Statistik – Kapitel 0

Karsten Keßler

19.01.2026

Inhaltsverzeichnis

0 Einführung

0.1 Einführung in die Statistik

Leitfrage:

Wie kommt man von den „rohen Zahlen” zu begründeten Aussagen und Entscheidungen?

Lernziele:
(i) grundlegende Begriffe der Statistik sauber verwenden,
(ii) statistische Fragestellungen präzise formulieren und
(iii) den Unterschied zwischen Deskription und Induktion erklären.

Notebooks
00 – Einführung

Rechenbeispiele
00 – Übungen

0.1.1 Was bedeutet „Statistik"?

Der Begriff **Statistik** wird in der Praxis in drei typischen Bedeutungen verwendet:

0.1.1.1 Statistik als Datenmaterial Veröffentlichte Zahlen, Tabellen oder Kennzahlen (z. B. Umsatzstatistik, Arbeitslosenstatistik, Wahlstatistik).
→ Statistik als **Ergebnis** einer Datenerhebung.

Informatik-Beispiele:

- Server-Uptime-Statistiken (99,9% Verfügbarkeit)
- GitHub-Repository-Metriken (Stars, Forks, Contributors)
- App-Store-Bewertungsstatistiken
- Website-Traffic-Reports (Visits, Bounce-Rate)
- Fehlerstatistiken aus Bug-Tracking-Systemen

0.1.1.2 Statistik als Methode Verfahren zur **Erhebung**, **Aufbereitung**, **Darstellung** und **Analyse** von Daten:
Häufigkeitstabellen, Diagramme, Mittelwert/Varianz, Modelle, Tests.

Informatik-Beispiele:

- Log-Analyse und Aggregation (ELK-Stack, Splunk)
- A/B-Testing-Frameworks
- Performance-Monitoring (Grafana, Prometheus, Datadog)

- Machine-Learning-Evaluationsmetriken (Accuracy, F1-Score)
- Code-Qualitätsmetriken (SonarQube)

0.1.1.3 Statistik als Wissenschaft Mathematische Disziplin, die Modelle für **Zufall und Unsicherheit** entwickelt und untersucht, wie zuverlässig Schlussfolgerungen aus Stichproben sind.

Informatik-Bezüge:

- Wahrscheinlichkeitstheorie für Kryptographie
- Stochastische Prozesse für Netzwerk- und Warteschlangenanalyse
- Bayessche Inferenz für Machine Learning
- Statistische Lerntheorie (PAC-Learning, VC-Dimension)
- Information Theory (Shannon-Entropie)

Merksatz
In dieser Vorlesung steht Statistik vor allem als **Methode** und **Wissenschaft** im Vordergrund.

0.1.2 Warum brauchen wir Statistik?

In vielen realen Situationen sind Ergebnisse **nicht deterministisch** und Daten **unvollständig**:

- In der Qualitätskontrolle wird nicht jedes Teil geprüft.
- In Umfragen werden nicht alle Personen befragt.
- In IT-Systemen schwanken Antwortzeiten oder Ausfallraten.
- In der Wirtschaft variieren Nachfrage, Preise und Lieferzeiten.

Statistik liefert Werkzeuge, um

- Daten verständlich zu strukturieren,
- Zufall von Struktur zu trennen,
- Unsicherheit zu quantifizieren,
- Entscheidungen nachvollziehbar zu begründen.

Wichtig:
Statistik liefert selten Gewissheit, sondern **begründete Aussagen mit Fehlerwahrscheinlichkeit**.

0.1.3 Warum Informatiker Statistik brauchen

Anwendungsbereich	Typische statistische Fragen
Software-Testing	Wie viele Tests reichen für ausreichende Abdeckung?
Performance-Optimierung	Ist die neue Version wirklich schneller?
Machine Learning	Wie gut generalisiert mein Modell auf neue Daten?
Security/Anomalie-Erkennung	Ist dieses Login-Verhalten verdächtig?
Kapazitätsplanung	Wie viele Server brauchen wir für Spitzenlasten?
A/B-Testing	Hat die UI-Änderung einen signifikanten Effekt?
Datenqualität	Wie gehen wir mit fehlenden oder fehlerhaften Daten um?
SLA-Monitoring	Erfüllen wir die vereinbarten Service Level?

0.1.4 Das Problem der Variabilität

Selbst bei identischen Bedingungen variieren IT-Metriken:

Beispiel: Dieselbe Datenbankabfrage, 5× ausgeführt:

Messung	Zeit (ms)
1	12,3
2	14,1
3	11,8
4	45,2
5	13,5

Fragen:

- Was ist die „typische" Antwortzeit?
  → *Mittelwert* oder *Median*? (siehe Kapitel 2)
- Ist Messung 4 ein Problem oder normal?
  → *Ausreißererkennung* mit Quartilen und Boxplots (siehe Kapitel 2)
- Wie zuverlässig ist ein einzelner Messwert?
  → *Konfidenzintervalle* für die mittlere Antwortzeit (siehe Kapitel 6)

→ Statistik liefert die Werkzeuge, um solche Fragen systematisch zu beantworten.

