**Kapitel 1: Einführung in Daten**

**Variable/Merkmal und Observation:** Eine Variable ist ein messbares Merkmal oder eine Eigenschaft, das/die in einer Studie gemessen oder beobachtet wird. Zum Beispiel könnten in einer Kundenumfrage die Variablen "Kaufverhalten", "Alter", "Einkommen" und "Zufriedenheit" sein. Jeder Kunde, der an der Umfrage teilnimmt, stellt eine Observation dar.

*Fallbeispiel 1: Einführung in Daten* Nehmen wir an, du führst eine Umfrage zur Bewertung von Filmen durch. Hierbei sind die Variablen die Merkmale, die du von den Teilnehmern erfassen möchtest, wie "Filmtitel", "Bewertung" und "Alter". Jede Person, die an der Umfrage teilnimmt, stellt eine Observation dar. Diese Personen sind die Grundgesamtheit, über die du Informationen sammeln möchtest.

**Kategoriale und metrische Variablen:** Kategoriale Variablen unterteilen Einheiten in verschiedene Kategorien oder Gruppen. Sie beschreiben Eigenschaften, die nicht quantifizierbar sind, wie Geschlecht, Farben oder Bildungsstufen. Im Gegensatz dazu haben metrische Variablen messbare Werte, sei es diskret (ganze Zahlen) oder stetig (beliebige Werte), wie Alter, Einkommen oder Temperatur.

*Fallbeispiel 2: Einführung in Daten* Nehmen wir an, du sammelst Daten über Autos, darunter "Farbe" (z. B. rot, blau, grün) als kategoriale Variable und "Geschwindigkeit" als metrische Variable. Farben sind Kategorien, die die Autos beschreiben, während Geschwindigkeiten numerische Werte sind, die gemessen werden können.

1. **Nominale Variablen:** Nominale Variablen sind Kategorien, die keine natürliche Reihenfolge oder Rangordnung haben. Das bedeutet, dass sie sich in keiner bestimmten Weise hierarchisch anordnen lassen. Ein Beispiel für nominale Variablen sind Farben, Geschlechter (männlich, weiblich, divers), Ländernamen oder Augenfarben. Diese Kategorien können nicht sinnvoll in eine Reihenfolge gebracht werden, da keine Wertung oder Rangfolge impliziert ist.
2. **Ordinale Variablen:** Ordinale Variablen hingegen haben eine bestimmte Reihenfolge oder Rangfolge zwischen den Kategorien, aber der Abstand zwischen den Kategorien ist nicht notwendigerweise gleichmäßig oder quantifizierbar. Ein Beispiel für ordinale Variablen ist Bildungsniveau (z.B. Grundschule, Mittelschule, Hochschule), da hier eine gewisse Rangfolge besteht, aber der tatsächliche Unterschied im Bildungsniveau zwischen den Kategorien nicht gleichmäßig gemessen werden kann.

**Abhängige und unabhängige Variablen:** Die abhängige Variable ist das Merkmal, das du untersuchst oder messt, um festzustellen, ob es durch Veränderungen in anderen Variablen beeinflusst wird. Die unabhängige Variable ist eine potenzielle Einflussgröße, von der vermutet wird, dass sie die abhängige Variable beeinflusst.

*Fallbeispiel 3: Einführung in Daten* Angenommen, du möchtest den Einfluss von Übung auf die Herzfrequenz untersuchen. Die Herzfrequenz ist die abhängige Variable, da sie von der Menge an körperlicher Übung abhängig ist. Die Menge an körperlicher Übung ist die unabhängige Variable, die möglicherweise die Herzfrequenz beeinflusst.

**Beobachtungsstudie und Experiment:** Eine Beobachtungsstudie beobachtet Ereignisse, ohne aktiv in den Ablauf einzugreifen. Sie zielt darauf ab, bestehende Zusammenhänge zu identifizieren. Ein Experiment hingegen ist ein geplantes Vorgehen, bei dem der Forscher bewusst gesteuerte Veränderungen vornimmt, um die Auswirkungen auf eine oder mehrere Variablen zu analysieren.

*Fallbeispiel 4: Einführung in Daten* Du möchtest den Einfluss von Sonnenlicht auf das Pflanzenwachstum untersuchen. Eine Beobachtungsstudie würde darin bestehen, Pflanzen in unterschiedlichem Sonnenlicht zu beobachten, um festzustellen, ob es einen Zusammenhang zwischen Lichtintensität und Wachstum gibt. Ein Experiment hingegen würde bedeuten, einige Pflanzen gezielt dem Sonnenlicht auszusetzen und andere im Schatten zu halten, um den Einfluss des Lichts auf das Wachstum zu analysieren.

