# Beschreibende Statistik
# Daten darstellen und charakterisieren

Roland Heynkes

1. April 2006, Aachen

Die beschreibende (descriptive) Statistik versucht, große und unübersichtliche, experimentell sowie durch Beobachtung oder Befragung gewonnene Datenmengen durch graphische Darstellung auf einen Blick verständlich zu machen. Wichtige Werkzeuge für diesen Zweck sind Diagramme, von denen es verschiedene Typen in vielen Variationen gibt, die aber immer Daten oder allgemein Informationen graphisch darstellen.

Wenn Häufigkeitsverteilungen quantitativer Merkmale wie Größen, Gewichte, Temperaturen oder Geschwindigkeiten eine große Anzahl von Meßwerten enthalten, dann sind diese selten gleichmäßig über den gesamten Wertebereich verteilt. Meistens bilden die Daten Schwerpunkte mit großen Häufungen sehr ähnlicher Werte, während mit zunehmender Abweichung von solchen Schwerpunkten die Zahl entsprechender Daten abnimmt. Lage und Streuung der Daten beschreibt man mathematisch durch Maße, bei denen zwischen Lagemaßen und Streumaßen unterschieden wird.

# Inhaltsverzeichnis

# 1 Erste Grundbegriffe der beschreibenden Statistik

## 1.1 Merkmal

Merkmale nennt man die Eigenschaften der Merkmalsträger, die bei statistischen Untersuchungen jeweils von Interesse sind [4, 7]. Die manchmal auch Attribute genannten Merkmale sind rein mathematisch betrachtet Variable [6], die verschiedene Merkmalsausprägungen annehmen können [3, 6]. Das Merkmal auf dem Merkmalsträger Stimmzettel ist die angekreuzte Partei [2]. Man unterscheidet qualitative von den von metrischen oder metrisch (in Zahlen) skalierten quantitativen Merkmalen [2, 3].

### 1.1.1 qualitative Merkmale

Die auch Kategorien genannten Ausprägungen qualitativer Merkmale können entweder nur benannt (nominalskaliert) oder sinnvoll geordnet (ordinalskaliert) werden [2, 3].

**Nominale oder nominalskalierte Merkmale** können nur benannt werden [2, 3, 4]. Beispiele wären das Geschlecht (männlich/weiblich), der erlernte Beruf oder die gewählte Partei [2, 3, 4]. Selbst wenn man nominale Merkmale durch Zahlen kodiert, ergeben Mittelwertbildungen keinen Sinn. Praktikabel sind Nominalsysteme nur, wenn sie erschöpfend und eindeutig sind [4, S.7]. Dazu müssen alle theoretisch möglichen Ausprägungen genannt werden und diese müssen sich außerdem gegenseitig ausschließen (z.B. beim Familienstand: ledig/verheiratet/verwitwet/geschieden) [4].

**Ordinale oder ordinalskaliert Merkmale** können anhand einer Eigenschaft geordnet werden [2, 3, 4]. Man spricht auch von Rangmerkmalen, wenn die Ausprägungen eine natürliche bzw. logische Reihenfolge haben [2]. Beispiele wären Schulnoten, eine Gewichtsklassifizierung (z.B.: untergewichtig, normal, übergewichtig) oder militärische Ränge [2, 3, 4]. Normalerweise sind auch bei ordinalskalierten Merkmalen Mittelwertbildungen sinnlos, aber beispielsweise bei Schulnoten funktioniert es doch.

### 1.1.2 quantitative Merkmale

Mit quantitativen Merkmalen kann man rechnen, weil die Ausprägungen Zahlen sind und nicht nur mit Zahlen kodiert werden.

**Intervallskalierte Merkmale** erlauben Aussagen über das Ausmaß des Unterschiedes zwischen zwei Daten, weil die Abstände zwischen aufeinanderfolgenden Stufen gleich groß sind. Typische Intervalldaten sind z.B. Temperaturangaben in Celsius. [4]

**Verhältnisskalierte Merkmale** besitzen auf natürliche Nullpunkte bezogene Meßskalen und werden auch als absolutskaliert bezeichnet. So kann man nicht nur die Differenz, sondern auch einen sinnvollen Quotienten zweier Merkmalsausprägungen erfassen. Verhältnisskalierte Merkmale sind z.B. Körpergröße und Monatseinkommen. [4]

Man kann Merkmale auch nach der Zahl der möglichen Ausprägungen unterscheiden:

**Diskrete Merkmale** können wie die natürlichen Zahlen nur ganz bestimmte und daher abzählbare Werte annehmen, zwischen denen Lücken oder Stufen sind [3, 4].

**Stetige Merkmale** können wie Dezimalzahlen innerhalb jedes beliebigen Intervalls unendlich viele verschiedene Zwischenwerte annehmen (z.B.: Wellenlängen) [3, 4].

