Statistik für Data Scientists

Vorlesung 2A: Glossar

Prof. Dr. Siegfried Handschuh

Universität St. Gallen

1. Lagekennzahlen

- **Mittelwert ($\bar{x}$):** Das arithmetische Mittel, die Summe aller Werte geteilt durch ihre Anzahl. Er ist sehr empfindlich gegenüber **Ausreissern** (extremen Werten) und kann dadurch verzerrt werden.
- **Median ($Q_{0.5}$):** Der Wert, der eine geordnete Datenreihe in zwei gleich grosse Hälften teilt. Er ist das 50%-Quantil und im Gegensatz zum Mittelwert **robust** gegenüber Ausreissern.
- **Getrimmter Mittelwert:** Ein Mittelwert, bei dem ein bestimmter Prozentsatz der kleinsten und grössten Werte vor der Berechnung entfernt wird, um die Robustheit zu erhöhen.
- **Modus:** Der am häufigsten vorkommende Wert in einer Datenreihe. Er ist besonders nützlich für kategorische Daten und kann auf **Multimodalität** (mehrere "Gipfel" in der Verteilung) hinweisen.
- **Quantile und Perzentile:** Werte, die eine Datenverteilung in gleiche Teile zerlegen. **Quartile** teilen die Daten in vier Teile (Q_1, Q_2, Q_3), während Perzentile sie in 100 Teile teilen. Quantile sind robust und bilden die Grundlage für den **Interquartilsabstand (IQR)**.

2. Streuungskennzahlen

- **Varianz (s^2) und Standardabweichung (SD, s):** Masse für die mittlere quadratische Abweichung vom Mittelwert. Die Varianz ist in quadratischen Einheiten, die **Standardabweichung** in der ursprünglichen Einheit der Daten und daher leichter interpretierbar. Beide sind **ausreisserempfindlich**.
- **Interquartilsabstand (IQR):** Die Differenz zwischen dem dritten und dem ersten Quartil ($Q_3 - Q_1$). Der **IQR** misst die Spannweite der mittleren 50% der Daten und ist **robust** gegenüber Ausreissern.
- **Mittlere absolute Abweichung (MAD):** Der Median der absoluten Abweichungen vom Median. Die **MAD** ist eine sehr **robuste** Alternative zur Standardabweichung und nützlich bei schiefen Verteilungen.

3. Ausreisser

- **Ausreisser:** Werte, die signifikant vom Rest der Daten abweichen. Sie können Messfehler, Eingabefehler oder seltene, aber valide Beobachtungen sein.
- **Tukey Fences:** Eine einfache und **robuste** Methode zur Identifizierung von Ausreisserkandidaten basierend auf dem **IQR**. Werte, die unter $Q_1 - 1.5 \cdot IQR$ oder über $Q_3 + 1.5 \cdot IQR$ liegen, gelten als Ausreisser.
- **Modifizierter Z-Score:** Eine **robuste** Alternative zum klassischen Z-Score, die den Median und die **MAD** anstelle des Mittelwerts und der Standardabweichung verwendet. Er eignet sich besser für schiefe Verteilungen.

4. Histogramm und KDE

- **Histogramm:** Ein Diagramm, das die Häufigkeitsverteilung von Daten durch rechteckige Säulen darstellt. Jede Säule repräsentiert die Häufigkeit (oder Dichte) von Werten in einem bestimmten Intervall (**Bin**).
- **Bin-Wahl:** Die Entscheidung für die Breite der Intervalle (Bins) in einem Histogramm. Eine falsche Wahl kann die Interpretation verzerren. Regeln wie die **Freedman-Diaconis-Regel (FD)** helfen, eine optimale Bin-Breite zu finden.
- **Kernel Density Estimation (KDE):** Eine Methode zur Schätzung der Wahrscheinlichkeitsdichtefunktion einer Zufallsvariable. Im Gegensatz zum Histogramm erzeugt die **KDE** eine glatte, kontinuierliche Kurve, die die Form der Verteilung ohne diskrete Bins zeigt.
- **Bandbreite (h):** Ein Schlüsselparameter bei der **KDE**, der die Glättung der Kurve steuert. Eine kleine Bandbreite führt zu einer zackigen, unruhigen Kurve, während eine grosse Bandbreite die Kurve zu stark glättet.

5. Box- und Violinplots

- **Boxplot:** Eine grafische Darstellung, die die Verteilung einer Variable durch fünf Kennzahlen zusammenfasst: Minimum, erstes Quartil (Q_1), Median, drittes Quartil (Q_3) und Maximum (oft dargestellt durch **Whisker**, die Ausreisser ausschliessen). Er bietet einen schnellen Überblick über Lage, Streuung und Ausreisser.
- **Violinplot:** Eine Kombination aus **Boxplot** und **KDE**. Die Form des Plots repräsentiert die Dichtefunktion (ähnlich der **KDE**), während die inneren Markierungen (z.B. ein kleiner Kasten oder Linien) die **Quartile** und den Median zeigen. Er ist nützlich, um die Form der Verteilung zu visualisieren, insbesondere bei Gruppenvergleichen.
- **Strip- und Swarmplots:** Plots, die die individuellen Rohdatenpunkte visualisieren. Sie können Boxplots oder Violinplots ergänzen, um die tatsächliche Verteilung der Punkte zu zeigen, die durch die statistischen Zusammenfassungen verborgen bleiben könnten.

6. ECDF und QQ-Plot

- **Empirical Cumulative Distribution Function (ECDF):** Eine binfreie und vollständige Darstellung der kumulativen Verteilung von Daten. Die **ECDF** zeigt für jeden Wert, welcher Anteil der Daten kleiner oder gleich diesem Wert ist. Sie ist sehr informativ, auch bei kleinen Datensätzen, und erlaubt das direkte Ablesen von **Quantilen**.
- **Quantile-Quantile Plot (QQ-Plot):** Ein Diagramm, das die **Quantile** einer Stichprobe gegen die **Quantile** einer theoretischen Verteilung (z. B. der Normalverteilung) aufträgt. Liegen die Punkte auf einer geraden Linie, stimmen die Verteilungen gut überein. Abweichungen von der Linie zeigen Abweichungen wie **Schiefe** oder **schwere Tails** (d. h. eine grosse Anzahl von extremen Werten).
- **Schwere Tails (Heavy Tails):** Eine Eigenschaft einer Verteilung, bei der die Wahrscheinlichkeit für extreme Werte höher ist als bei einer Normalverteilung. Im **QQ-Plot** zeigt sich dies durch eine Krümmung am oberen und unteren Ende.