Einführung in die deskriptive Statistik

Umgang mit Datenmengen

Olaf Schimmel

1 Daten, Datentypen und Skalen

In der Statistik untersucht man wohldefinierte Objekte auf bestimmte Eigenschaften, die auch **Merkmale** genannt werden. Die Merkmale nehmen gewisse Merkmalswerte an, die man **Daten** bzw. **Merkmalsausprägungen** nennt. Es muss stets eine eindeutige **Messvorschrift** zugrunde liegen, nach der die Merkmale bestimmt werden.
Dabei unterteilt man die Daten in drei Typen: **nominale, ordinale und kardinale Daten**.

Datentyp	Skala	Ausprägungsmerkmale	Vergleichsmöglichkeiten
Kategorie	Nominalskala	Namen, Eigenschaften ungeordnet	nur $x_1 = x_2$ oder $x_1 \neq x_2$ möglich
Rangdaten	Ordinalskala	Reihenfolge möglich, Werte geordnet Differenzen sinnlos, denn Abstände müssen nicht gleich sein	$x_1 < x_2$, $x_1 > x_2$ oder $x_1 = x_2$
Intervalldaten Messwerte	Kardinalskala	Ordnung und Differenzen möglich	$x_1 - x_2$

Beispiel 1: 14 Schüler wurden nach verschiedenen Kriterien befragt.

Lieblingsfach: Kategoriedaten, (Nominalskala)
Note für das UMG: Rangdaten (Ordinalskala) bzw.
Kardinalskala (diskret)
Körpergröße: Messwerte (Kardinalskala)

Definition: Ein Merkmal heißt **diskret**, wenn es nur abzählbar viele Ausprägungen gibt. Es heißt **stetig**, wenn es nicht diskret ist.

Merke: Ob ein Merkmal diskret oder stetig ist, hängt vom Messverfahren ab.

Beispiel 2: Die Masse eines Menschen wird auf verschiedene Arten bestimmt.

Variante 1:
Digitale Personenwaage auf 0,1 kg genau. Hierbei handelt es sich um ein diskretes Merkmal.

Variante 2:
Die Masse wird mit einer Federwaage bestimmt, wobei man die Ausdehnung der Feder misst. Jeder Wert ist möglich. Hierbei handelt es sich zunächst um ein stetiges Merkmal. Es wird zum diskreten Merkmal, wenn man von vornherein eine bestimmte Messgenauigkeit festlegt, zum Beispiel die Ausdehnung immer in vollen Millimetern bestimmt.

2 Darstellungsformen

Merke: Ziel der Darstellung von Daten ist es, möglichst einfach wichtige Informationen aus ihnen herauslesen zu können. Man unterscheidet hierbei **Tabellen** und **Diagramme**.

Diagramme sind oft anschaulicher als Tabellen. Außerdem eignen sich Diagramme auch besser zur Manipulation von Statistiken.

Beispiele: **Zeitreihe**
für chronologisch geordnete Merkmale

Säulen-, Balkendiagramm
für Kardinaldaten geringer Anzahl

Kreisdiagramm
Zur Darstellung relativer Häufigkeiten, die sich zu einem Ganzen bzw. 100% ergänzen.

Stabdiagramm
Zur Darstellung relativer Häufigkeiten auf einem Balken. 100% entsprechen dabei der gesamten Balkenlänge.

3 Häufigkeiten und Häufigkeitsdichte

Wenn man statistische Erhebungen durchführt, werden oft Fragebögen benutzt. Dabei werden dieselben Fragen einer Gruppe von Probanden gestellt. Oft möchte man aus dieser Stichprobe Aussagen gewinnen, die sich verallgemeinern lassen, also möglichst zuverlässig auf die Gesamtheit zutreffen.

Definitionen:

Umfang n der Stichprobe...
...ist die Anzahl der Probanden, die man befragt.

Absolute Häufigkeit $H_n(x)$ ...
...gibt an, wie oft die Merkmalsausprägung x in einer Stichprobe vom Umfang n vorliegt.

Relative Häufigkeit $h_n(x)$...
...gibt den Anteil am Umfang n der Stichprobe an, mit der die Merkmalsausprägung x auftrat.
Es gilt: $h_n(x) = \dfrac{H_n(x)}{n}$

Gegeben sei eine Stichprobe vom Umfang n und ein Merkmal mit den Merkmalsausprägungen x_1, x_2, ..., x_k. Die Zuordnung, die jeder Merkmalsausprägung x_i ihre relative Häufigkeit $h_n(x_i)$ zuordnet, heißt **Häufigkeitsdichte** der Stichprobe.

Die flächentreue grafische Darstellung der Häufigkeitsdichte nennt man **Histogramm**.

