Beschreibung univariater Verteilungen

Inhaltsverzeichnis

Beschreibung univariater Verteilungen 2
Lernhinweise 2
Einführung 2
Theorie (1-4) 4
1. Verteilungsformen 4
2. Masse der zentralen Tendenz (Mittelwerte) 5
Einleitung .....
Der arithmetische Mittelwert [mean] .....
Der Modalwert [mode] .....
Der Median [median] .....
3. Masse zur Beschreibung der Variabilität (Streuung) 12
Einleitung .....
Variabilität V [range] .....
Exkurs: Perzentil-, Dezil-, Quartilwerte .....
Perzentil- bzw. Zentilwerte .....
Quartilwerte und Darstellung mit Boxplots .....
Darstellung des Boxplots mit SPSS .....
Interquartilweite QW und Quartilabweichung Q .....
Varianz s^2 und die Standardabweichung s .....
4. Zusammenfassung zum Lernschritt 20
Fallbeispiel 21
Lernkontrolle 22

Beschreibung univariater Verteilungen

Lernhinweise

Dieser Lernschritt informiert Sie darüber, wie univariate Datensätze bzw. die Verteilung der Daten mit Hilfe verschiedener sogenannter Kennwerte charakterisiert werden können. Es wird auf Masse der zentralen Tendenz (Lagemasse) und auf Masse zur Beschreibung der Variabilität (Streuungsmasse) sowie verschiedene Verteilformen univariater Datensätze eingegangen.

Benötigte Vorkenntnisse

- Die verschiedenen Massstabstypen (Skalenniveaus)
- Die tabellarische Darstellung der Ausprägungen eines Merkmals (univariate Verteilungen)
- Die grafische Darstellung der Ausprägungen eines Merkmals (univariate Verteilungen)

Lernziele

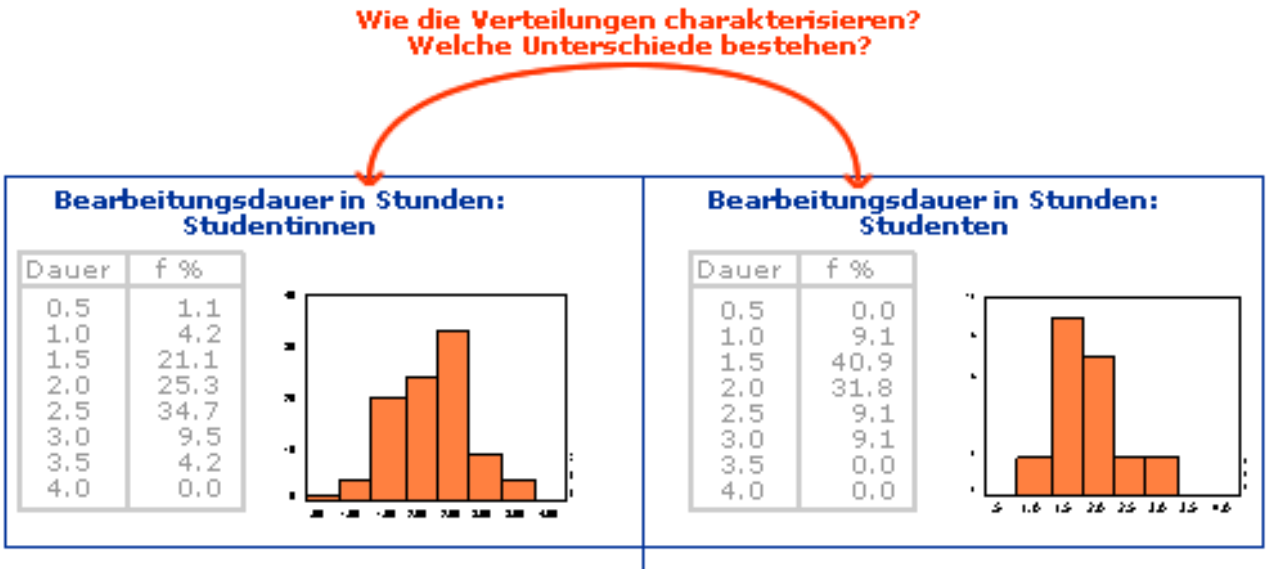
- Sie kennen drei Kennwerte zur Charakterisierung der zentralen Tendenz und die wichtigsten Kennwerte zur Beschreibung der Streuung einer Häufigkeitsverteilung.
- Sie wissen, welches Skalenniveau für die Bestimmung der verschiedenen Kennwerte vorausgesetzt wird.
- Sie können für eine Häufigkeitsverteilung die geeigneten Kennwerte bestimmen.

Geschätzte Bearbeitungszeit
Dieser Lernschritt kann in 60 bis 90 Minuten durchgearbeitet werden.

Hinweise zur Bearbeitung
Beim Anklicken des "next"-Buttons (oben rechts und unten links auf der Seite) werden Sie nach der in der oben dargestellten Rubriken-Reihenfolge durch den Lernschritt geführt: (1) Lernhinweise, (2) Einführung, (3) Theorie, (4) Fallbeispiel, (5) Lernkontrolle.
Ein Fallbeispiel und eine Lernkontrolle ist noch nicht verfügbar.

Einführung

Das Problem, Häufigkeitsverteilungen miteinander zu vergleichen...
Empirisch erhobene Daten können anhand von Häufigkeitsverteilungen tabellarisch zusammengefasst oder grafisch dargestellt werden (s. Abbildung). Oft will man Datensätze vergleichen. Es interessiert beispielsweise, ob Studentinnen für die Bearbeitung eines Lernschritts durchschnittlich mehr Zeit aufwenden als Studenten. Mit tabellarisch oder grafisch dargestellten Häufigkeitsverteilungen sind eindeutige Aussagen zu derartigen Fragen nur schwer oder kaum möglich.



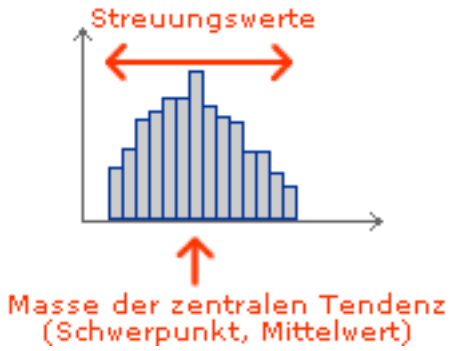
Wie können Häufigkeitsverteilungen beschrieben werden?
Die Form der Verteilung kann beschrieben werden, indem beispielsweise auf die Symmetrie oder die Steilheit eingegangen wird. Ein Datensatz kann damit aber nur ungenau charakterisiert werden. Es stellt sich deshalb die Frage:

- Wie kann ein Datensatz auf das Wesentliche reduziert werden?
- Wie können solche Häufigkeitsverteilungen möglichst knapp charakterisiert werden?
- Wie können Datensätze verdichtet werden, dass Vergleiche zwischen verschiedenen Datensätzen möglich sind?

Die Lösung: Die Einführung von Kennwerten zur Beschreibung univariater Verteilungen.
Das Bestreben der Deskriptiven Statistik, Beobachtungsdaten knapp zu charakterisieren, hat zur Entwicklung einer Anzahl von Kennwerten geführt. Diese sollen die Daten möglichst gut repräsentieren und zur Beschreibung der Verteilung verwendet werden können. Dabei werden viele Einzelinformationen zu wenigen, aber aussagekräftigen Grössen verdichtet.

Nachteil von Kennwerten
Da durch die Verdichtung der Daten zu Kennwerten notwendigerweise Informationen verloren gehen, muss dennoch häufig auf den Datensatz oder die Häufigkeitsverteilung zurückgegriffen werden. Der Nachteil von Informationsverlust wird jedoch oft in Kauf genommen, weil Kennwerte leichter vergleichbar und mitteilbar sind als Häufigkeitsverteilungen.

Zwei Gruppen von Kennwerten zur Beschreibung von Häufigkeitsverteilungen
Häufigkeitsverteilungen können auf zwei verschiedene Arten charakterisiert werden:



Masse der zentralen Tendenz (Mittelwerte)
Auf unterschiedliche Weise kann der „Schwerpunkt“ einer Verteilung beschrieben werden. Dies ermitteln die sogenannten Masse zur zentralen Tendenz, auch Lagemasse genannt.

Streuungswerte
Häufig interessiert auch die Frage, wie dicht die einzelnen Daten beieinander liegen, oder wie stark sie streuen. Man spricht von einer starken Streuung, wenn die Daten weit auseinander liegen.

