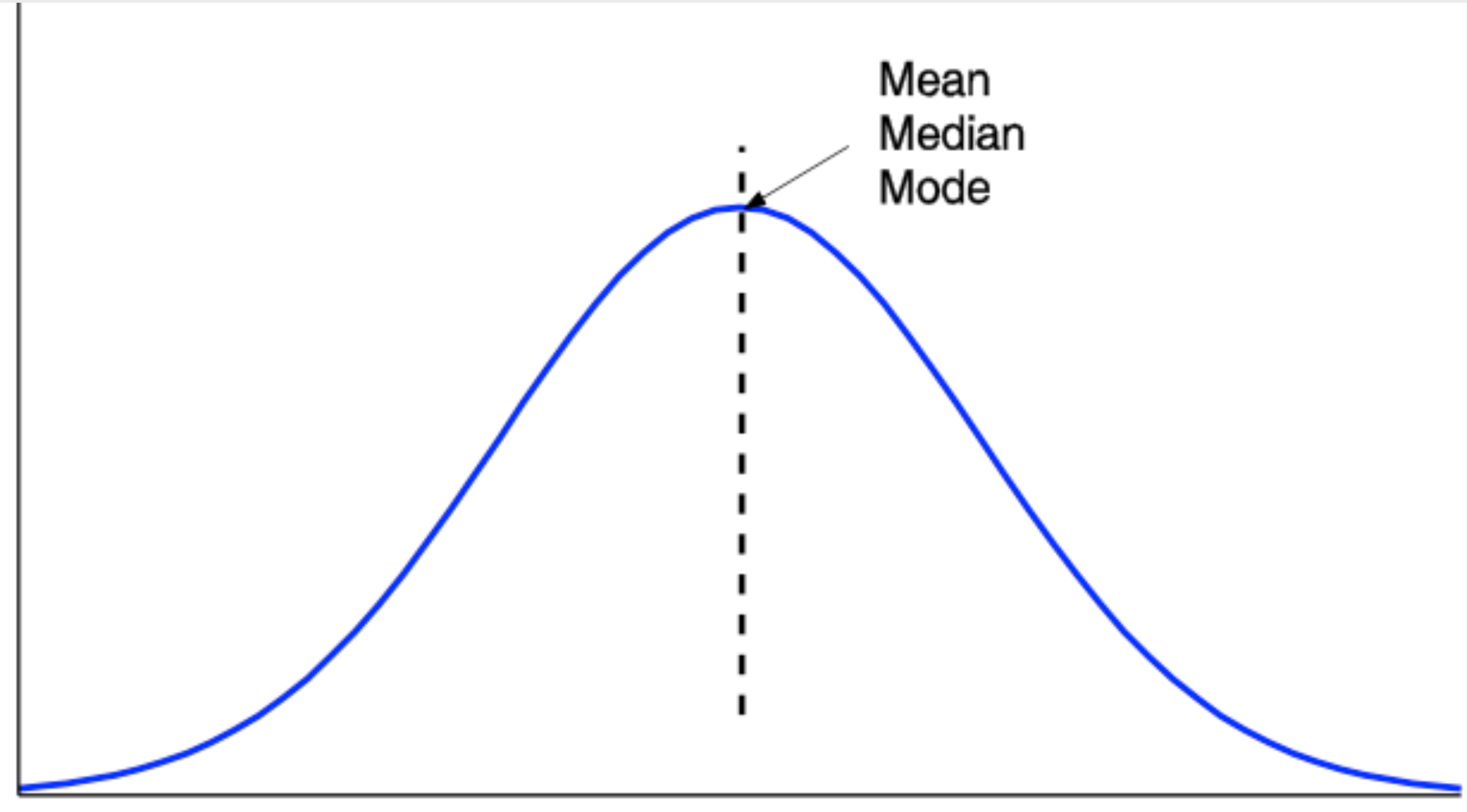
Visualisierung: Formen und Bedeutung

Die Form der Verteilung erzählt die Geschichte des Datensatzes.