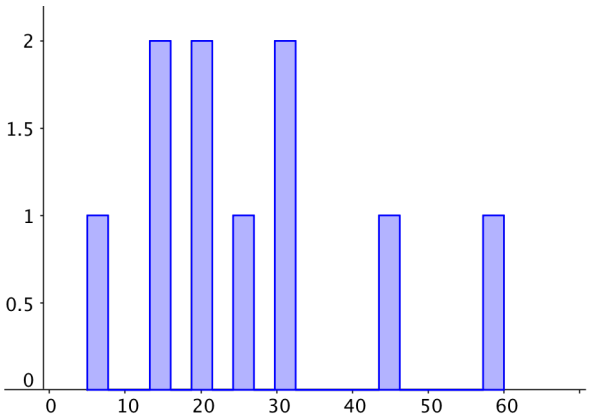
Beispiel:

10 Schüler werden nach der Länge des Schulweges in Minuten gefragt.

Urliste: 20, 15, 5, 15, 30, 30, 45, 20, 25, 60

geordnete Liste: 5, 15, 15, 20, 20, 25, 30, 30, 45, 60

Histogramm mit den absoluten Häufigkeiten:



Definition: Gegeben sei eine Stichprobe vom Umfang n und ein Merkmal mit den Merkmalsausprägungen x_1, x_2, ..., x_k mit $x_i \in R$. Die Zuordnung, die jeder Merkmalsausprägung x_i die relative Häufigkeit $h_n(x \leq x_i)$ zuordnet, heißt **empirische Verteilungsfunktion** $\hat{F}$ der Stichprobe.

Bemerkung: Im Gegensatz zur Häufigkeitsdichte summiert die empirische Verteilungsfunktion $\hat{F}$ die relativen Häufigkeiten aller Ausprägungen bis zum Wert x_i auf. Daher ist die empirische Verteilungsfunktion eine monoton wachsende Treppenkurve mit dem Minimalwert 0 und dem Maximalwert 1.

Beispiel: Zeiten für Schulwege

x_i	5	15	20	25	30	45	60
$h_n(x_i)$	0.1	0.2	0.2	0.1	0.2	0.1	0.1
$\hat{F}(x_i)$	0.1	0.3	0.5	0.6	0.8	0.9	1.0

4 Lageparameter

Zu jeder Häufigkeitsverteilung gibt es verschiedene Möglichkeiten Lageparameter anzugeben. Diese haben sehr unterschiedliche Aussagekraft und sind je nach Fragestellung mehr oder weniger gut geeignet, Aussagen zu treffen.

Definition: Gegeben sei eine Stichprobe vom Umfang n mit den Merkmalsausprägungen x_i mit $i \in \{1, 2, ..., n\}$.

Der **Modalwert (Modus)** x_{mod} ist der am häufigsten auftretende Wert. Treten mehrere Werte mit derselben Häufigkeit auf, so sind alle diese Werte Modalwerte (Modi).

Der **Median** x_{med} ist der mittlere Wert der geordneten vollständigen Liste aller Werte. Ist die Anzahl aller Werte gerade, dann ist der Median der arithmetische Mittelwert der beiden mittleren Werte in der geordneten Liste.

$$x_{med} = \begin{cases} x_{\frac{n+1}{2}} & \text{wenn n ungerade.} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{wenn n gerade.} \end{cases}$$

Das arithmetische Mittel: $\overline{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n)$ ist der **Schwerpunkt** der Daten.
andere Schreibweise:

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Definitionen: **andere Mittelwerte**

geometrisches Mittel: $x_{geom} = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n}$

harmonisches Mittel: $x_{har} = \dfrac{n}{\dfrac{1}{x_1} + \dfrac{1}{x_2} + ... + \dfrac{1}{x_n}}$

quadratisches Mittel: $x_{quad} = \sqrt{\dfrac{x_1^2 + x_2^2 + ... + x_n^2}{n}}$

Beispiel: Zeiten für Schulwege

$x_{mod} \in \{15, 20, 30\}$

$x_{med} = \dfrac{1}{2}(20 + 25) = 22,5$

$\overline{x} = \dfrac{1}{10} \cdot (5 + 2 \cdot 15 + 2 \cdot 20 + 25 + 2 \cdot 30 + 45 + 60) = 26,5$

$x_{geom} = \sqrt[10]{5 \cdot 15^2 \cdot 20^2 \cdot 25 \cdot 30^2 \cdot 45 \cdot 60} = 22,06$

$x_{har} = \dfrac{10}{\dfrac{1}{5} + \dfrac{2}{15} + \dfrac{2}{20} + \dfrac{1}{25} + \dfrac{2}{30} + \dfrac{1}{45} + \dfrac{1}{60}} = 17,27$