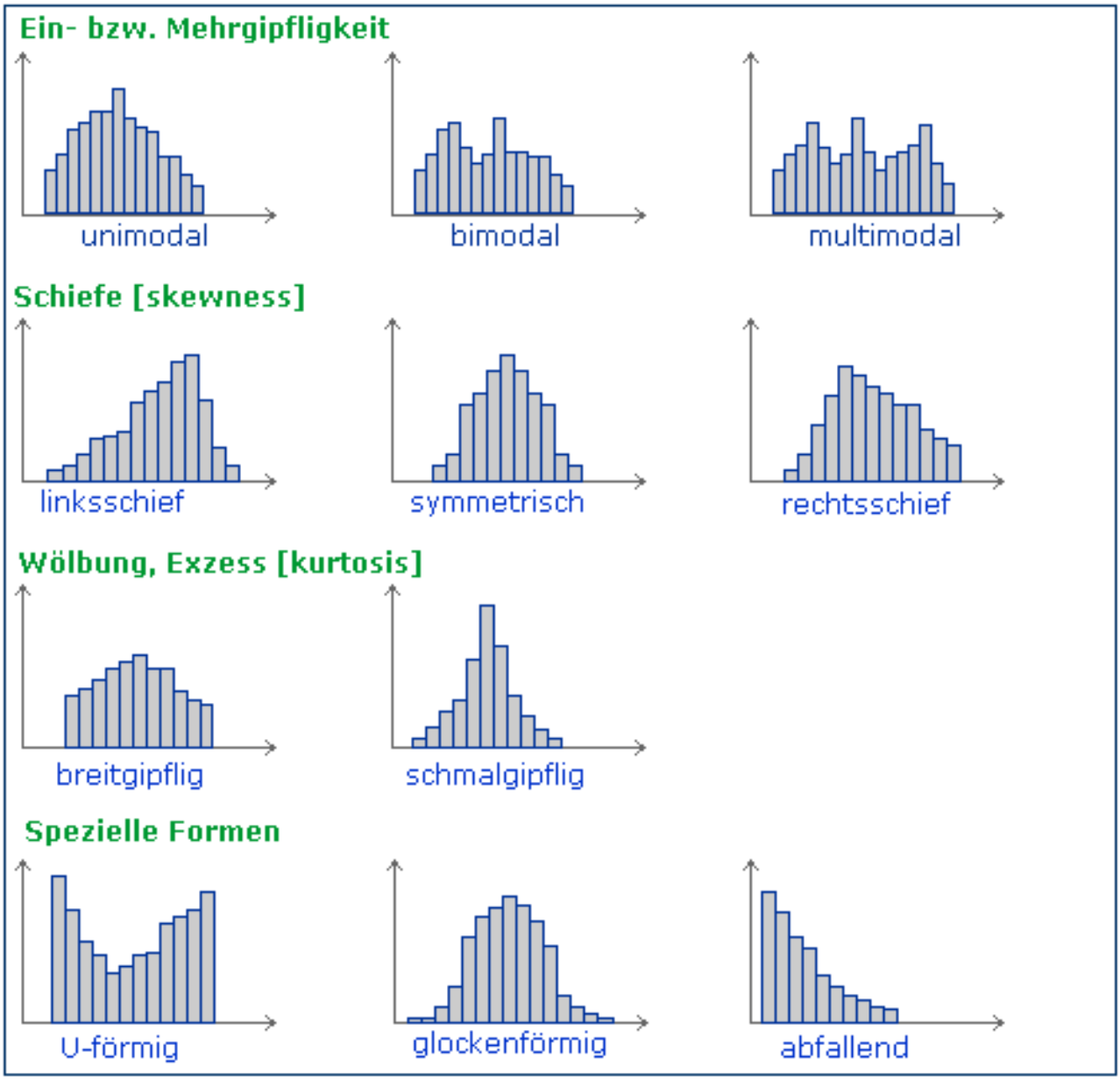
Theorie (1-4)

Inhaltsübersicht:

- 1. Verteilungsformen
- 2. Masse der zentralen Tendenz (Mittelwerte)
- 3. Masse zur Beschreibung der Variabilität (Streuung)
- 4. Zusammenfassung zum Lernschritt

1. Verteilungsformen

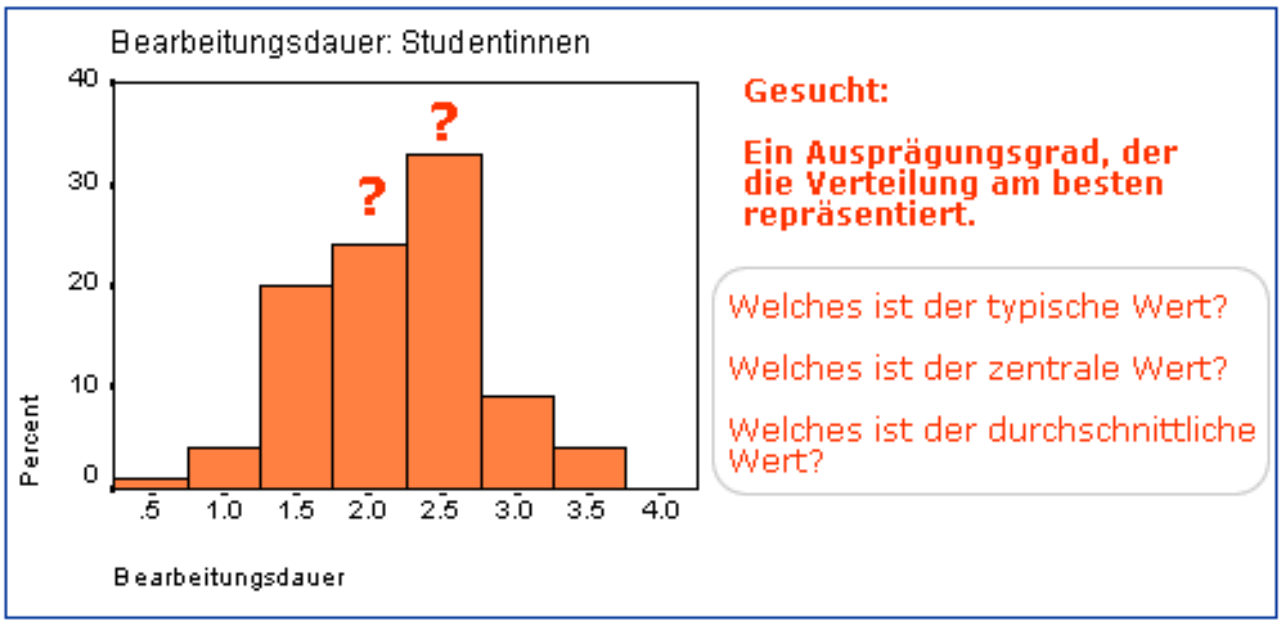
Verschiedene Formen einer Verteilung
Eine Häufigkeitsverteilung kann mit einem Kennwert (Masse der zentralen Tendenz, Streumasse) charakterisiert werden, oder man beschreibt die Häufigkeitsverteilung verbal.
Es lassen sich einige typische Verteilungsformen unterscheiden:



2. Masse der zentralen Tendenz (Mittelwerte)

Die Masse der zentralen Tendenz bezeichnen in unterschiedlicher Weise den „Schwerpunkt“ einer Verteilung. Im Englischen werden diese Kennwerte mit „measures of central tendency“ oder „representative values“ bezeichnet, was darauf hinweist, dass sie den typischen, den zentralen oder durchschnittlichen Ausprägungsgrad einer Verteilung repräsentieren.

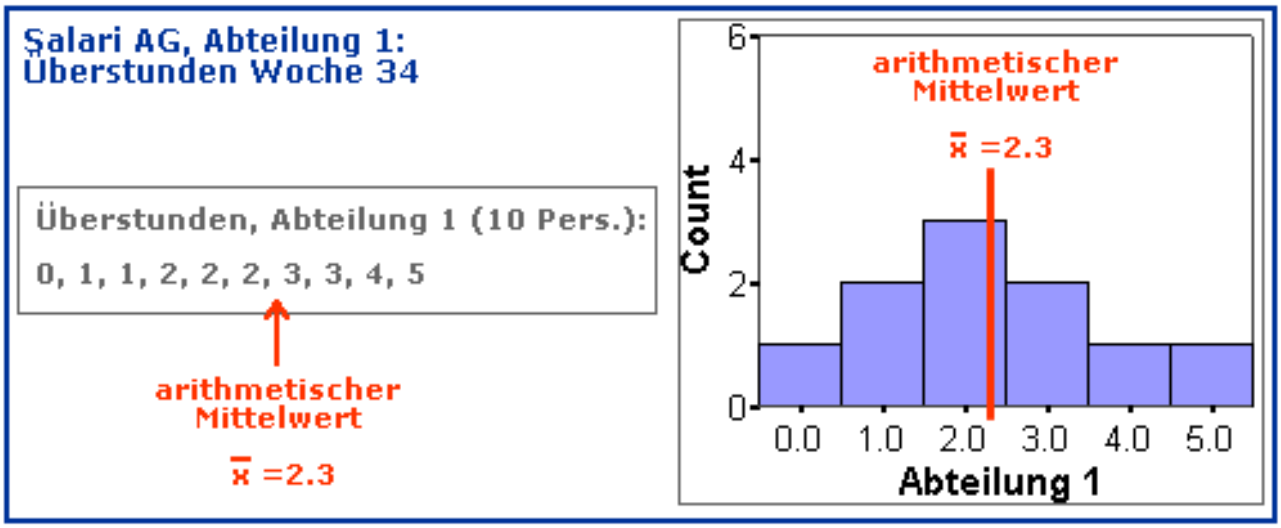
Wie lässt sich beispielsweise die Bearbeitungsdauer von 95 Studentinnen für einen bestimmten webbasierten Lernschritt mit einem einzigen Kennwert beschreiben?