0.1.5 Grundbegriffe

0.1.5.1 Statistische Einheit (Merkmalsträger) Die **statistische Einheit** ist das Objekt, über das Daten erhoben werden (kleinste Untersuchungseinheit).

Allgemeine Beispiele: Person, Bestellung, Produkt, Bauteil, Messung

Informatik-Beispiele:

Kontext	Statistische Einheit
Web-Analytics	Ein Seitenaufruf (Page View)
API-Monitoring	Ein einzelner HTTP-Request
Datenbank-Analyse	Eine Query / Transaktion
User-Research	Ein Nutzer / Eine Session
Log-Analyse	Eine Log-Zeile / Ein Event
Software-Testing	Ein Testfall / Ein Test-Run
Code-Qualität	Eine Datei / Funktion / Klasse
DevOps	Ein Deployment / Ein Build

0.1.5.2 Merkmal und Merkmalsausprägung

- **Merkmal:** Eigenschaft der Einheit (z. B. *Lieferzeit*)
- **Merkmalsausprägung:** konkret beobachteter Wert (z. B. *3,2 Tage*)

Mini-Beispiel:
Einheit = Lieferung · Merkmal = Lieferzeit · Ausprägung = 3,2 Tage

Informatik-Beispiel:

Einheit	Merkmal	Ausprägung
HTTP-Request	Response-Zeit	142 ms
HTTP-Request	Statuscode	200
HTTP-Request	HTTP-Methode	GET
Server	CPU-Auslastung	73,5 %
Bug-Ticket	Priorität	High
Code-Commit	Lines of Code	247

0.1.5.3 Grundgesamtheit und Stichprobe

- **Grundgesamtheit:** alle relevanten Einheiten, über die eine Aussage getroffen werden soll
- **Stichprobe:** tatsächlich beobachtete Teilmenge

Beispiel:
Grundgesamtheit = alle Lieferungen eines Jahres
Stichprobe = 120 zufällig ausgewählte Lieferungen

Informatik-Beispiele:

Grundgesamtheit	Stichprobe
Alle Requests eines Tages (10 Mio.)	10.000 zufällig geloggte Requests
Alle User einer Plattform (1 Mio.)	500 User für eine Usability-Studie
Alle möglichen Eingaben eines Programms	Ausgewählte Testfälle
Alle Commits eines Projekts	Die letzten 100 Commits
Alle potenziellen Kunden	A/B-Test mit 5% der Besucher

Repräsentativität:
Die Stichprobe soll die Grundgesamtheit möglichst gut abbilden.
Verzerrungen (Bias) führen zu falschen Schlussfolgerungen.

0.1.5.4 Bias in der Informatik Typische Verzerrungen bei IT-Daten:

Bias-Art	Beschreibung	Beispiel
Survivorship Bias	Nur „überlebende" Daten werden analysiert	Nur erfolgreiche Requests im Log
Selection Bias	Systematisch verzerrte Auswahl	Nur Power-User geben Feedback
Temporal Bias	Zeitliche Verzerrung	Daten nur von Werktagen, nicht Wochenenden
Sampling Bias	Nicht-repräsentative Stichprobe	Requests nur aus einem Rechenzentrum
Measurement Bias	Messung beeinflusst Ergebnis	Profiling verlangsamt die Anwendung

Beispiel Survivorship Bias:

Ein Monitoring-System loggt nur Requests, die erfolgreich beantwortet wurden. Requests mit Timeout werden nicht erfasst.

→ Die gemessene durchschnittliche Response-Zeit ist **systematisch zu niedrig**, weil die langsamsten Requests (die mit Timeout) fehlen.

0.1.6 Messniveaus (Skalentypen)

Das **Messniveau** bestimmt, welche statistischen Auswertungen sinnvoll und zulässig sind.

0.1.6.1 Nominalskala

- Kategorien ohne natürliche Ordnung
- Erlaubte Operationen: Gleichheit (=,)
- Beispiele: Farbe, Land, Produkttyp
- Auswertungen: Häufigkeiten, Modus, Balkendiagramme

Informatik-Beispiele:

- HTTP-Methode (GET, POST, PUT, DELETE, PATCH)
- Betriebssystem (Windows, macOS, Linux, Android, iOS)
- Programmiersprache (Python, Java, C++, JavaScript)
- Browser (Chrome, Firefox, Safari, Edge)
- Error-Typ (Timeout, NotFound, ServerError, AuthError)
- Serverstandort (Frankfurt, Dublin, Virginia, Singapore)

0.1.6.2 Ordinalskala

- Geordnete Kategorien ohne interpretierbare Abstände
- Erlaubte Operationen: Gleichheit + Ordnung (<, >, ,)
- Beispiele: Schulnoten, Zufriedenheit (1–5), Rankings
- Auswertungen: Median, Quartile, Rangmaße
  (Mittelwert nur mit Vorsicht!)