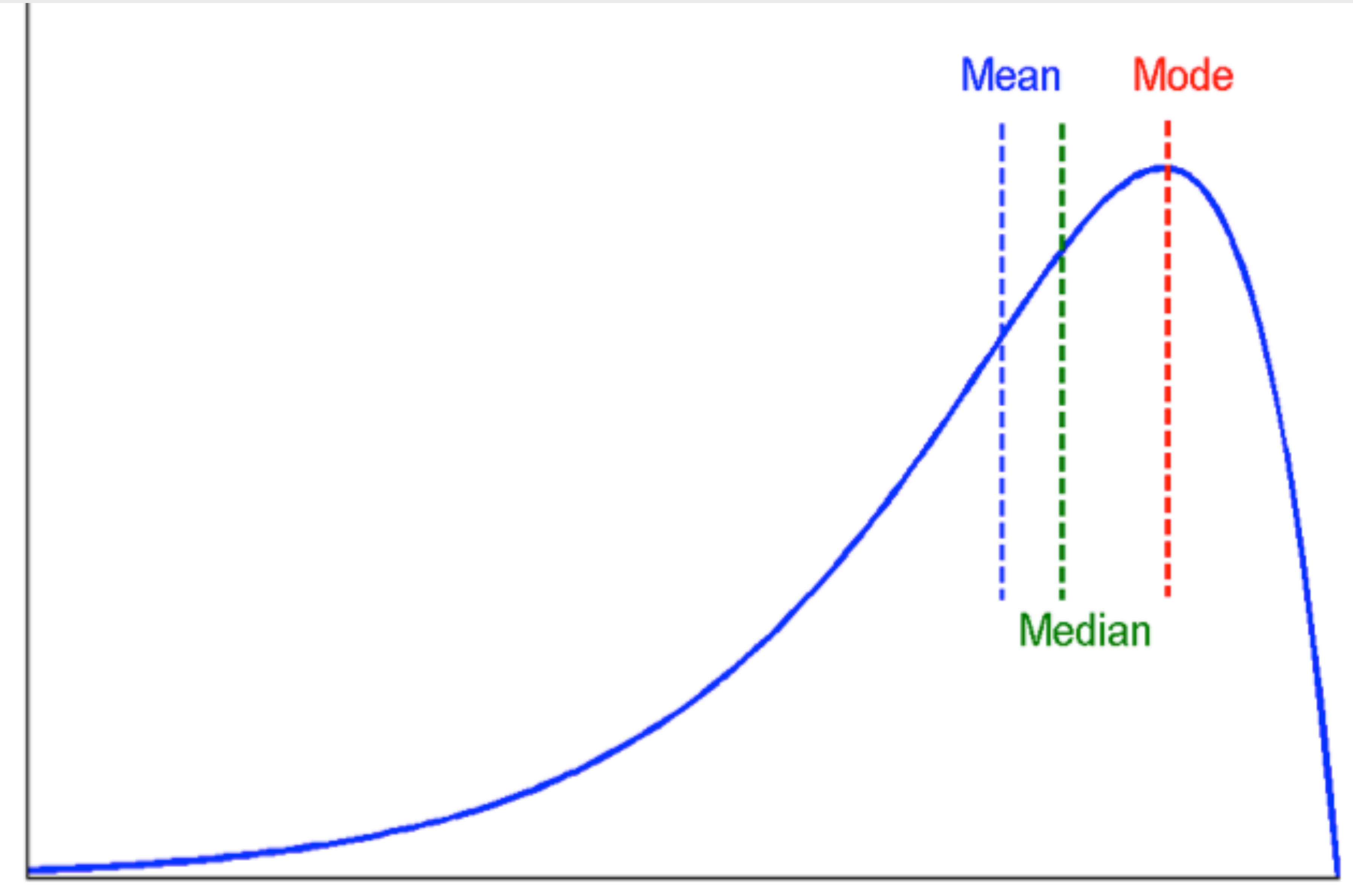
- Symmetrisch: Normal $\rightarrow$ Fehler verteilen sich gleichmässig
- Rechtssteil: Einkommen, Antwortzeiten, Poisson – viele kleine, wenige grosse Werte
- Linkssteil: z. B. Fehlerbewertungen 1–5 (Schulnoten)
- In der Praxis: Data Scientists nutzen Histogramme, QQ-Plots und KDEs, um die Verteilungsform zu prüfen

Mini-Check: Wie würde eine rechtsschiefe Verteilung aussehen?