**Grundgesamtheit und Stichprobe:** Die Grundgesamtheit ist die Gesamtheit der Elemente, über die Informationen gesammelt werden sollen. Da es oft nicht praktikabel ist, alle Elemente einer Grundgesamtheit zu untersuchen, wählen Forscher eine Stichprobe aus, um Rückschlüsse auf die Gesamtheit zu ziehen.

*Fallbeispiel 5: Einführung in Daten* Stell dir vor, du möchtest das Durchschnittseinkommen aller Angestellten in einem Unternehmen ermitteln. Die Grundgesamtheit wäre die Gesamtheit aller Mitarbeiter im Unternehmen. Da es jedoch zeitaufwändig und teuer sein könnte, Informationen von allen Mitarbeitern zu sammeln, würdest du stattdessen eine Stichprobe von beispielsweise 100 Mitarbeitern auswählen, um auf Grundlage dieser Stichprobe Rückschlüsse auf die gesamte Belegschaft zu ziehen.

**Stichprobenauswahl:** Es gibt verschiedene Methoden, um eine Stichprobe aus der Grundgesamtheit zu ziehen, darunter einfache Zufallsstichproben, stratifizierte Stichproben, Cluster-Stichproben und Mehrstufen-Stichproben.

*Fallbeispiel 6: Einführung in Daten* Du möchtest die Musikpräferenzen von Schülern in einer Schule ermitteln. Anstatt die gesamte Schülerpopulation zu befragen, könntest du eine einfache Zufallsstichprobe von 200 Schülern ziehen, um die Vielfalt der Präferenzen zu untersuchen. Du könntest auch eine stratifizierte Stichprobe verwenden, indem du die Schüler nach Klassenstufen aufteilst und aus jeder Stufe eine bestimmte Anzahl auswählst.

**Grundprinzipien statistischer Experimente:** Die Grundprinzipien für gute statistische Experimente sind Kontrolle, Randomisierung, Reproduzierbarkeit und Vergleich.

*Fallbeispiel 7: Einführung in Daten* Angenommen, du möchtest die Wirksamkeit eines neuen Medikaments zur Schmerzlinderung untersuchen. Du wählst zwei Gruppen von Patienten aus - eine erhält das Medikament und die andere ein Placebo. Beide Gruppen haben ähnliche Merkmale, um die Kontrolle zu gewährleisten. Du verabreichst die Substanzen zufällig und vergleichst dann die Schmerzreduktion zwischen den Gruppen, um die Wirksamkeit zu beurteilen. Durch die Randomisierung wird sichergestellt, dass die Ergebnisse nicht durch unkontrollierte Variablen verzerrt werden.

**Reduzierung menschlichen Einflusses:** Um den menschlichen Einfluss zu minimieren, werden oft Doppelblindstudien durchgeführt, bei denen weder die Teilnehmer noch die Forscher wissen, welcher Gruppe eine bestimmte Person zugewiesen wurde. Dadurch werden Vorurteile und Erwartungshaltungen reduziert, die die Ergebnisse beeinflussen könnten.

*Fallbeispiel 8: Einführung in Daten* Angenommen, du möchtest den Einfluss eines neuen Trainingsprogramms auf die Ausdauer testen. Um menschlichen Einfluss zu minimieren, könntest du eine Doppelblindstudie durchführen, bei der weder die Teilnehmer noch die Forscher wissen, welche Gruppe das intensive Training erhält und welche Gruppe das herkömmliche Training oder keine Änderung erhält. Dies hilft, Vorurteile und Erwartungshaltungen zu reduzieren, die die Ergebnisse beeinflussen könnten.

**Kapitel 2 Summarizing Data**

**2.1 Examining numerical Data**

**1. Scatterplots für gepaarte Daten:** Erklärung: Ein Scatterplot ist eine Grafik, die Punkte im Koordinatensystem darstellt, wobei jede Achse eine Variable repräsentiert. Wenn die Daten gepaart sind, werden die Punkte entsprechend den jeweiligen Paaren von Datenpunkten platziert, um die Beziehung zwischen den beiden Variablen zu visualisieren.

Beispiel: Eine Firma möchte untersuchen, ob es einen Zusammenhang zwischen der Anzahl der verbrachten Stunden für die Vorbereitung auf eine Prüfung und der tatsächlich erzielten Note gibt. Ein Scatterplot könnte erstellt werden, indem für jeden Schüler ein Punkt platziert wird, wobei die x-Achse die Stunden der Vorbereitung und die y-Achse die erzielte Note darstellen. Dies ermöglicht, zu sehen, ob höhere Vorbereitungsstunden mit besseren Noten korrelieren.