Informatik-Beispiele:

- Bug-Priorität (Critical, High, Medium, Low)
- User-Rating (1–5 Sterne)
- Log-Level (DEBUG, INFO, WARNING, ERROR, CRITICAL)
- Story Points / T-Shirt-Sizes (XS, S, M, L, XL)
- Service-Tier (Free, Basic, Pro, Enterprise)

- Reifegrad (Alpha, Beta, Stable, LTS)

Achtung: Der Abstand zwischen „3 Sterne" und „4 Sterne" ist nicht notwendigerweise gleich dem Abstand zwischen „4 Sterne" und „5 Sterne"!

0.1.6.3 Metrische Skala

- Numerisch mit interpretierbaren Abständen
- Erlaubte Operationen: Alle arithmetischen Operationen
- Beispiele: Zeit, Gewicht, Umsatz, Temperatur
- Auswertungen: Mittelwert, Varianz/Stdabw, Korrelation, Regression

Unterscheidung:

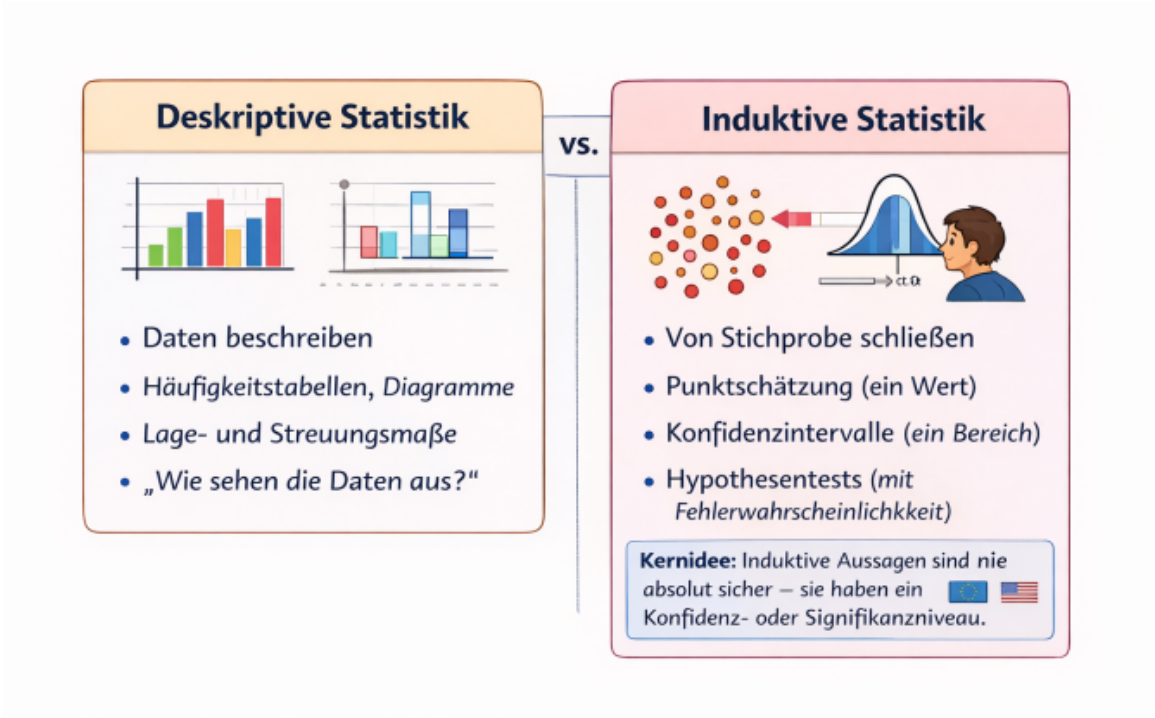
- **Intervallskala:** Kein natürlicher Nullpunkt (z. B. Temperatur in °C, Datum)
- **Verhältnisskala:** Natürlicher Nullpunkt (z. B. Zeit, Gewicht, Geld)

Informatik-Beispiele:

Merkmal	Skalentyp	Begründung
Response-Zeit (ms)	Verhältnis	0 ms = keine Zeit, 200 ms = 2× so lang wie 100 ms
Speicherverbrauch (GB)	Verhältnis	0 GB = kein Verbrauch
CPU-Auslastung (%)	Verhältnis	0% = keine Last
Dateigröße (Bytes)	Verhältnis	0 Bytes = leer
Lines of Code	Verhältnis	0 LOC = keine Zeilen
Timestamp (Unix-Zeit)	Intervall	Abstände interpretierbar, Nullpunkt willkürlich
Temperatur (°C)	Intervall	0°C „keine Temperatur"

Merksatz:
Nicht jede Zahl ist metrisch (z. B. Postleitzahlen).

0.1.7 Deskriptive vs. induktive Statistik



0.1.7.1 Deskriptive Statistik

Beschreibt den **vorliegenden Datensatz**:

- Häufigkeitstabellen und Diagramme
- Lage- und Streuungsmaße
- Leitfrage: *Wie sehen die Daten aus?*

0.1.7.2 Induktive Statistik (schließende Statistik)

Schließt von der **Stichprobe auf die Grundgesamtheit**:

- Punktschätzungen
- Konfidenzintervalle
- Hypothesentests

Kernidee: Induktive Aussagen sind nie absolut sicher – sie besitzen ein **Konfidenz- oder Signifikanzniveau**.