Verschiedene Masse der zentralen Tendenz

Die beschreibende Statistik in den Sozialwissenschaften kennt eine Anzahl von Kennwerten, die auf unterschiedlichen Vorstellungen von der „Mitte“ oder dem „Schwerpunkt“ basieren:

- Der arithmetische Mittelwert
- Der Modus
- Der Median



Andere Bezeichnungen

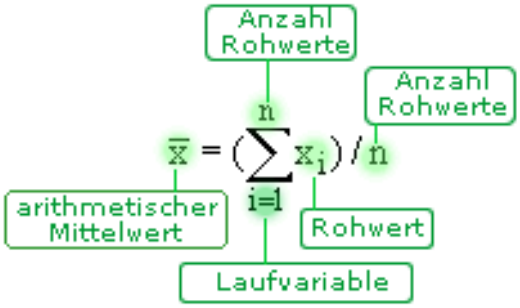
Umgangssprachlich: Mittelwert, Durchschnitt

Charakterisierung

Die Summe aller Ausprägungswerte wird durch die Anzahl aller Beobachtungen dividiert.

Bestimmung des arithmetischen Mittelwertes aus den **Rohwerten**:

$$\bar{x} = \left(\sum_{i=1}^{n} x_i\right) / n$$



Voraussetzungen

Die Addition von Ausprägungsgraden zur Bestimmung des arithmetischen Mittelwertes ist zulässig, wenn das Merkmal mindestens intervall-skaliert ist.

Eignung

Der arithmetische Mittelwert ist die am häufigsten verwendete Kennzahl. Da der arithmetische Mittelwert von allen Ausprägungswerten abhängig ist, kann bei schiefen Verteilungen oder durch einzelne Extremwerte, die für die Verteilung nicht repräsentativ sind, bei kleinen Stichproben ein verzerrter Eindruck entstehen.
Deshalb eignet er sich für eingipflige und symmetrische Häufigkeitsverteilungen sowie für Verteilungen ohne klar erkennbare Konzentration auf einen Ausprägungsgrad. Er eignet sich nicht bei mehrgipfligen und asymmetrischen Verteilungen.

Hinweise und zusätzliche Erklärungen

Unterschiedliche Bezeichnungen des arithmetischen Mittelwertes.

Wenn die Daten aus einer Stichprobenerhebung stammen, wird der arithmetische Mittelwert mit

$\bar{x}$

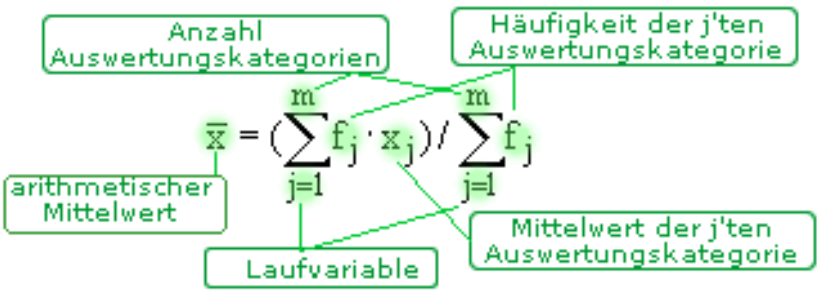
(„x-quer“) bezeichnet. Man spricht dabei von einem Verteilungskennwert.
Im Zusammenhang mit Populationen oder theoretischen Verteilungen wird der arithmetische Mittelwert mit

μ

(„my“) bezeichnet. Wir sprechen von einem Verteilungsparameter.

Berechnung des arithmetischen Mittelwertes aus einer primären oder sekundären Häufigkeitsverteilung

$$\bar{x} = \left(\sum_{j=1}^{m} f_j \cdot x_j\right) / \sum_{j=1}^{m} f_j$$

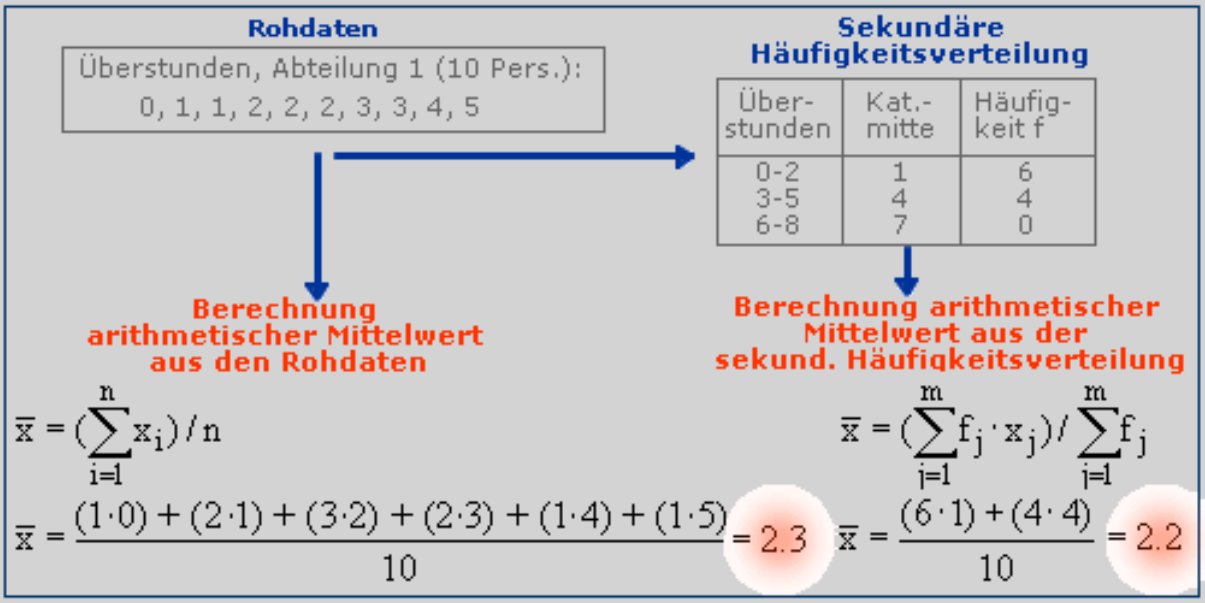


Zu beachten bei Berechnung des arithmetischen Mittelwertes aus einer sekundären Häufigkeitsverteilung

Der aus den Rohwerten oder einer primären Häufigkeitsverteilung berechnete arithmetische Mittelwert kann vom arithmetischen Mittelwert einer sekundären Häufigkeitsverteilung, bei der die Ausprägungsgrade kategorisiert wurden, abweichen, da durch die Kategorisierung Informationen verloren gehen. Es soll darum angegeben werden, auf welchen Datensatz sich der arithmetische Mittelwert bezieht. Nach Möglichkeit soll sich der arithmetische Mittelwert auf die Rohdaten oder die primäre Häufigkeitsverteilung beziehen.

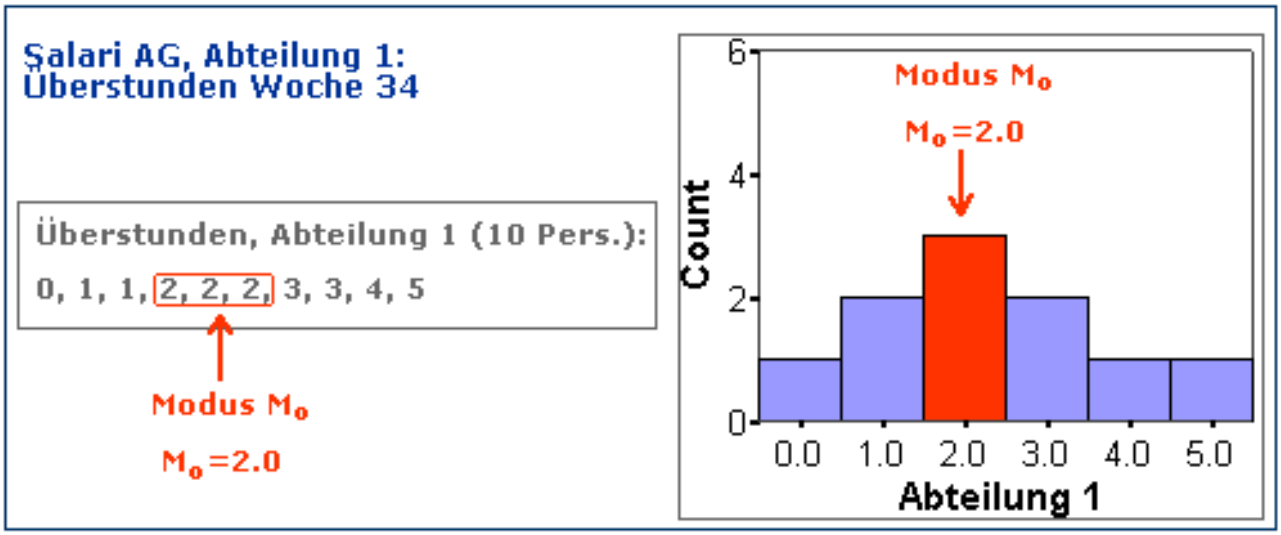
Beispiel: Ermittlung des arithmetischen Mittelwertes aus den Rohwerten und aus einer sekundären Häufigkeitsverteilung

Der arithmetische Mittelwert für die Überstunden in der Abteilung 1 der "Salari AG" wurden auf zwei Arten berechnet:
Einmal wurde aus den Rohdaten der arithmetische Mittelwert direkt ermittelt (linker Teil der Abbildung).
Beim zweiten Fall wurden die Rohdaten zuerst kategorisiert und danach erst der arithmetische Mittelwert berechnet (rechter Teil der Abbildung). Wie schon zuvor erwähnt, können die beiden arithmetischen Mittelwerte voneinander abweichen, da durch die Kategorisierung Informationen verlorengehen. Im Beispiel besteht dieser Unterschied: Für den arithmetischen Mittelwert erhält man je nach Art der Berechnung "2.3" oder "2.2".