$x_{quad} = \sqrt{\dfrac{5^2 + 2 \cdot 15^2 + 2 \cdot 20^2 + 25^2 + 2 \cdot 30^2 + 45^2 + 60^2}{10}} = 30,53$

Bemerkungen: Jeder der Mittelwerte hat für bestimmte Sachverhalte seine Berechtigung. Der am häufigsten vorkommende Mittelwert ist das arithmetische Mittel. Das geometrische Mittel verwendet man beispielsweise zur Berechnung der durchschnittlichen prozentualen Veränderung pro Zeitraum aus den Veränderungen der einzelnen Zeiträume (Finanzmathematik). Das harmonische Mittel kann man zur Berechnung der Durchschnittsgeschwindigkeit aus den Durchschnittsgeschwindigkeiten auf gleichlangen Teilstrecken benutzen.

Satz: **Ungleichung der Mittelwerte**

Für positive Zahlen $x_1, x_2, ...x_n$ gilt:

$$x_{min} \leq x_{har} \leq x_{geom} \leq \overline{x} \leq x_{quad} \leq x_{max}$$

5 Streuungsparameter

Streuungsparameter geben Auskunft darüber, wie die Daten um die Lageparameter herum verteilt sind.

Definition: Der einfachste Streuungsparameter ist die **Spannweite S**. Es gilt: $S = x_{max} - x_{min}$

Gegeben sei eine Datenmenge mit n Zahlen $x_1; x_2; ...; x_n$. Die Zahlen $x_{\frac{1}{k}}; x_{\frac{2}{k}}; ...x_{\frac{k-1}{k}}$, die die geordnete Liste in k gleichgroße Teilmengen zerlegen, heißen **k-Quantile**. Bei gerader Anzahl der Listenelemente wird wie beim Median verfahren.

Die 4-Quantile $x_{0,25}, x_{0,5}, x_{0,75}$ nennt man **Quartile**.

Die Differenz $q_a = x_{0,75} - x_{0,25}$ heißt **Quartilsabstand**
Der Quartilsabstand bildet die Länge der Box bei Boxplots.

Beispiel: $1, 1, 1, 2, 3, 4, 4, 4, 4, 5, 6, 6, 6, 6, 6, 6, 7, 8, 75, 100$

5-Quantile: $x_{0,2} = 2,5$; $x_{0,4} = 4$; $x_{0,6} = 6$; $x_{0,8} = 6,5$.

Quartile: $x_{0,25} = 3,5$; $x_{0,5} = x_{med} = 5,5$; $x_{0,75} = 6$.

Quartilsabstand: $q_a = 6 - 3,5 = 2,5$

Der Vorteil von Quantilen ist ihre relativ leichte Berechnung, besonders bei großen Datenmengen. Nachteil ist, dass sie keine Aussage über weit abweichende Einzelwerte geben. Diese werden in keiner Weise berücksichtigt.

Definition: Gegeben sei eine Datenmenge mit n Zahlen $x_1; x_2; ...; x_n$ und dem Median$\overline{x}_{med}$. Dann heißt die Zahl

$$\tilde{x}_{abs} = \frac{1}{n} \sum_{i=1}^{n} |x_i - x_{med}|$$

mittlere absolute Abweichung

Beispiel: $\tilde{x}_{abs} = \frac{1}{20}(3 \cdot 4.5 + 3.5 + 2 \cdot 2.5 + 5 \cdot 1.5 + 7 \cdot .5 + 69.5 + 94.5)$

$\tilde{x}_{abs} = 9.85$

Der Nachteil dieser Streuung ist ihre recht aufwändige Berechnung. Außerdem möchte man gern größere Abweichungen stärker berücksichtigen und sich lieber am Schwerpunkt der Daten orientieren und nicht am Median. Deshalb wurde in der Statistik ein Streuungsparameter eingeführt, der dies berücksichtigt. Es handelt sich um die Varianz und die Standartabweichung.

Definition: Gegeben sei eine Datenmenge X mit n Zahlen $x_1; x_2; ...; x_n$ und dem arithmetischen Mittel $\overline{x}$. Dann heißt die Zahl

$$V(X) = var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Varianz der Datenmenge X.

Die Wurzel aus der Varianz

$$\sigma(X) = \sqrt{V(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

heißt **Standartabweichung** der Datenreihe X.

Beispiel: $V(x) = \frac{1}{20}(3 \cdot 4.5^2 + 3.5^2 + 2 \cdot 2.5^2 + 5 \cdot 1.5^2 + 7 \cdot 0.5^2 + 69.5^2 + 94.5^2)$

$V(x) = 692.95$

$\sigma(X) = 26.32$