0.1.8 Typischer Statistik-Workflow

Der typische Ablauf einer statistischen Analyse folgt diesen Schritten:

1. Fragestellung präzisieren
2. Grundgesamtheit definieren
3. Stichprobenplan / Datenerhebung
4. Daten bereinigen (Fehler, Missing Values, Ausreißer)
5. Deskriptive Auswertung
6. Modellannahmen formulieren
7. Induktiv schließen
8. Interpretation im Kontext

0.1.9 Übergang: Von Statistik zur Wahrscheinlichkeitsrechnung

Viele statistische Modelle basieren auf **Zufallsexperimenten** und auf der Frage, wie viele mögliche Ergebnisse es gibt und wie wahrscheinlich diese sind.

Daher starten wir als mathematische Grundlage mit der **Wahrscheinlichkeitsrechnung**. Zuerst betrachten wir kombinatorische Zählverfahren (Permutation, Kombination, Variation), die später direkt in Wahrscheinlichkeiten übergehen.

0.2 Kombinatorische Hilfsmittel

0.2.1 Kombinatorik als Grundlage

Bevor wir mit Wahrscheinlichkeiten rechnen können, müssen wir oft wissen: Wie viele verschiedene Möglichkeiten gibt es?

IT-Beispiel: Ein Login-System akzeptiert 8-stellige Passwörter aus Kleinbuchstaben. Wie viele verschiedene Passwörter sind möglich?

→ Die Kombinatorik liefert die Werkzeuge für solche Zählprobleme.

0.2.2 Urnenmodell

Ausgangssituation ist eine Urne, in der sich n Kugeln befinden. Die Kugeln können sich durch Farbe oder Nummer unterscheiden.

Zentrale Fragestellung:
Auf wie viele verschiedene Arten lassen sich Kugeln aus der Urne ziehen?

Dabei sind zwei grundlegende Entscheidungen zu treffen:

- Ziehen **mit Zurücklegen** oder **ohne Zurücklegen**
- Beachtung oder Nichtbeachtung der **Reihenfolge**

0.2.2.1 Ziehen ohne Zurücklegen Beim Ziehen ohne Zurücklegen wird die gezogene Kugel nicht wieder in die Urne zurückgelegt. Die Anzahl der in der Urne verbleibenden Kugeln nimmt daher nach jedem Zug ab.

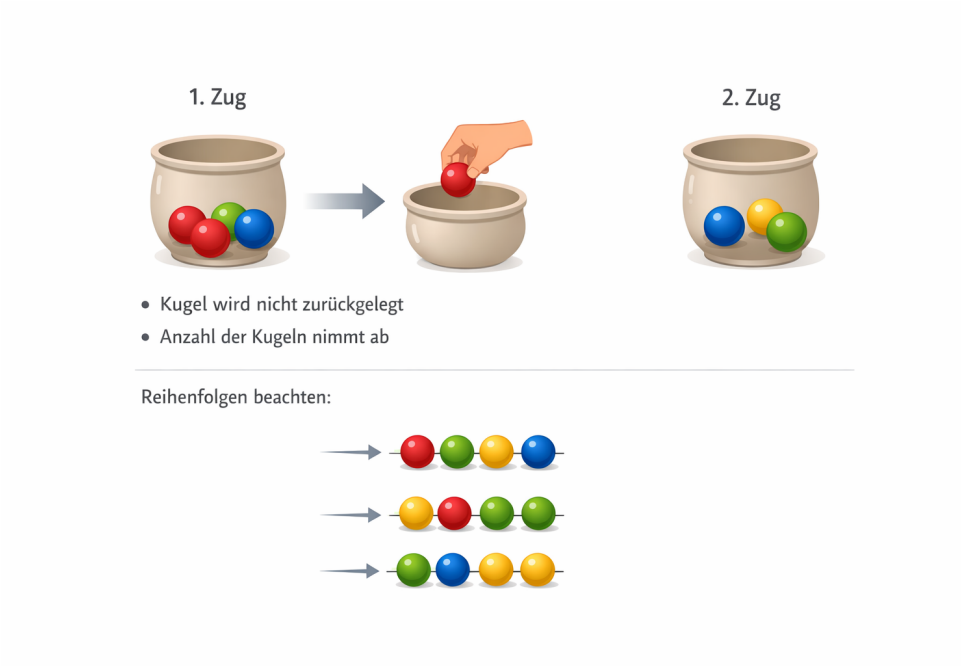


Abbildung 1: Ziehen ohne Zurücklegen. Nach jedem Zug wird eine Kugel entfernt, die Anzahl der verbleibenden Kugeln nimmt ab. Unterschiedliche Ziehungsreihenfolgen führen zu verschiedenen Ergebnissen.

Die möglichen Ergebnisse hängen davon ab, - welche Kugeln gezogen werden und - in welcher Reihenfolge die Ziehungen erfolgen.