Bemerkung zum Kategorienmittelwert xj

Die Kategorienmitte wird als Repräsentant für die Merkmalsausprägung angesehen, d.h., dass beispielsweise für alle Ausprägungen der Kategorie "0-2" Überstunden die Kategorienmitte "1" für die Berechnung angewendet wird.

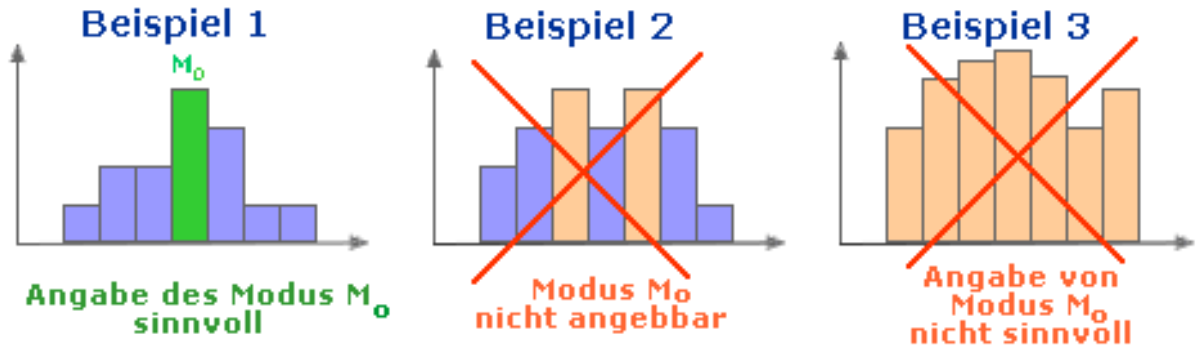


Andere Bezeichnungen
Modus M_o, häufigster Wert, dichtester Wert

Charakterisierung
Der Modus Mo ist derjenige Ausprägungsgrad der Merkmalsdimension, der am häufigsten beobachtet wurde.

Voraussetzungen
Der Modalwert M_o kann prinzipiell bei jedem Skalenniveau angewendet werden.

Eignung
Die drei nachfolgenden Beispiele zeigen, dass der Modus M_o nicht immer angebbar ist oder die Angabe als nicht sinnvoll erachtet wird.



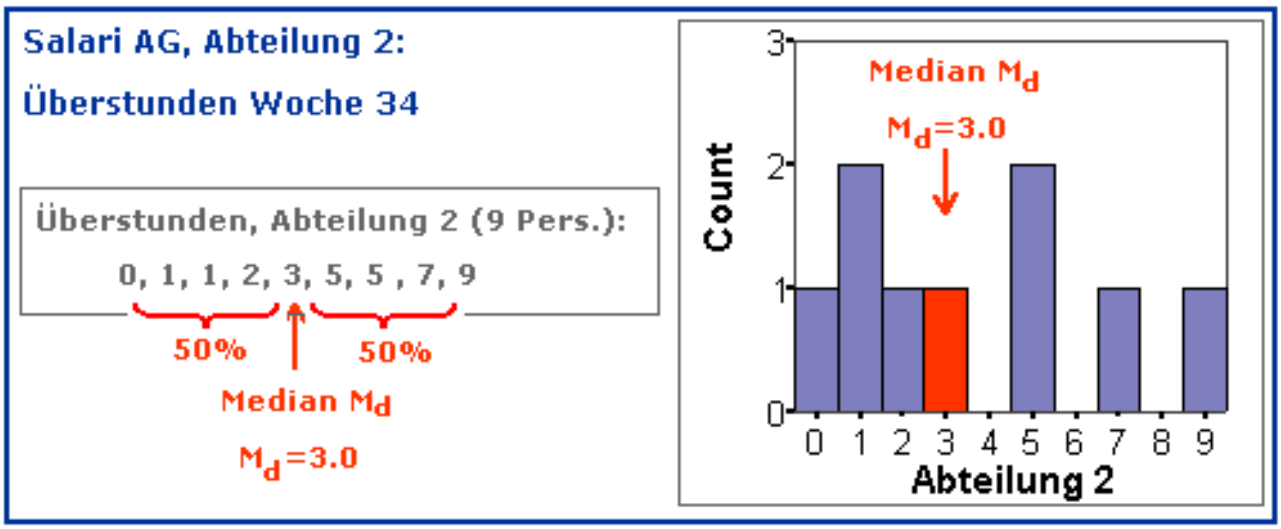
Beispiel 1
Die Angabe des Modus M_o ist dann sinnvoll, wenn seine Häufigkeit die anderen Häufigkeiten dominiert oder die Verteilung in der „Umgebung“ des Modus M_o eine erkennbare Konzentration aufweist.

Beispiel 2
Bei Beispiel 2 kann der Modalwert M_o nicht bestimmt werden.

Beispiel 3

Die Bestimmung des Modus M_o bei Beispiel 3 wird als nicht sinnvoll erachtet, weil sich die zum Modus M_o gehörende Häufigkeit nicht deutlich genug von den anderen Häufigkeiten abhebt.

Der Modus M_o kann auch für nominal skalierte Daten angegeben werden. Er bezeichnet in diesem Fall indessen einfach die Ausprägungs-Kategorie, die am häufigsten vorkommt. Von einer zentralen Tendenz kann aber nicht gesprochen werden.



Andere Bezeichnungen

Zentralwert, zentraler Wert

Charakterisierung

Nachdem die Ausprägungsgrade in eine Rangreihe gebracht wurden, kann der Median M_d bestimmt werden: Liegen über einem Wert genau so viele Ausprägungen wie unter diesem Wert, so wird dieser Wert als Median M_d bezeichnet.

Voraussetzungen

Da zur Bestimmung des Medians M_d die Ausprägungswerte in eine Rangordnung gebracht werden müssen, kann der Median M_d nur bestimmt werden, wenn das Merkmal mindestens ordinal skaliert ist.

Eignung

Im Gegensatz zum arithmetischen Mittel wirken sich einzelne extrem hohe oder extrem tiefe Ausprägungsgrade (Ausreisser oder Extremwerte) nicht auf den Median M_d aus. Es spielt somit keine Rolle, ob einzelne Werte weit oberhalb oder unterhalb des Medians M_d liegen. Dies bedeutet, dass der Median M_d or allem dann besonders nützlich ist, wenn stark asymmetrische (schiefe) Verteilungen oder Verteilungen mit einigen sehr extremen Werten beschrieben werden sollen.

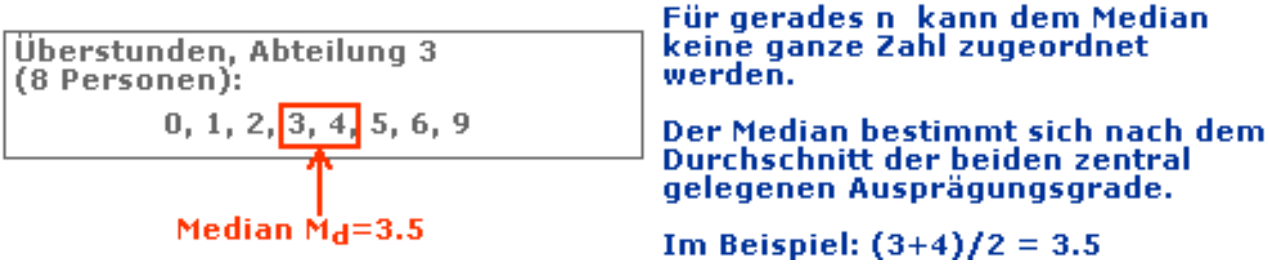
Hinweise und zusätzliche Erklärungen

Medianbestimmung bei ungerader Zahl von Beobachtungen und nur einmal vorkommenden Ausprägungsgraden bei intervall-skalierten Merkmalen

Die Berechnung des Medians M_d ist einfach, wenn ein Datensatz eine ungerade Anzahl von Beobachtungen aufweist.

