

DSAI Study Guide: Big Data & Stochastik (Statistics)

Komplette Prüfungszusammenfassung — Englisch (Original-Unterlagen) & Deutsch (Tafelbild + Heftnotizen erklärt)

Part 1: Big Data

1.1 Definition of Big Data

- **Big Data** consists of the words “big” and “data”.
- **Formal Definition:** Big Data is **high-volume, high-velocity, and/or high-variety** information assets that demand cost-effective, innovative forms of information processing that enable **enhanced insight, decision making, and process automation**.
→ Deutsch: Datenbestände mit großem Volumen (Volume), hoher Geschwindigkeit (Velocity) und/oder großer Vielfalt (Variety), die kosteneffiziente, innovative Formen der Informationsverarbeitung erfordern, um bessere Erkenntnisse, Entscheidungsfindungen und Prozessautomatisierung zu ermöglichen.

1.2 The 3 V's of Big Data (Volume, Velocity, Variety)

The 3 V's	Meaning (Bedeutung)	Lecture Points & Examples (aus den Unterlagen)
1. Volume (Datenmenge)	Massive amounts of data collected continuously from many sources.	• Sensor data (Sensordaten) • Quality testing data (Qualitätsprüfungsdaten) • E-mails from customers and suppliers (Kunden-/Lieferanten-E-Mails) • Evaluation data on portals (web shops, web pages / Bewertungen) • Content of social networks (Inhalte sozialer Netzwerke)
2. Velocity (Geschwindigkeit)	Speed at which data is generated and processed in real-time.	• Velocity of creation of data (Geschwindigkeit der Datenerzeugung) • Velocity of managing data (in real-time) (Echtzeit-Verarbeitung) • Data in Use: active data stored in databases under constant changes (aktive Daten in Datenbanken unter ständiger Änderung) • Data in Motion: data stored temporarily in a computer that will soon be read or updated (temporäre Daten im Arbeitsspeicher/Stream) • Data at Rest: inactive data stored in databases (data warehouse ...) or archives (inaktive Daten im Data Warehouse oder Archiv)
3. Variety (Datenvielfalt)	Different types, formats, structures, and sources of data.	• Heterogeneous data sources and data formats (text, videos, images, etc. / verschiedenartige Datenquellen & Formate) • Structured and non-structured data (strukturierte & unstrukturierte Daten) • Internal and external data (interne & externe Daten)

1.3 Big Data - Principles (Die 4 Grundprinzipien)

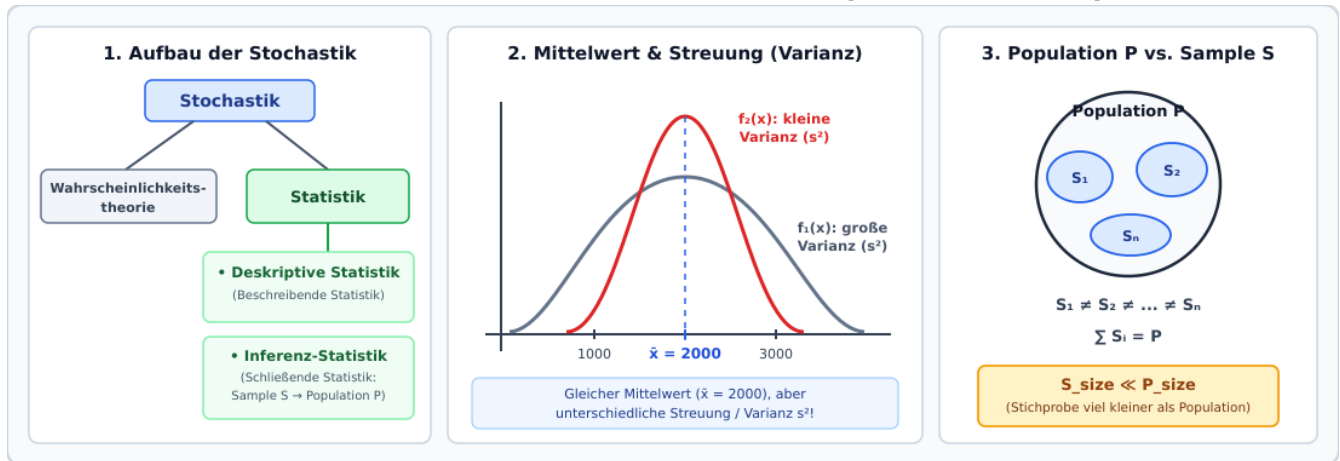
Principle	English Definition & Lecture Example	Deutsche Erklärung
1. “More”	Not only collecting samples, but collecting complete information .	Nicht nur kleine Stichproben, sondern möglichst alle Daten erfassen.
2. “Messy”	Accept that data quality will be low; amount of data will help to get correct results.	Geringere Datenqualität akzeptieren – die riesige Datenmenge gleicht Fehler aus.
3. “Correlation”	Analyze what, but not why (do not analyze the principle of yield management of airlines; find the best time for booking).	Das „ Was “ (Muster) analysieren, nicht zwingend das „ Warum “ (z. B. bester Flugbuchungs-Zeitpunkt).
4. “Datafication”	Everything is information (vibration of engine of a car) — collect it and use it (forecast that engine breaks down).	Alles messbar machen & nutzen (z. B. Motorvibrationen → Motorschaden vorhersagen).