0.2.2.2 Ziehen mit Zurücklegen Beim Ziehen mit Zurücklegen wird die gezogene Kugel nach jedem Zug wieder in die Urne zurückgelegt. Die Anzahl der Kugeln in der Urne bleibt somit konstant.

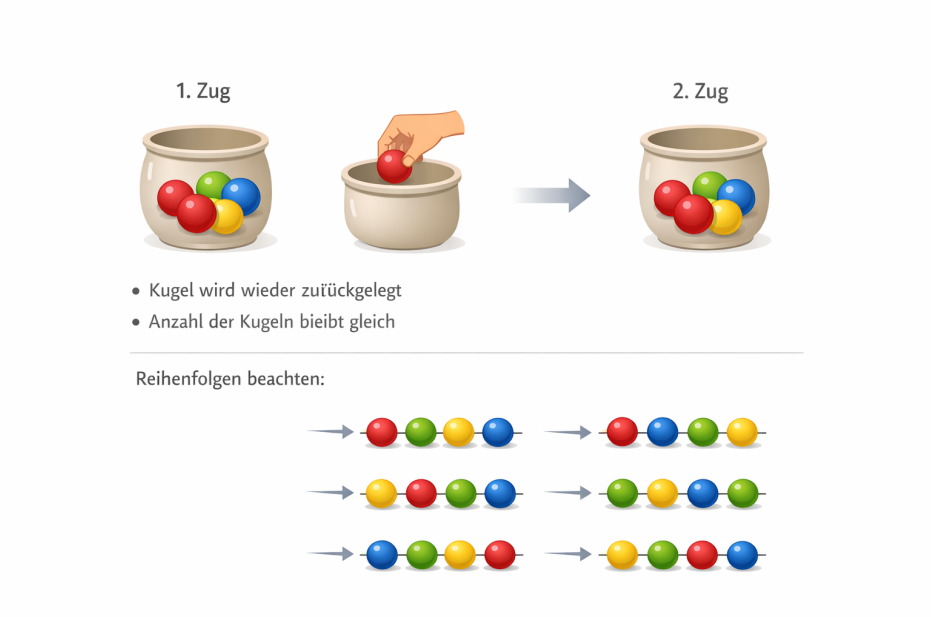


Abbildung 2: Ziehen mit Zurücklegen. Die gezogene Kugel wird nach jedem Zug zurückgelegt, sodass jede Ziehung unter identischen Bedingungen erfolgt.

Jede Ziehung erfolgt unter denselben Bedingungen wie die vorherige. Die einzelnen Ziehungen sind daher unabhängig voneinander.

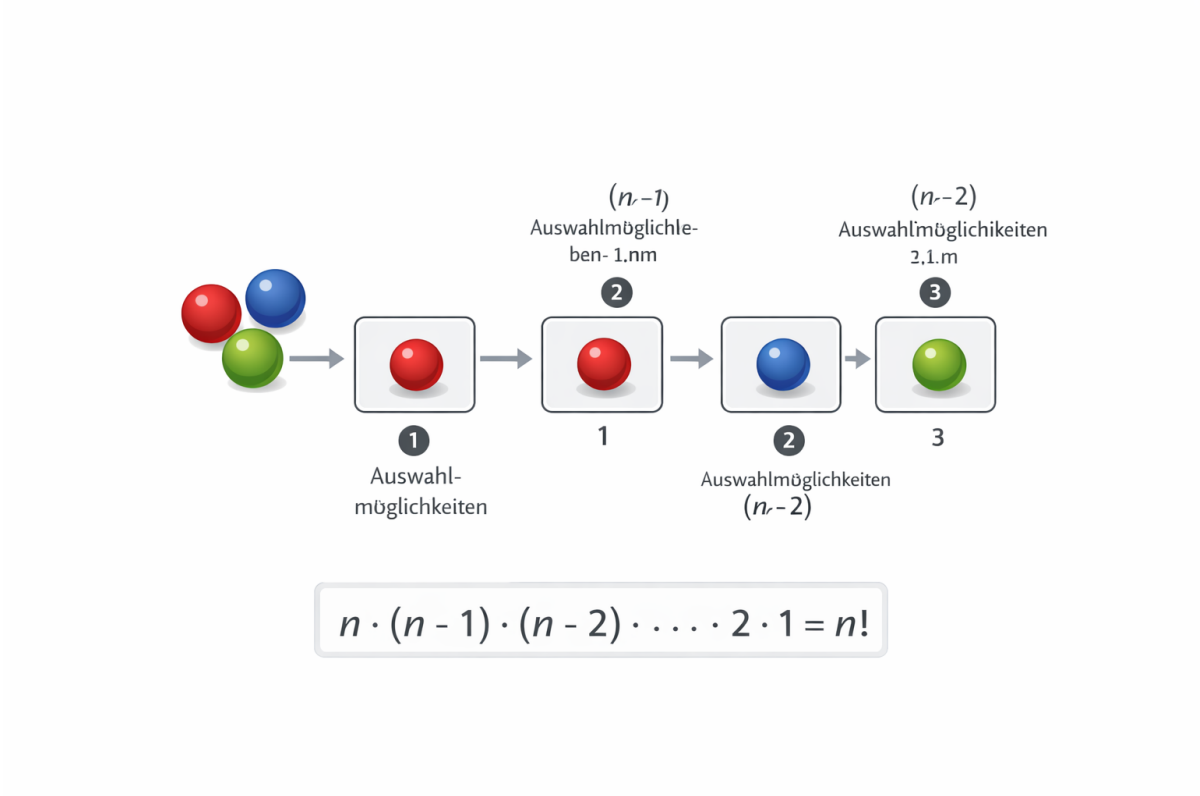
Die Kombination der beiden Entscheidungen - mit oder ohne Zurücklegen - mit oder ohne Beachtung der Reihenfolge

führt zu den grundlegenden kombinatorischen Begriffen: - **Permutation** - **Kombination** - **Variation**