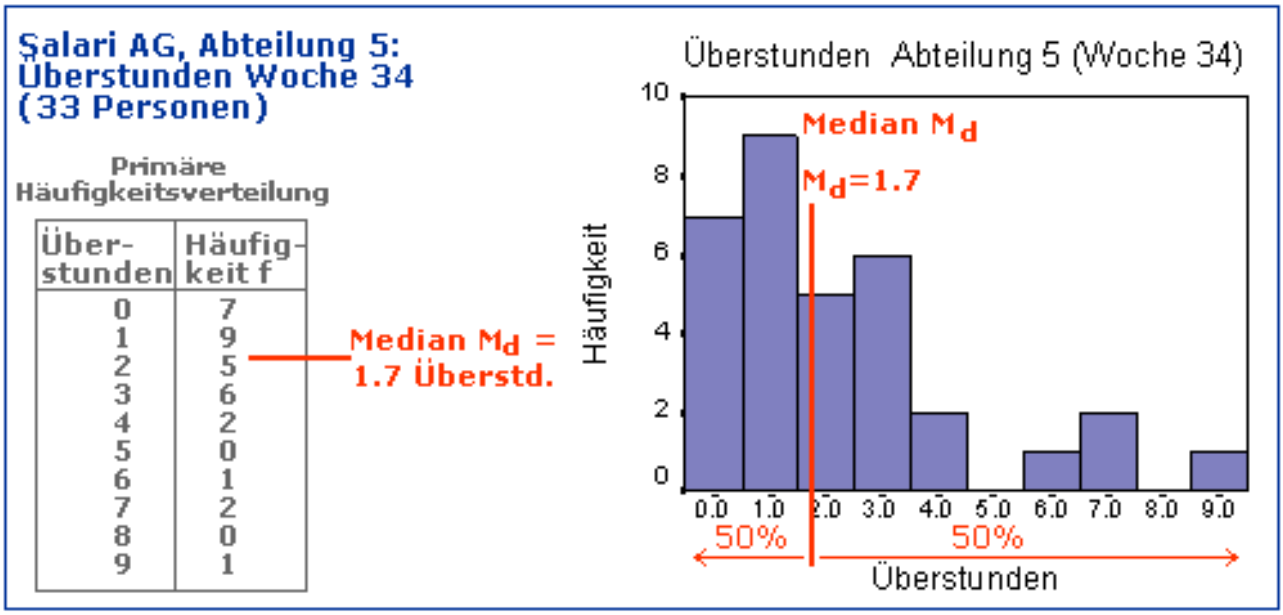
Medianbestimmung bei gerader Zahl von Beobachtungen und nur einmal vorkommenden Ausprägungsgraden bei intervall-skalierten Merkmalen

Bei ungerader Zahl von Beobachtungen berechnet sich der Median M_d als arithmetischer Mittelwert der beiden zentralen Werte der geordneten Datenreihe.



Medianbestimmung bei Häufigkeitsverteilungen bei intervall-skalierten Merkmalen

Bei Häufigkeitsverteilungen kann der Median M_d nicht so einfach bestimmt werden. Mit einer linearen Interpolation kann der Median M_d jedoch auch in diesem Fall entweder manuell oder mit der Statistiksoftware SPSS näherungsweise bestimmt werden. Wie aus der Abbildung ersichtlich ist, entspricht der Median M_d in diesem Fall keinem tatsächlich beobachteten Ausprägungsgrad.



Medianbestimmung bei ordinal skalierten Merkmalen

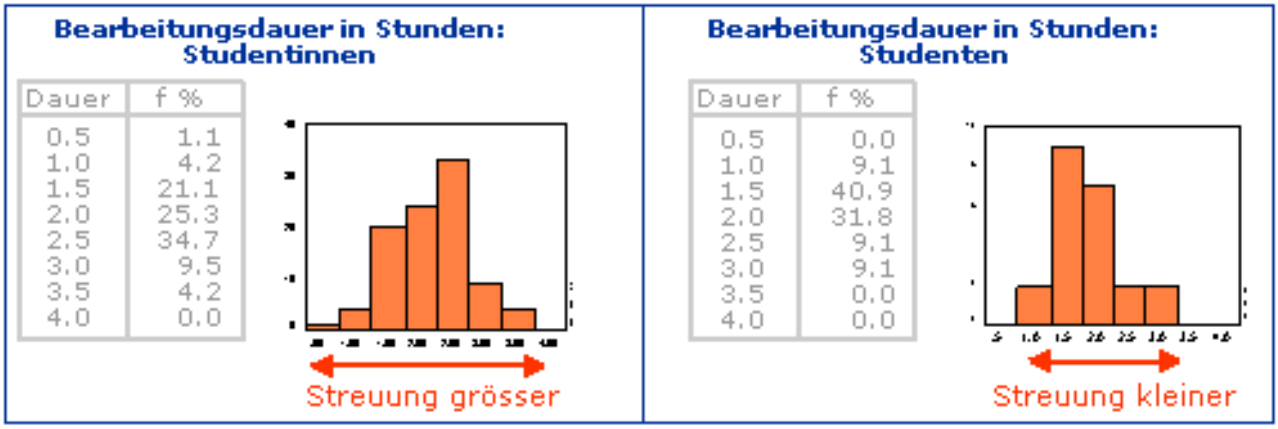
Bei ordinal skalierten Daten mit einer graden Anzahl von Beobachtungen und nur einmal vorkommenden Ausprägungsgraden ist es nicht möglich, die Werte der beiden mittleren Fälle zu halbieren wie bei intervall-skalierten Daten. Man beschränkt sich deshalb darauf anzugeben, zwischen welche zwei Beobachtungen der Median M_d zu liegen kommt.

Bei Häufigkeitsverteilungen kann bei ordinal skalierten Merkmalen eine lineare Interpolation nicht durchgeführt werden. Die Suche nach dem Median M_d beschränkt sich deshalb auf die Identifizierung der Kategorie, in die der Median M_d fällt.

3. Masse zur Beschreibung der Variabilität (Streuung)

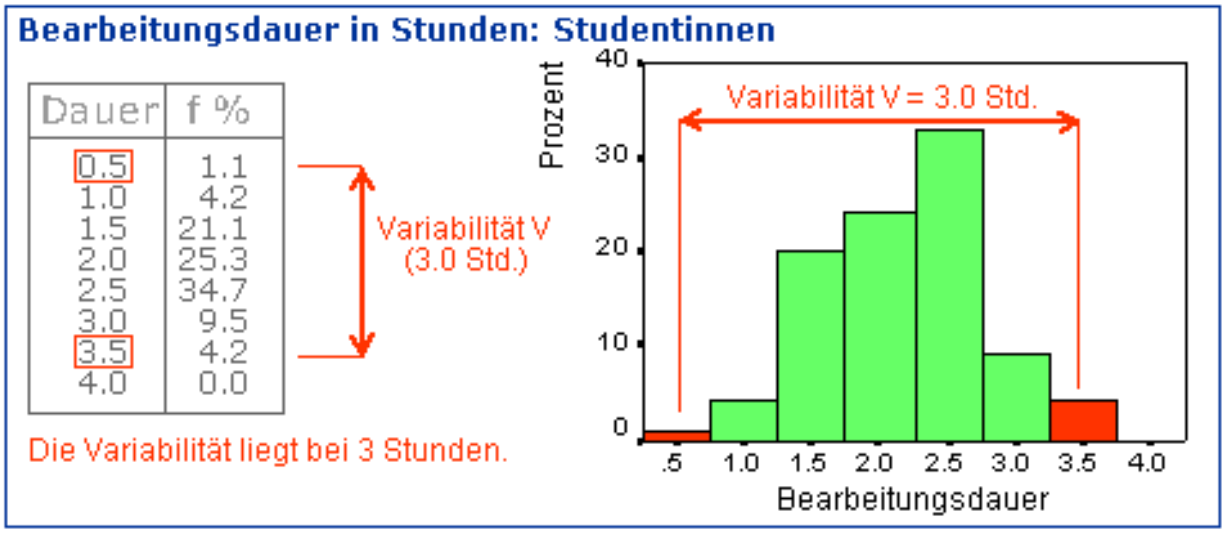
Arithmetischer Mittelwert, Median und Modus informieren über einen wichtigen Aspekt von Verteilungen, nämlich über die zentrale Tendenz von Beobachtungswerten. Sie geben jedoch keinen Aufschluss über die Homogenität bzw. Heterogenität der Beobachtungswerte.

Werden beispielsweise Studentinnen und Studenten nach der Bearbeitungsdauer für einen bestimmten Lernschritt befragt, kann es durchaus sein, dass sich die beiden Gruppen unterscheiden. Die folgende Abbildung zeigt deutlich, dass die beiden Gruppen, Studentinnen und Studenten, tatsächlich keine identische Häufigkeitsverteilung aufweisen, sondern dass die Ausprägungen bei Studentinnen stärker streuen. Es sind deshalb Kennwerte gesucht, die diese Streuung der Ausprägungsgrade in Form eines einzigen Kennwerts ausdrücken können.