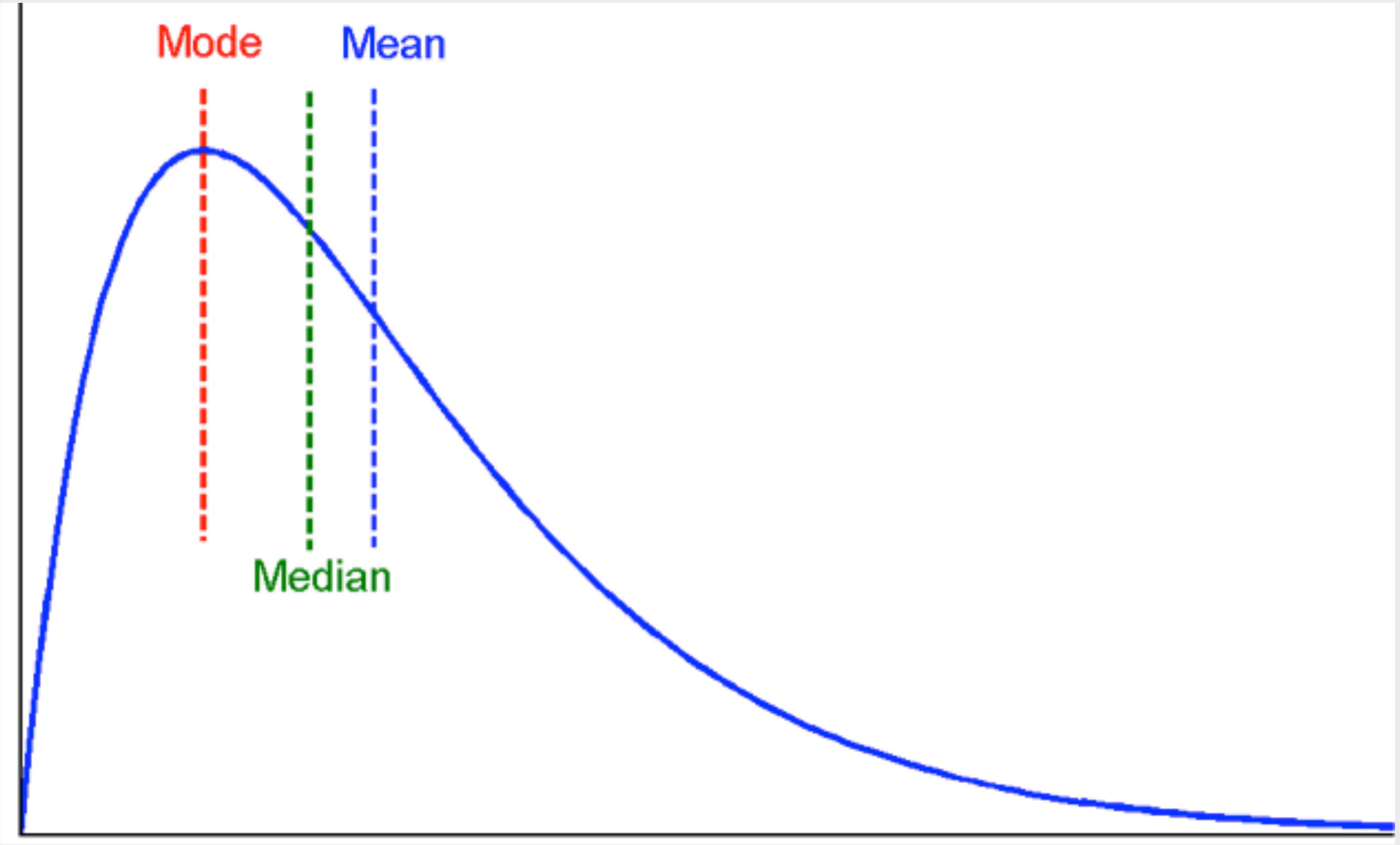
Symmetrisch



Linksschief



Rechtschief



Take-Away: Vom Zufall zur Form

Jede Verteilung erzählt eine Geschichte darüber, wie Daten entstehen

- Diskret vs. kontinuierlich: zählen vs. messen
- Formen zeigen Mechanismen: Symmetrie, Schiefe, Ausreisser
- Bekannte Verteilungen sind wiederkehrende Naturgesetze von Daten
- Data Science nutzt sie, um Modelle zu prüfen und anzupassen

Erwartungswert &

Gesetz der grossen Zahlen (LLN)

Was ist der Erwartungswert?