Diese werden in den folgenden Abschnitten systematisch eingeführt.

0.2.3 Permutation

Beim Ziehen aller n Kugeln aus der Urne, wobei **jede Kugel genau einmal gezogen** wird und **die Reihenfolge der Ziehungen beachtet wird**, spricht man von einer **Permutation**.



Zur Veranschaulichung betrachten wir eine Urne mit drei verschiedenfarbigen Kugeln.

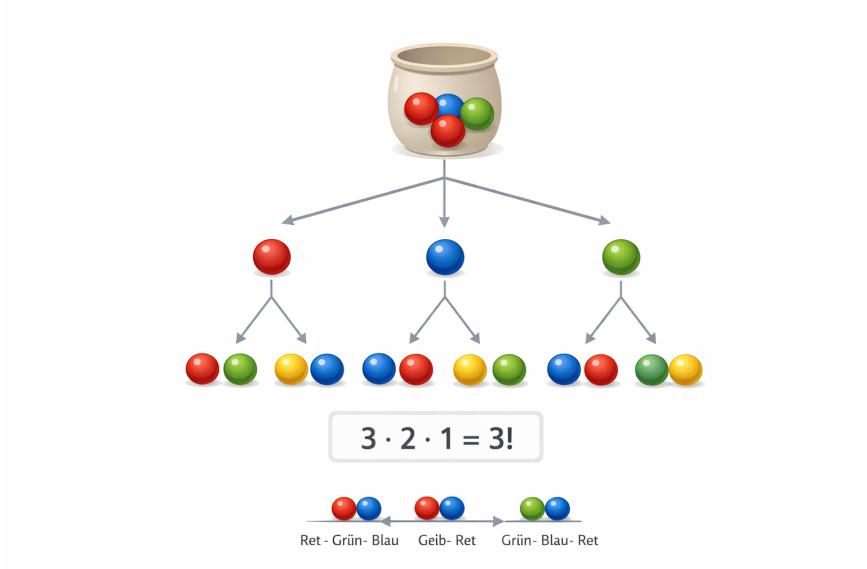


Abbildung 3: Systematische Aufzählung der möglichen Ziehungsfolgen bei drei Kugeln. Jede unterschiedliche Reihenfolge stellt eine eigene Permutation dar.

Im ersten Zug stehen drei Kugeln zur Auswahl, im zweiten Zug noch zwei, im dritten Zug nur noch eine Kugel.

Damit ergibt sich für die Anzahl der möglichen Ziehungsfolgen:

$$3 \cdot 2 \cdot 1 = 3!$$

Allgemein gilt für n verschiedene Kugeln:

$$P(n) = n \cdot (n-1) \cdot (n-2) \cdot ... \cdot 2 \cdot 1 = n!$$

Die Schreibweise $n!$ bezeichnet die **Fakultät** von n.

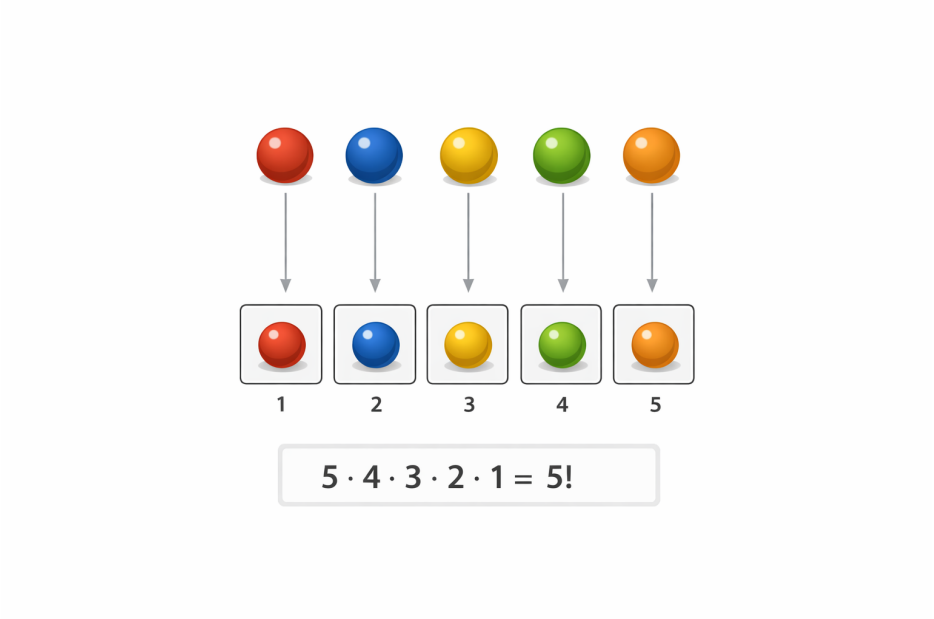


Abbildung 4: Permutation von 5 verschiedenen Kugeln. Jede Kugel wird genau einer Position zugeordnet. Die Anzahl aller möglichen Anordnungen beträgt $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5! = 120$.