Diese unterschiedliche Streuung von Beobachtungswerten kann mit verschiedenen Kennwerten beschrieben und ausgedrückt werden

- Die Variabilität V
- Die Interquartilweite QW und die Quartilabweichung Q
- Die Varianz s2 und die Standardabweichung s



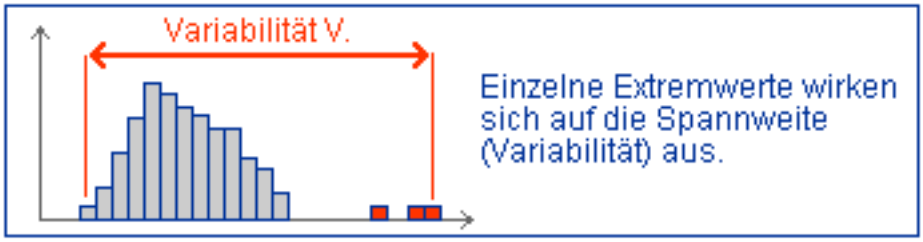
Andere Bezeichnungen
Spannweite, Variationsbreite

Charakterisierung und Voraussetzungen

Bei intervall-skalierten Merkmalen
Die Variabilität entspricht der Differenz zwischen dem grössten und dem kleinsten beobachteten Merkmalswert (V = Maximaler Ausprägungsgrad – minimaler Ausprägungsgrad).

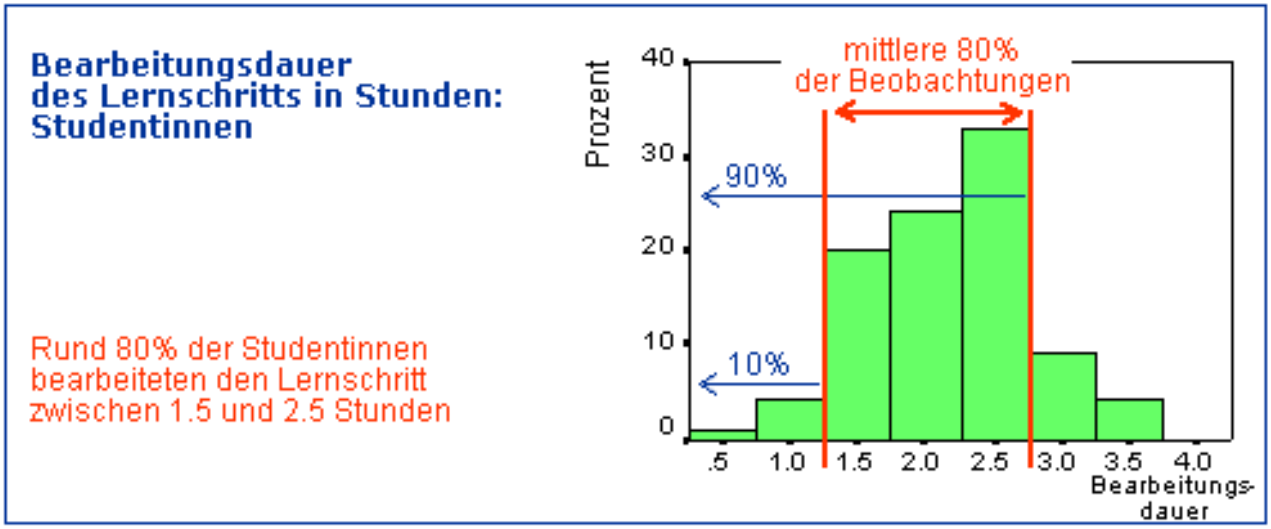
Bei ordinal skalierten Merkmalen
Für ordinal skalierte Merkmale kann die Variabilität ermittelt werden, indem nicht die Differenz, sondern die Spannweite durch die Nennung des grössten und kleinsten Merkmalswertes angegeben wird. In diesem Fall wird beispielsweise angegeben, dass bei Psychologie-Studierenden die Angabe „Repetition des Lernstoffs“ zwischen „wöchentlich“ und „in den Semesterferien“ streut.

Eignung
Die Spannweite kann einfach und rasch berechnet werden. Sie kann jedoch durch das Vorhandensein von einzelnen Extremwerten, die für die Verteilung nicht repräsentativ sind, leicht verzerrt werden. Einzelne Extremwerte an den beiden Enden einer Datenreihe können die Spannweite nämlich erheblich vergrössern und mehr Variabilität suggerieren als tatsächlich vorhanden ist. Die Variabilität gibt aber lediglich die Länge des Streubereichs an und beschreibt nicht, wie die einzelnen Beobachtungen in diesem Bereich streuen.



Einleitung

Der Variabilität kann entnommen werden, in welchem Bereich sich Merkmalsausprägungen befinden. Weil dieses Mass stark von Extremwerten abhängt, kann es in gewissen Fällen angebracht sein, nur einen eingeschränkten Streubereich zu betrachten, z.B. nur die mittleren 80 % aller Merkmalsausprägungen. Dieser Bereich wird durch Merkmalsausprägungen begrenzt, welche die oberen 10% sowie die unteren 10% der Häufigkeitsverteilung abschneiden. Es werden somit Grenzwerte gesucht, unterhalb derer 10% und 90% Prozent aller Merkmalsausprägungen liegen. Solche Grenzwerte sind: Die Perzentil- oder Zentilwerte, die Dezilwerte und die Quartilwerte.



Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Perzentil- bzw. Zentilwerte

Ein Perzentilwert P_b bezeichnet denjenigen Ausprägungsgrad des Merkmals, unterhalb dessen b Prozent der Beobachtungen liegen. Perzentilwerte P_b werden auch als Zentilwerte Z_b bezeichnet und weisen einen Wertebereich von $P_1 – P_{99}$ bzw. $Z_1 – Z_{99}$ auf.

Beispiele

- P_{90} bzw. Z_{90} bezeichnet den Ausprägungsgrad des Beobachtungsmerkmals, unterhalb dessen 90% aller Beobachtungen liegen.
- P_{50} bzw. Z_{50} bezeichnet den Ausprägungsgrad des Beobachtungsmerkmals, unterhalb dessen 50% aller Beobachtungen liegen (und entspricht somit dem Median M_d).

Dezilwerte

Die Dezilwerte D_q (Bereich $D_1 – D_9$) entsprechen den Perzentilwerten in 10er-Abstufungen.

Beispiele

- D_9 entspricht dem Ausprägungsgrad des Beobachtungsmerkmals, unterhalb dessen 90% aller Beobachtungen liegen.
- $D_5=P_{50}=M_d$

Die Statistiksoftware SPSS berechnet die Perzentil- und Dezilwerte anhand einer linearen Interpolation, falls gewünscht. Diese Werte entsprechen dadurch nicht mehr tatsächlich beobachteten Ausprägungsgraden. So wird im obigen Beispiel bei P_{90} der Wert „2.9“ Stunden angegeben, obwohl nur die Ausprägungsgrade „0.5, 1.0, 1.5, 2.0, etc.“ Stunden gewählt werden konnten.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Die Quartile Q_1 , Q_2 , Q_3 beschreiben diejenigen Ausprägungsgrade, unterhalb denen 25%, 50% bzw. 75% aller Beobachtungen liegen.

Zusammenhang mit Perzentilen und Median

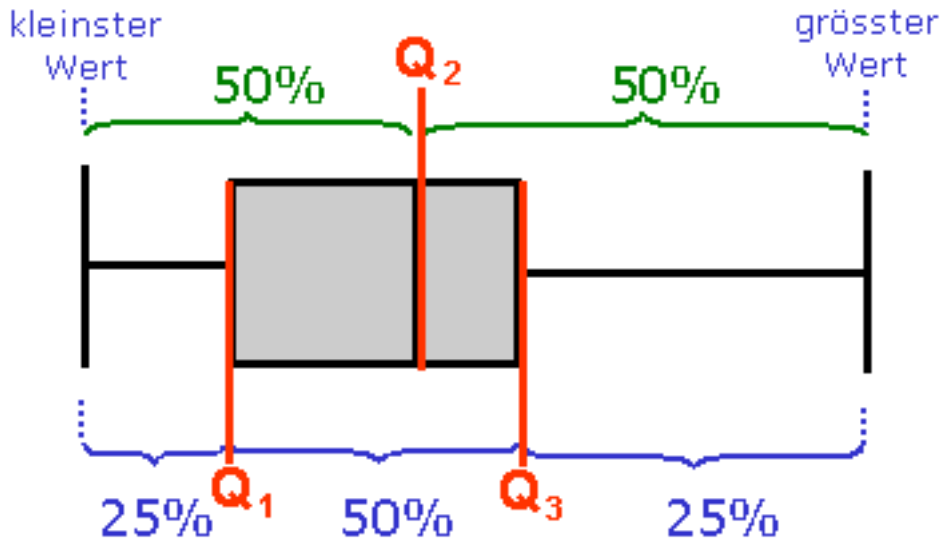
- $Q_1=P_{25}$
- $Q_2=P_{50}=M_d$
- $Q_3=P_{75}$

Berechnung der Grenzwerte manuell oder mit SPSS

Wenn ein gewünschter Perzentilwert, Zentilwert oder Quartilwert aus einer primären Häufigkeitsverteilung nicht abgelesen werden kann, so kann anhand einer linearen Interpolation der gesuchte Grenzwert manuell bestimmt werden. Die Statistiksoftware SPSS berechnet dies, falls gewünscht. Der Grenzwert entspricht dadurch nicht mehr einem tatsächlich beobachteten Ausprägungsgrad. So weist in unserem Beispiel in der obigen Grafik Q_1 den Wert „1.7“ Stunden auf.