1.4 Big Data - Technology (Technologien)

- **Parallel Computing:** Usage of hundreds or thousands of parallel running computers (Einsatz hunderter/tausender parallel arbeitender Computer).
- **NoSQL (“Not Only SQL”):** Flexible, non-relational database systems for massive and unstructured data.
- **MapReduce (Hadoop map):** Parallel and distributed algorithm (paralleler und verteilter Algorithmus zur Datenverarbeitung).
- **In-memory analytics:** Analyzing data directly in fast system RAM (Datenanalyse direkt im Arbeitsspeicher für Echtzeit-Analysen).

Part 2: Stochastik (Probability Theory & Statistics)

2.1 Tafelbild-Überblick: Stochastik, Mittelwert/Varianz & Population vs. Sample



Wichtiger Hinweis zum Heftaufschrieb (Blatt 1): Auf der Tafel steht im Kasten unten links bei der Glockenkurve **S_size << P_size** (die Stichprobengröße *Sample size* ist viel kleiner als die *Population size*). Im Heft wurde das versehentlich als „Size << Prize“ abgeschrieben. Für die Prüfung unbedingt **S_size << P_size** merken!

2.2 Statistiken / Metriken (Statistical Metrics)

Sei x_i ein einzelner **Datenpunkt** und n die Anzahl der Elemente (# Populationselemente bzw. Stichprobenumfang):

Metrik (Metric)	Formel & Symbole	Erklärung & Bedeutung für die Prüfung
Mittelwert (Mean)	$\bar{x} = (1/n) \cdot \sum x_i$ Sample: $\bar{x}$ („x-Quer“), Population: μ	Durchschnitt aller Werte (Summe aller Datenpunkte x_i geteilt durch die Anzahl n).
Varianz (Variance / Streuung)	$s^2 = (1/n) \cdot \sum (x_i - \bar{x})^2$ Sample: s^2 , Population: σ^2	Misst die Streuung der Datenpunkte um den Mittelwert $\bar{x}$. Durch das Quadrieren heben sich positive und negative Abweichungen nicht auf. • Breite, flache Kurve $f_1(x)$: große Varianz (große Streuung). • Schmale, hohe Kurve $f_2(x)$: kleine Varianz (kleine Streuung).
Standardabweichung (Std. Deviation)	$s = \sqrt{s^2}$ Population: $\sigma = \sqrt{\sigma^2}$	Quadratwurzel der Varianz. Hat wieder dieselbe Einheit wie die ursprünglichen Datenpunkte x_i (z. B. cm statt cm^2).
Median (Zentralwert)	Mittlerer Wert der aufsteigend sortierten Daten	Wird besonders für die Ausreißer-Erkennung (Outlier Detection) genutzt, da der Median unempfindlich gegenüber extremen Ausreißern ist.