IT-Anwendung:

Bei einer CI/CD-Pipeline müssen 4 Stages durchlaufen werden:
Build → Test → Deploy → Notify

Wie viele verschiedene Reihenfolgen sind theoretisch möglich?

$$P(4) = 4! = 24$$

(In der Praxis sind natürlich nicht alle Reihenfolgen sinnvoll – man würde nicht vor dem Testen deployen!)

0.2.3.1 Permutation mit gleichen Elementen Sind nicht alle Kugeln verschieden, so verringert sich die Anzahl der unterscheidbaren Permutationen.

Befinden sich unter den n Kugeln n_1 identische Kugeln, so ergibt sich:

$$P(n; n_1) = \frac{n!}{n_1!}$$

Sind mehrere Gruppen identischer Kugeln vorhanden, z. B. $n_1, n_2, \dots, n_k$, so gilt:

$$P(n; n_1, n_2, \dots, n_k) = \frac{n!}{n_1!\, n_2! \cdots n_k!}, \qquad n_1 + n_2 + \cdots + n_k = n$$

Abbildung 5 illustriert, dass durch identische Kugeln mehrere Ziehungsfolgen nicht unterscheidbar sind.

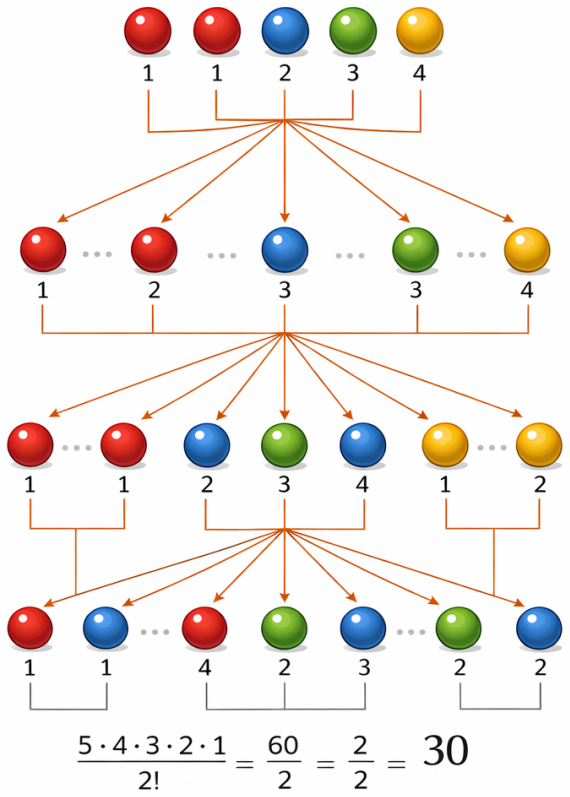


Abbildung 5: Permutationen mit gleichen Elementen. Durch identische Kugeln entstehen weniger unterscheidbare Anordnungen; die Gesamtzahl reduziert sich durch Division durch die Fakkultäten der gleichen Elemente.

IT-Anwendung:

Ein Logging-System erzeugt einen String aus 8 Zeichen: 3× ‘E’ (Error), 3× ‘W’ (Warning), 2× ‘I’ (Info).

Wie viele unterscheidbare Strings sind möglich?

$$P(8;3,3,2) = \frac{8!}{3! \cdot 3! \cdot 2!} = \frac{40\,320}{6 \cdot 6 \cdot 2} = 560$$

0.2.4 Kombinationen

Werden aus einer Urne mit n Kugeln k **Kugeln gezogen**, wobei - **nicht zurückgelegt** wird und - **die Reihenfolge keine Rolle spielt**,

so spricht man von einer **Kombination**.

Zur Herleitung betrachten wir eine Darstellung mit k Einsen („Kugel gezogen”) und $(n-k)$ Nullen („Kugel nicht gezogen”). Die Anzahl der möglichen Anordnungen dieser Symbole entspricht der Anzahl der Kombinationen.

Da Vertauschungen der Einsen untereinander sowie der Nullen untereinander keine neuen Kombinationen erzeugen, erhält man:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Die Schreibweise $\binom{n}{k}$ heißt **Binomialkoeffizient**.