Zur Angabe von Quartilwerten bei Boxplots

Wie unter „Grafische Darstellung univariater Verteilungen“ angesprochen, können Boxplots (mit Angabe des ersten, zweiten und dritten Quartils sowie des kleinsten und grössten Wertes) einen Datensatz knapp, aber informativ zusammengefasst werden.

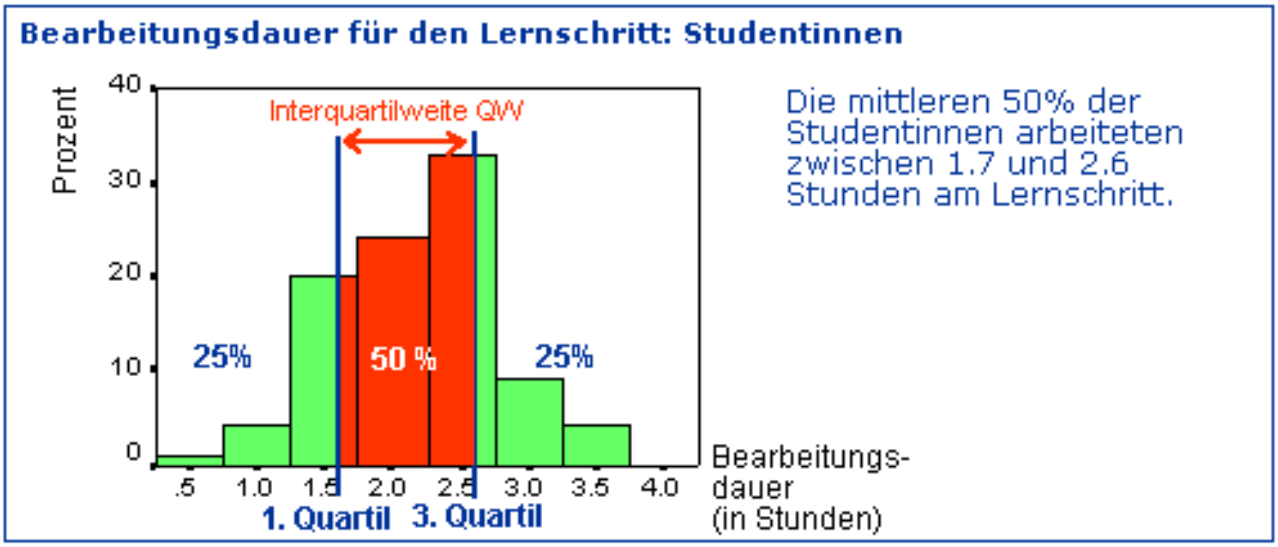


Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Wenn einige Merkmalsausprägungen von der „Box“ (dem Kästchen in der Mitte) weit entfernt zu liegen kommen (sogenannte Ausreisser und Extremwerte), würden in einer Grafik kurze Boxen mit extrem langen „Schweifen“ gezeichnet werden müssen. Da solche grafischen Darstellungen aus einem langen Schweif für einzelne, wenige Werte und einer kleinen Box für 50% der Ausprägungen bestehen und dadurch wichtige Informationen nicht auf den ersten Blick erfasst werden können, stellt das Statistikprogramm SPSS den Boxplot wie folgt dar:
„Schweife werden nur bis zu einer maximalen Länge von 1.5 Boxlängen (Interquartilweiten) aufgeführt. Werte, die ausserhalb zu liegen kommen, werden speziell gekennzeichnet und wie folgt bezeichnet:

- **Ausreisser (outliers):** Werte, die zwischen 1.5 und 3 Boxlängen (Interquartilweiten) ausserhalb der Box liegen.
- **Extremwerte (extremes):** Werte, die mehr als 3 Boxlängen (Interquartilweiten) ausserhalb der Box liegen.

Die Angabe der Extremwerte und der Ausreisser kann bei SPSS unterdrückt werden, so dass nur durch eine zusätzliche Erläuterung ersichtlich ist, dass Daten ausgeblendet wurden!



Andere Bezeichungen
Für Interquartilweite: der zentrale Quartilabstand
Für Quartilabweichung: der mittlere Quartilabstand

Charakterisierung
Mit der Interquartilweite QW und der Quartilabweichung Q wird auf unterschiedliche Weise die „Breite“ der Ausprägungsbereiches beschrieben, in dem die zentral gelegenen 50% aller Merkmalsausprägungen liegen.

Interquartilweite QW
Der Interquartilweite entspricht dem Bereich zwischen dem 1. und 3. Quartil, d.h. sie entspricht der Entfernung zwischen den beiden Merkmalswerten, welche die in der Rangordnung zentral gelegenen 50% der Merkmalsausprägungen eingrenzen ($QW = Q_3-Q_1$).

Quartilabweichung Q
Quartilabweichung ist ein anderes übliches Mass und entspricht der Hälfte der Interquartilweite ($Q= QW/2= (Q_3-Q_1)/2$).

Voraussetzungen
Die Berechnung der Interquartilweite setzt ein mindestens intervall-skaliertes Merkmal voraus.

Eignung
Die Ermittlung der Interquartilweite und der Quartilabweichung ist vor allem sinnvoll, wenn der Kernbereich einer Häufigkeitsverteilung interessiert, d.h. wenn die zentral gelegenen 50% der Merkmalsausprägungen interessieren. Im Unterschied zur Variabilität tritt dabei das Ausreisser-Problem nicht auf, da die oberen und unteren 25% der Merkmalsausprägungen abgeschnitten werden.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Charakterisierung

Mit der Varianz „s^2“ bzw. der Standardabweichung „s“ soll die mittlere „Abweichung“ der Ausprägungen eines Merkmals vom arithmetischen Mittelwert aller Merkmalsausprägungen ermittelt werden.

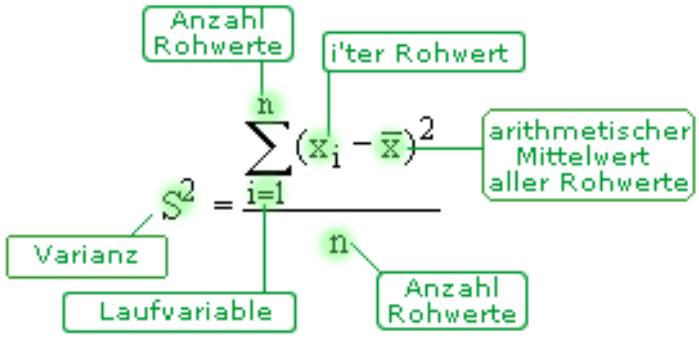
Bemerkung zur Berechnung

Die einzelnen Abweichungen vom Mittelwert können nicht einfach aufaddiert werden, da diese sowohl positive wie negative Vorzeichen aufweisen und ihre Summe deshalb gleich Null wäre.
Deshalb werden die „Abweichungen“ der einzelnen Ausprägungen des Merkmals vom arithmetischen Mittelwert quadriert. Damit haben die quadrierten „Abweichungen“ immer ein positives Vorzeichen und können dann aufsummiert werden.

Formel zur Berechnung der Varianz s^2

Der Varianz entspricht die Summe der quadrierten Abweichungen der Merkmalswerte vom arithmetischen Mittelwert, dividiert durch die

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



Anzahl der Merkmalsträger.

Formel zur Berechnung der Standardabweichung s

Da man bei Angabe der Varianz Ergebnisse mit dem Quadrat der Masseinheit der zugrundeliegenden Daten erhält (z.B. $[0.6\ \text{Std.}]^2$) und sich aus einem solchen Wert kaum Interpretationen in bezug auf die Realität ableiten lassen, wird an Stelle der Varianz überall dort, wo es um inhaltliche Aussagen geht, die Standardabweichung verwendet. Man erhält die Standardabweichung s, indem man die Quadratwurzel aus der Varianz berechnet.

$$s = \sqrt{s^2}$$

Voraussetzungen

Das Merkmal muss mindestens intervall-skaliert sein, da die Abstände zwischen den Merkmalsausprägungen und dem arithmetischen Mittelwert zu berechnen sind.

Eignung

Die Varianz s^2 bzw. die Standardabweichung s werden als sinnvolles Mass zur Beschreibung der Variabilität angesehen, wenn die Daten eingipflig und näherungsweise symmetrisch sind.

Hinweise und zusätzliche Erklärungen

Welche Bedeutung haben s^2 und s?