Ein Bild, das Text, Screenshot, Schrift, Reihe enthält.

Automatisch generierte Beschreibung

**2. Dot Plots und der Durchschnitt:** Erklärung: Ein Dot Plot ist eine Art von Diagramm, in dem jeder Datenpunkt durch einen Punkt auf einer horizontalen Achse dargestellt wird. Der Durchschnitt (Mittelwert) ist die Summe der Datenpunkte geteilt durch die Anzahl der Datenpunkte.

Beispiel: In einer Klasse haben Schüler einen Mathetest gemacht. Ein Dot Plot könnte erstellt werden, indem für jeden Schüler ein Punkt auf der Achse platziert wird, der seine Note repräsentiert. Der Durchschnitt der Noten könnte als Punkt auf dem Dot Plot dargestellt werden, um den durchschnittlichen Testwert visuell zu zeigen.

Ein Bild, das Text, Screenshot, Schrift, Reihe enthält.

Automatisch generierte Beschreibung

**3. Histogramme und Form:** Erklärung: Ein Histogramm ist eine grafische Darstellung der Verteilung von Daten, wobei die X-Achse in Intervalle unterteilt ist und die Y-Achse die Häufigkeit der Datenpunkte in jedem Intervall darstellt.

Beispiel: Ein Unternehmen möchte die Altersverteilung seiner Mitarbeiter analysieren. Ein Histogramm könnte erstellt werden, wobei die Altersgruppen in Intervallen auf der X-Achse abgebildet werden und die Höhe der Balken die Anzahl der Mitarbeiter in jedem Altersintervall zeigt.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

**4. Varianz und Standardabweichung:** Erklärung: Die Varianz misst die durchschnittliche Abweichung der Datenpunkte vom Durchschnitt. Die Standardabweichung ist die Quadratwurzel der Varianz und gibt an, wie stark die Datenpunkte um den Durchschnitt streuen.

Beispiel: Eine Gruppe von Schülern hat Mathematiktests gemacht, und ihre Noten wurden auf einer Skala von 0 bis 100 gemessen. Die Varianz würde anzeigen, wie weit die Noten im Durchschnitt von der durchschnittlichen Testnote abweichen. Die Standardabweichung würde die durchschnittliche Abweichung in derselben Einheit wie die ursprünglichen Noten messen.

**5. Boxplots, Quartile und Median:** Erklärung: Ein Boxplot ist eine grafische Darstellung, die Quartile (25. Perzentil, Median, 75. Perzentil) und mögliche Ausreißer anzeigt. Die Box repräsentiert den Bereich, in dem die meisten Datenpunkte liegen, während die "Whisker"-Linien die Streuung der Daten und potenzielle Ausreißer zeigen.

Beispiel: Eine Schule möchte die Verteilung der Testergebnisse für einen Mathematiktest analysieren. Ein Boxplot würde die Quartile (unteres Quartil, Median, oberes Quartil) zeigen und anzeigen, ob es Ausreißer gibt, die stark von der allgemeinen Verteilung abweichen.

**6. Robuste Statistik:** Erklärung: Robuste Statistik bezieht sich auf Methoden, die weniger empfindlich gegenüber Ausreißern oder abweichenden Werten sind. Diese Methoden bieten stabilere Analysen, insbesondere wenn die Daten nicht perfekt normal verteilt sind oder Ausreißer enthalten.

Beispiel: Ein Forscher untersucht das durchschnittliche Gehalt einer Gruppe von Angestellten. Da es einige leitende Angestellte gibt, die viel höhere Gehälter als der Rest der Gruppe verdienen, entscheidet er sich dafür, den Median anstelle des Durchschnitts zu verwenden, um eine stabilere Schätzung des durchschnittlichen Gehalts zu erhalten. Dies minimiert die Verzerrung durch die Ausreißer.

(a) About 30. (b) Since the distribution is

right skewed the mean is higher than the median.

(c) Q1: between 15 and 20, Q3: between 35 and 40,

IQR: about 20. (d) Values that are considered to be

unusually low or high lie more than 1.5 IQR away

from the quartiles. Upper fence: Q3 + 1.5   IQR =

37:5 + 1:5   20 = 67:5; Lower fence: Q1 - 1.5   IQR

= 17:5􀀀1:5 20 = 􀀀12:5; The lowest AQI recorded

is not lower than 5 and the highest AQI recorded is

not higher than 65, which are both within the fences.

Therefore none of the days in this sample would be

considered to have an unusually low or high AQI.

Für die Artbei kann RMarkdown gebraucht werden