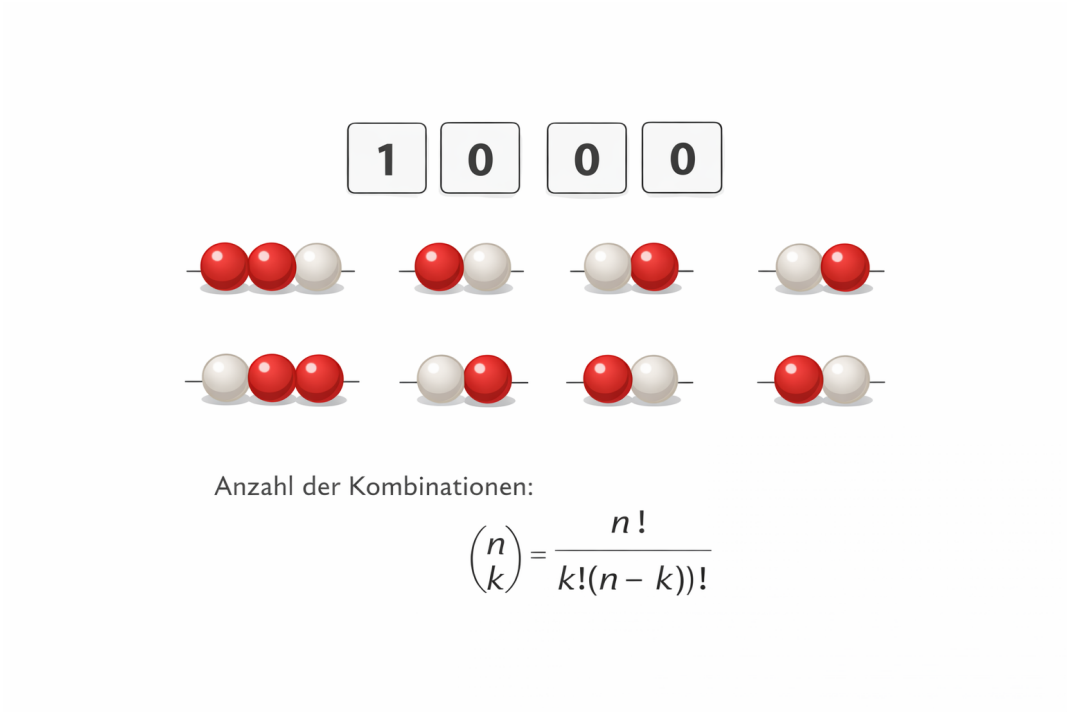


Abbildung 6: Kombinationen. Beim Ziehen ohne Zurücklegen und ohne Beachtung der Reihenfolge werden unterschiedliche Ziehungsfolgen als gleiches Ergebnis gezählt.

IT-Anwendung:

Für einen A/B-Test sollen aus 10 neuen Features genau 3 ausgewählt werden, um sie gemeinsam zu testen. Die Reihenfolge spielt keine Rolle.

Wie viele verschiedene Feature-Kombinationen gibt es?

$$\binom{10}{3} = \frac{10!}{3! \cdot 7!} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$$

0.2.5 Variationen

Werden aus n Kugeln k **Kugeln gezogen**, wobei - **die Reihenfolge berücksichtigt wird**, so spricht man von einer **Variation**.

0.2.5.1 Variation ohne Wiederholung Beim Ziehen **ohne Zurücklegen** ergibt sich:

$$V(n;k) = \frac{n!}{(n-k)!}$$

IT-Anwendung:

Ein System vergibt temporäre IDs aus einem Pool von 100 verfügbaren IDs. In einer Session werden 5 IDs nacheinander vergeben (ohne Wiederverwendung).

Wie viele verschiedene ID-Sequenzen sind möglich?

$$V(100;5) = \frac{100!}{95!} = 100 \cdot 99 \cdot 98 \cdot 97 \cdot 96 \approx 9{,}03 \cdot 10^9$$

0.2.5.2 Variation mit Wiederholung Beim Ziehen **mit Zurücklegen** stehen bei jeder Ziehung n Möglichkeiten zur Verfügung. Für k Ziehungen ergibt sich daher:

$$V_w(n;k) = n^k$$

IT-Anwendung:

Ein Passwort-System verlangt 8-stellige Passwörter aus Kleinbuchstaben (26 Zeichen). Wie viele verschiedene Passwörter sind möglich?

$$V_w(26;8) = 26^8 \approx 2{,}09 \cdot 10^{11}$$

Das sind etwa 209 Milliarden Möglichkeiten.

0.2.6 Tabellarische Zusammenfassung

Die wichtigsten kombinatorischen Begriffe lassen sich wie folgt zusammenfassen:

Auswahlart	ohne Wiederholung	mit Wiederholung
Kombination (Reihenfolge egal)	$\binom{n}{k}$	$\binom{n+k-1}{k}$
Variation (Reihenfolge wichtig)	$\frac{n!}{(n-k)!}$	n^k

Hinweis:
Die Formel für Kombination *mit* Wiederholung ($\binom{n+k-1}{k}$) wird in dieser Vorlesung nicht verwendet und ist daher nur der Vollständigkeit halber aufgeführt.


ohne Zurücklegen
mit Zurücklegen
ohne Zurücklegen
Reihenfolge wichtig
Reihenfolge egal
Reihenfolge wichtig