In der beschreibenden Statistik haben diese Masse wegen ihrer kleinen Anschaulichkeit und der Schwierigkeit, diese Masszahlen zu interpretieren, keine so grosse Bedeutung. In der schliessenden Statistik (prüf- und entscheidungsstatistische Verfahren) haben sie als rechentechnische Grösse eine herausragende Bedeutung.
Wie später unter „Wahrscheinlichkeits-Verteilungen“ belegt wird, liegen bei normalverteilten Daten zwischen den Grenzen „

$\bar{x}-s$

“ und „

$\bar{x}+s$

“ 68,3 % aller Beobachtungen; zwischen den Grenzen „

$\bar{x}-2s$

“ und „

$\bar{x}+2s$

“ 95.5 % aller Beobachtungen.

Unterschiedliche Bezeichnungen der Varianz und der Standardabweichung

Mit s^2 und s wird die Varianz und die Standardabweichung für Daten bezeichnet, die aus einer Stichprobe stammen. Man spricht dabei von Verteilungskennwerten.
Werden die Daten an einer Population erhoben, so wird die Varianz mit

σ^2

(sigma Quadrat) und die Standardabweichung mit

σ

(sigma) bezeichnet, und man spricht von Verteilungsparametern.

Berechnung der Varianz aus einer Häufigkeitsverteilung

Wie schon unter „Berechnung des arithmetischen Mittelwertes aus einer Häufigkeitsverteilung“ erwähnt, gehen durch die Kategorisierung der Daten Informationen verloren, so dass die Masszahl nur näherungsweise bestimmt werden kann. Dies gilt auch für die Varianz. Nach Möglichkeit soll sich die Varianz deshalb auf die Rohdaten beziehen.

$$S^2 = \frac{\sum_{j=1}^{m} f_j \cdot (x_j - \bar{x})^2}{\sum_{j=1}^{m} f_j}$$

Häufigkeit der j'ten Auswertungskategorie
Mittelwert der j'ten Auswertungskategorie
Anzahl Auswertungskategorien
arithmetischer Mittelwert aller Rohwerte
Varianz
Anzahl Auswertungskategorien
Laufvariable

4. Zusammenfassung zum Lernschritt

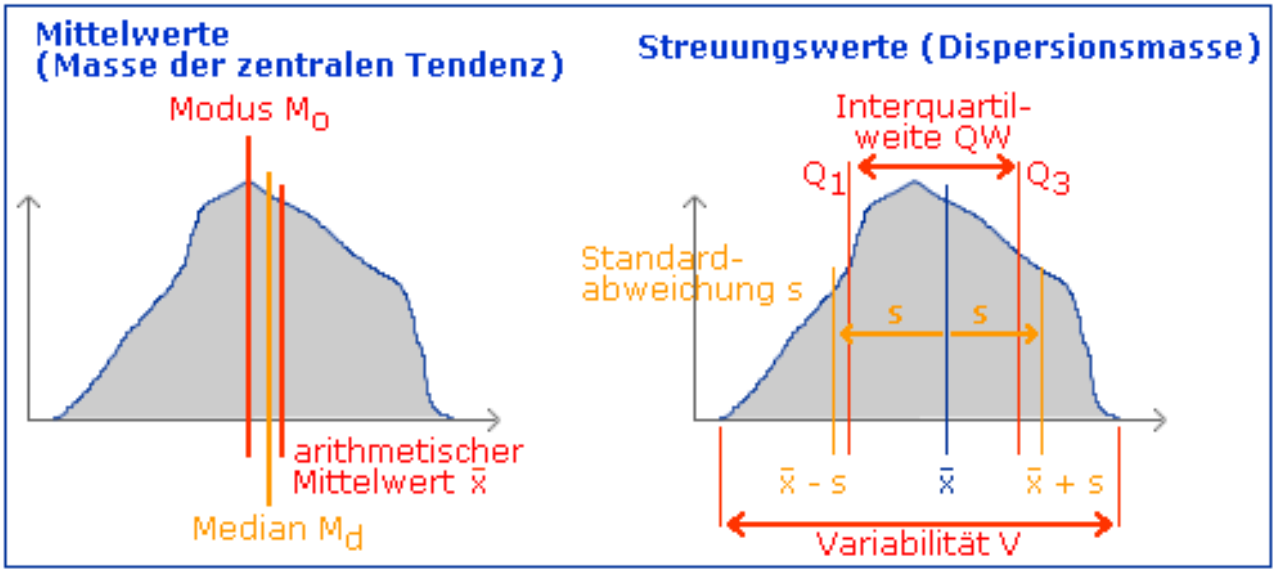
Verteilungsformen

Zur Charakterisierung der Form einer Häufigkeitsverteilung werden häufig folgende Begriffe verwendet: Ein- bzw. Mehrgipfligkeit (unimodal, bimodal und multimodal), Wölbung (schmalgipflig oder breitgipflig), Schiefe (symmetrisch oder asymmetrisch, linkssteil oder rechtssteil) oder auch spezielle Formen (u-förmig, abfallend oder glockenförmig).

Kennwerte zur Beschreibung von Häufigkeitsverteilungen

Eine Reihe von Kennzahlen, welche die Daten möglichst gut repräsentieren, können zur Beschreibung von Verteilungen verwendet werden. Dabei werden viele Einzelinformationen zu wenigen, aber aussagekräftigen Grössen verdichtet.

2 Gruppen von Kennwerten zur Beschreibung von Häufigkeitsverteilungen



1. Masse der zentralen Tendenz (Mittelwerte)
Auf unterschiedliche Weise kann der „Schwerpunkt“ oder „Mittelwert“ einer Verteilung berechnet werden. Dies ermitteln die sogenannten Masse der zentralen Tendenz, auch Lagemasse genannt. Sie beschreiben den typischen, den zentralen oder durchschnittlichen Wert einer Verteilung.

- **Der arithmetische Mittelwert**

  $\bar{x}$

  [mean].
  Die Summe aller Ausprägungswerte wird durch die Anzahl aller Beobachtungen dividiert.
- **Der Modalwert Mo [mode].** Der Modus ist derjenige Ausprägungsgrad der Merkmalsdimension, der am häufigsten beobachtet wurde.
- **Der Median Md [median].** Dem Median entspricht der Grenzwert, unterhalb dessen 50% der Beobachtungen liegen.

2. Streuungswerte (Dispersionsmasse)
Häufig interessiert die Frage, wie dicht die einzelnen Daten beieinander liegen, oder wie stark sie streuen. Man spricht von einer grossen Streuung, wenn die Daten weit auseinander liegen. Diese Masszahlen charakterisieren somit die Variabilität oder Heterogenität der Beobachtungen.

- **Die Variabiliät V [range].** Die Variabilität entspricht der Differenz zwischen dem grössten und dem kleinsten beobachteten Merkmalswert (V = Maximaler Ausprägungsgrad – minimaler Ausprägungsgrad).
- ie Interquartilweite QW [interquartil range] und Quartilabweichung Q. Mit der Interquartilweite QW und der Quartilabweichung Q wird auf unterschiedliche Weise die „Breite“ des Wertebereiches beschrieben, in dem die zentral gelegenen 50% der Merkmalsausprägungen liegen [QW = Q_3-Q_1 bzw. Q= QW/2= (Q_3-Q_1)/2].
- **Die Varianz s^2 [variance] und die Standardabweichung s [standard deviation]**. Der Varianz entspricht die Summe der quadrierten Abweichungen der Merkmalswerte vom arithmetischen Mittelwert, dividiert durch die Anzahl der Beobachtungen. Man erhält die Standardabweichung s, indem man die Quadratwurzel aus der Varianz s^2 berechnet.

Zusammenfassung (als pdf 151 KB Grösse) zum Ausdrucken.

Fallbeispiel

Experiment: Können kleine Quadrate mit der Maus ebenso schnell angeklickt werden wie grössere Quadrate? (Vergleich zweier Häufigkeitsverteilungen)
Im folgenden Experiment soll untersucht werden, ob kleine Quadrate mit der Computer-Maus ebenso schnell angeklickt werden können wie grössere Quadrate.
Mit dem Experiment kann begonnen werden, sobald unten die Schaltfläche "Start" erscheint. Dies kann, abhängig von Ihrer Internetverbindung, bis 1 Minute dauern.

Durchführung des Experiments
Sobald Sie "Start" klicken, erscheint rechts ein Quadrat, in welches Sie mit der Computer-Maus so schnell wie möglich klicken sollen. Die dazu benötigte Zeit wird in Millisekunden gemessen.

Nach 2 Probedurchgängen und je 12 Versuchen werden Ihre Daten analysiert und Ihnen diese in einem Histogramm präsentiert.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Lernkontrolle

Mit dem Klick auf die Schaltfläche "Start" wird ein interaktives Histogramm heruntergeladen. Dies kann, abhängig von Ihrer Internetverbindung, bis 1 Minute dauern.
Die Häufigkeitsverteilung können Sie durch Klicken ins Histogramm verändern und dabei die Veränderung von Mittelwert, Median, Standardabweichung und Schiefe beobachten.
Die unten aufgeführten Fragen sollen Ihnen aufzeigen, ob sie den Lernstoff begriffen haben. Das interaktive Histogramm kann Ihnen bei der Lösungsfindung nützlich sein.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]