2.3 Daten-Beschaffenheit (Qualitative vs. Quantitative Data) & Data Science Tabellen

Qualitative Daten (Categorical Data)	Quantitative Daten (Numerical Data)
• Kategorien / Eigenschaften („Label wie Kategorie“) • Beispiele: Haarfarbe (<i>blond, schwarz, ...</i>), Auto, Bewertung (<i>gut, mittel, schlecht</i>) • Man kann damit nicht direkt rechnen! • ⇒ Encoding: Kategorien werden in Zahlen übersetzt (z. B. <i>gut = 1, mittel = 2, schlecht = 3</i>), um damit zu rechnen.	• „Zahlen“ (Messwerte, Zählwerte, Körpergröße, Preise) • Man kann direkt damit rechnen („kann rechnen“ → Mittelwert, Varianz, Histogramm).

Statistical Terms in Data Science Language (Tabellen-Konvention):

- A database contains **Instances** (*Instanzen / Datensätze*) and their **Features** (*Merkmale / Eigenschaften*).
- Data are typically given in **2-dimensional Data tables** using the convention: **Instances are in rows (Zeilen), Features in columns (Spalten)**.

Part 3: Statistics - Visualization Histogram (Histogramm)

3.1 Definition & 4-Step Construction Recipe

- A **histogram** is a chart that plots the **distribution of a numerical variable's values** as a series of bars.
- Each bar covers a range of numeric values called a **bin (or class)** (*Klasse / Intervall*).
- A bar's **height** indicates the **frequency (count)** of data points with a value within the corresponding bin.

Step 1 (Range / Spannweite): Calculate the range of the data set: **Range = max – min**

Step 2 (Bins & Bin Width): Divide the range by the number of groups (#groups) you want, then **round up** (immer aufrunden!):

- **Bin (class) width** = $\lceil \text{Range} / \# \text{groups} \rceil$
- Other ways to get the optimal number of bins (groups) [with n = sample size]:
 - **Sturges' rule:** $\# \text{groups} = 1 + 3.322 \cdot \log_{10}(n)$
 - **Freedman-Diaconis rule**

Step 3 (Create groups): Use the class width (bin width) to create your groups (intervals) starting from min.

Step 4 (Frequency / Count): Find the frequency for each group (count) and draw the histogram bars.

3.2 Worked Example 1 (from Handout Big Data.docx.pdf)

- **Step 1 (Range):** $\text{max} - \text{min} = 65 - 12 = 53$
- **Step 2 (Bin Width):** $\# \text{groups} = 6$ (decided) $\Rightarrow \text{Range} / \text{Groups} = 53 / 6 = 8.83 \approx 9$ (rounded up) $\Rightarrow 9$ is called the bin (class) width.
- **Step 3 & 4:** Use the class width (here: 9) to create the 6 groups and find the frequency (count) for each group.

3.3 Worked Example 2 (from Handwritten Notes WA0004)

- **Gegeben:** Stichprobe mit $n = 27$ Werten, **min = 152, max = 191**.
- **Step 1 (Range):** $\text{Range} = 191 - 152 = 39$
- **Step 2 (Sturges' Rule & Bin Width):** $1 + 3,322 \cdot \log_{10}(27) = 5,75 \approx 6$ **Gruppen (Bins)**

Prüfungs-Hinweis zu Blatt WA0004: Im Heft steht vor der Formel „Bin Width = $1 + 3,322 \cdot \log(27) = 5,75$ “. Diese Formel (Sturges' Rule) berechnet jedoch die **Anzahl der Gruppen (5,75 → gerundet 6 Gruppen)**! Die **Klassenbreite (Bin Width)** ergibt sich danach aus $\text{Range} / \text{Gruppen} = 39 / 6 = 6,5 \rightarrow$ aufgerundet **7 Werte pro Intervall** (152 bis 158 umfasst genau 7 ganze Zahlen: 152, 153, 154, 155, 156, 157, 158).

Gruppe	Intervall (je 7 Zahlenwerte)	Count (Häufigkeit)	Bedeutung
1	152 - 158	1	1 Datenpunkt liegt zwischen 152 und 158
2	159 - 165	6	6 Datenpunkte liegen zwischen 159 und 165
3	166 - 172	5	5 Datenpunkte liegen zwischen 166 und 172
4	173 - 179	8	8 Datenpunkte (höchster Balken)
5	180 - 186	4	4 Datenpunkte liegen zwischen 180 und 186
6	187 - 193	3	3 Datenpunkte liegen zwischen 187 und 193
Summe (n)	Gesamtbereich: 152 - 193	27	$1 + 6 + 5 + 8 + 4 + 3 = 27$ Datenpunkte