Der Erwartungswert ist der «Schwerpunkt» einer Verteilung: das mathematische Mittel aller möglichen Ausgänge.

- Diskret: $E(X) = \sum x_i \cdot Pr(X = x_i)$
- Kontinuierlich: $E(X) = \int x \cdot f(x)\, dx$
- Beispiel Würfel: $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$

Kein einzelner Wurf ergibt 3.5: es ist der langfristige Durchschnitt

Mini-Check: Was ist der Erwartungswert einer fairen Münze ($X = 1$ bei Kopf, 0 bei Zahl)?

Varianz und Standardabweichung

Die Varianz zeigt, wie weit Zufallswerte vom Erwartungswert entfernt liegen.

$$Var(X) = E[(X - E(X))^2]$$

$$SD = \sqrt{Var(X)}$$

- Beispiel: Zwei Maschinen mit gleichem Durchschnitt, aber eine streut stärker
- In Data Science: Stabilität von Modellmetriken (z. B. Cross-Val-Scores)

Mini-Check: Was bedeutet eine kleine Standardabweichung?

Gesetz der grossen Zahlen (LLN)

Law of Large Numbers (LLN): Viele Zufälle ergeben Regelmässigkeit.

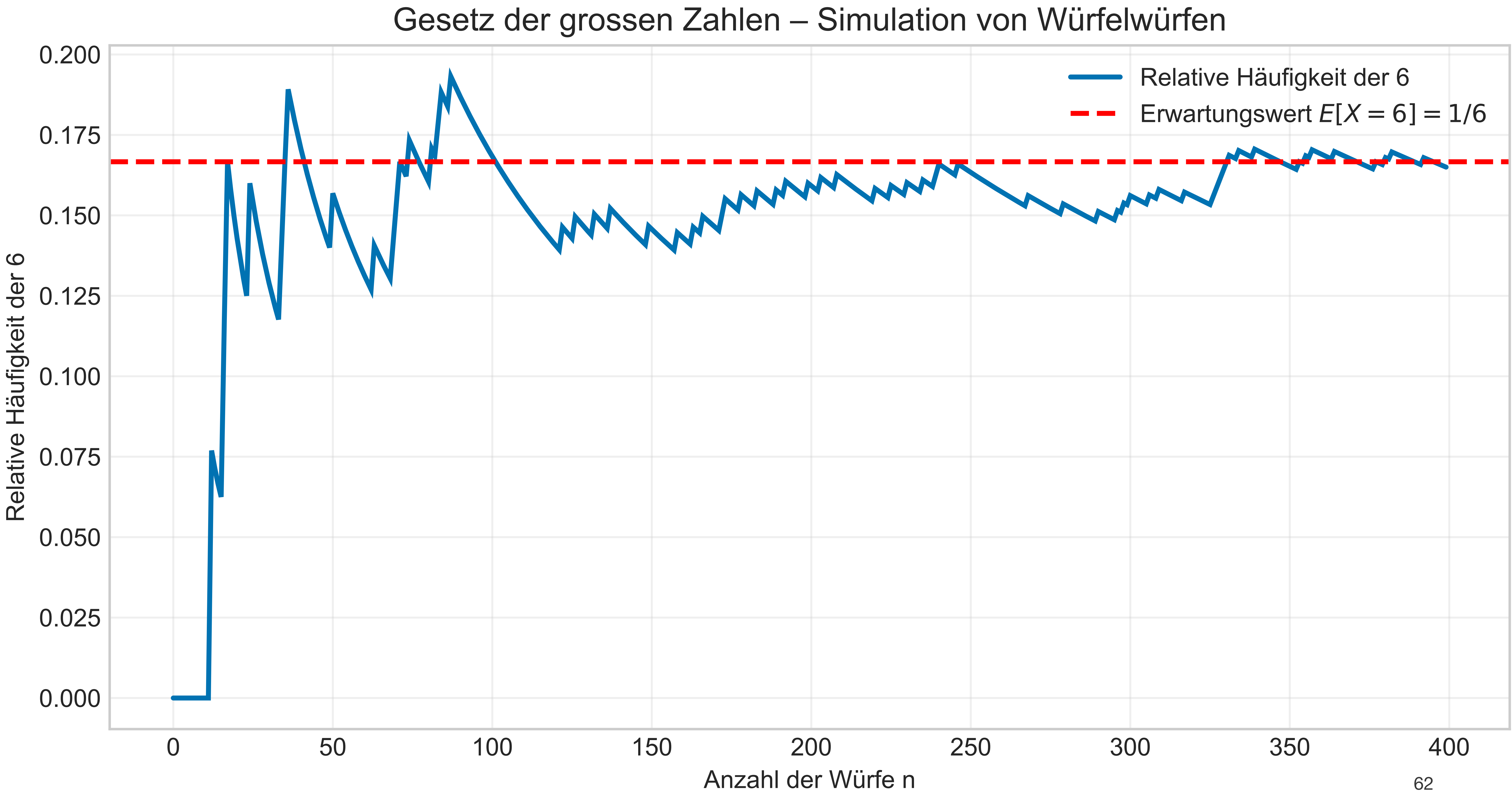
Wenn $X_1, \ldots, X_n$ unabhängig und identisch verteilt sind, gilt:

$$\overline{X}_n = \frac{1}{n} \sum X_i \;\rightarrow\; E(X)$$

Mit wachsendem n konvergiert der Stichprobenmittelwert gegen den Erwartungswert

- Beispiel: Münzwurf: Anteil Kopf $\rightarrow$ 0.5 bei grossem n
- Grundlage von Modellevaluation, Monte-Carlo und Predictive Stability

Mini-Check: Was passiert mit dem Stichprobenmittel, wenn $n \rightarrow \infty$?



Gesetz der grossen Zahlen – Simulation von Würfelwürfen
Relative Häufigkeit der 6
Erwartungswert $E[X = 6] = 1/6$
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Relative Häufigkeit der 6
0
50
100
150
200
250
300
350
400
Anzahl der Würfe n

Zufall wird vorhersagbar

Das Gesetz der grossen Zahlen erklärt, warum Durchschnittswerte so mächtig sind.

- Einzelbeobachtung = Lärm, viele Beobachtungen = Signal
- Deshalb arbeiten Data Scientists mit Aggregaten (Mittelwerten, Anteilen)
- LLN bedeutet: «Rauschen löscht sich im Schnitt aus»
- Beispiel: Online-Experiment – Klickrate stabilisiert sich nach tausenden Usern

Mini-Check: Was passiert, wenn dein Sample zu klein ist?

Von Wahrscheinlichkeit zu Data Science

Wahrscheinlichkeiten machen Vorhersage messbar und damit verbesserbar.

- Erwartungswert: Grundlage für Loss-Functions (MSE, Cross-Entropy)
- Varianz: Vertrauensbereich und Unsicherheitsquantifizierung
- LLN: Validierung und Reproduzierbarkeit

Mini-Check: Wie spiegelt sich das LLN im Training von ML-Modellen wider?

Take-Away: Wenn Zufall sich beruhigt

Je mehr Daten, desto klarer das Muster: darum funktioniert Statistik

- Erwartungswert: Zentrum des Zufalls
- Varianz: Unsicherheit um dieses Zentrum
- Gesetz der grossen Zahlen (LLN): Mittelwerte konvergieren $\rightarrow$ Zufall wird berechenbar
- Data Science: Grosse Stichproben $\rightarrow$ stabile Modelle

Zusammenfassung & Ausblick

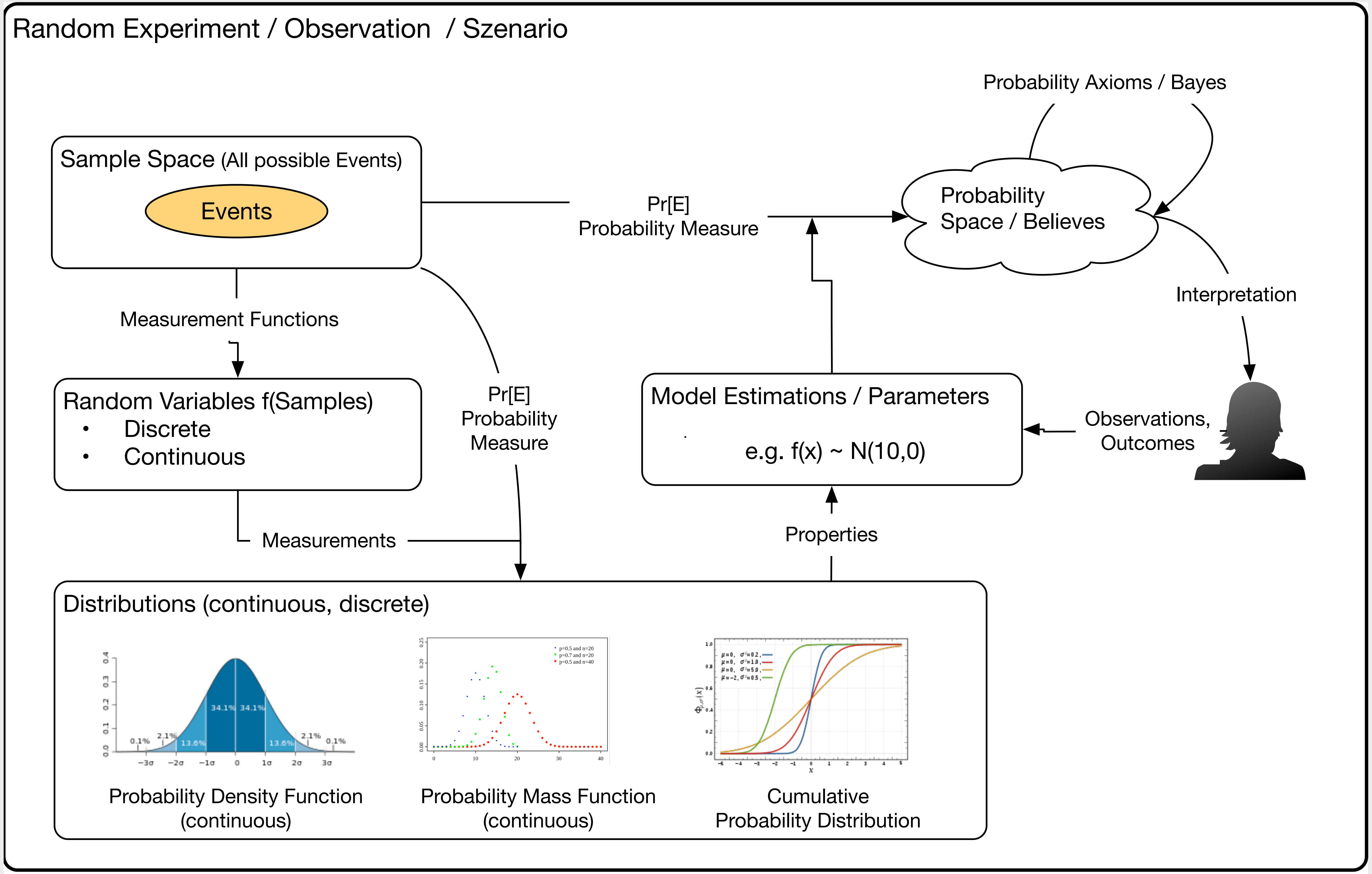


Image sources: Wikipedia

Gesamt-Take-Away: Die Ordnung im Zufall

Data Science ist die Kunst, Zufall in Struktur zu verwandeln: mit Wahrscheinlichkeit als Werkzeug

- Wahrscheinlichkeit gibt Chaos eine Grammatik
- Axiome liefern die Regeln, Bayes das Denken, Verteilungen die Formen, LLN das Vertrauen
- Mit genügend Daten verschwindet der Zufall nicht, er stabilisiert sich
- Das ist die Grundlage jeder Vorhersage, jedes Modells, jeder Entscheidung

Quiz - Aktive Wiederholung

Kahoot Quiz VL5: Wahrscheinlichkeit & Verteilungen

Abschluss & Ausblick

Heute gelernt
- Axiome, Ereignisse, Komplement, Vereinigung, Unabhängigkeit
- Bedingte Wahrscheinlichkeit und Bayes als Denken mit Information
- Zufallsvariablen, Formen von Verteilungen
- Erwartungswert, Varianz, Gesetz der grossen Zahlen

Nächste Sitzung: VL6
- Schätzen und Konfidenzintervalle
- Vorbereitung und Details: siehe syllabus.md