## 1.2 Merkmalsträger

Merkmalsträger sind statistische Einheiten wie Personen, Ereignisse oder Stimmzettel, für die Ausprägungen von Merkmalen (Eigenschaften) erhoben werden können [3, 4, 6].

## 1.3 Merkmalsausprägungen

Merkmalsausprägungen (Ausprägungen), Merkmalswerte oder Kategorien nennt man die verschiedenen Werte oder Formen, die ein Merkmal (eine Variable) annehmen [2, 3, 6] und die man per Datenerhebung feststellen kann [4, 7]. Merkmale kann man deshalb auch als die Summe der möglichen Ausprägungen betrachten [2, 3, 6]. Auf einem Stimmzettel (Merkmalsträger) gibt es für das Merkmal der angekreuzten Partei z.B. die Ausprägungen SPD, CDU, CSU, FDP, Grüne und Linke [2, 4].

## 1.4 Erhebungen

Statistisch auszuwertende Daten können experimentell, durch Beobachtung oder durch Befragung gewonnen werden [3, 6]. Kann oder will man aus praktischen Gründen nicht alle interessierenden Personen befragen (Vollerhebung), muß man eine möglichst repräsentative Stichprobe auswählen (Teilerhebung) [3].

## 1.5 Datenreihe

Datenreihe oder Messreihe nennt man die Menge aller Ergebnisse einer statistischen Untersuchung [4, 7]. Eine besondere Datenreihe ist die sogenannte Urliste, die die Daten in der Reihenfolge enthält, in der sie angefallen sind [3, 4].

## 1.6 Grundgesamtheit

Die auch Grundpopulation genannte Grundgesamtheit besteht laut einem Mathematikschulbuch von Schroedel aus den Merkmalsträgern [2]. Diese Definition ist also abhängig davon, was man jeweils als die Menge aller Merkmalsträger versteht. Als Merkmalsträger nennt dieses Mathematikbuch bei Wahlen die abgegebenen Stimmzettel oder bei Umfragen alle Befragten, aber nicht alle Wahlberechtigten oder alle potentiell Befragbaren [2]. Im Gegensatz dazu definieren zwei interaktive Lehr-Lernprogramme zur Statistik [6, 7] die „Grundgesamtheit“ oder „Population“ als die Menge aller bezüglich des zu untersuchenden Merkmals gleichartigen Objekte, Individuen oder Ereignisse, die überhaupt zur betrachteten Menge gehören können. Demnach hängt es von der Fragestellung ab, wer oder was dazu gehört [6]. Das können die gesamte Bevölkerung oder alle wahlberechtigten Personen eines Landes oder auch nur alle Jugendlichen unter 16 Jahren sein [6]. Eine uneindeutige mittlere Position nimmt Frau Dr. Lauer ein, die zur Grundgesamtheit alle für die statistische Untersuchung relevanten Merkmalsträger (z.B. alle Wahlberechtigten im Politbarometer) zählt [4, 7]. Bei Umfragen wird aus der Grundgesamtheit eine möglichst repräsentative Stichprobe ausgewählt [3, 4, 5, 6, 7]. ILMES und Dr. Handl definieren die Grundgesamtheit als die Menge der Objekte, für welche die Aussagen einer Untersuchung gelten sollen [3, 5]. Dementsprechend heißt es aufpassen und nicht etwa von einer Umfrage unter Schülern auf die Anzahl der Kinder pro Haushalt schließen [2,

113], weil ja die Stichprobe keine kinderlosen Haushalte umfaßt. Im Zusammenhang mit Bundestagswahlen gelten in allen mir bekannten Internettexten alle Wahlberechtigten als Grundgesamtheit. Aus meiner Sicht ergibt dies auch mehr Sinn als eine Grundgesamtheit aller ausgezählten Stimmzettel, weil es schließlich bei einer Wahl um die Ermittlung des Willens der Wahlberechtigten geht und demnach von den abgegebenen Stimmen auf eine Volksentscheidung geschlossen werden soll. Hinzu kommt, daß eine Wahl aufgrund der Freiwilligkeit der Beteiligung in Deutschland immer eine Teilerhebung und nicht wie eine Volkszählung eine Vollerhebung ist. Demnach sind die Wähler eine Stichprobe der Wahlberechtigten, die wiederum eine Stichprobe aller Deutschen sind.

## 1.7 Häufigkeit

Die absolute Häufigkeit einer Merkmalsausprägung entspricht der Anzahl der Merkmalsträger mit dieser Merkmalsausprägung, also z.B. der Zahl der Stimmzettel, auf denen das Kreuz bei der SPD gemacht wurde [2, 3]. Den Anteil einer bestimmten Ausprägung an allen Ausprägungen eines Merkmals in einer Grundgesamtheit nennt man relative Häufigkeit und gibt sie in Prozentzahlen oder als Bruchteile von 1 an [2, 3, 4]. Absolute und relative Häufigkeiten werden in sogenannten Häufigkeitstabellen zusammengefasst [3]. Wächst von einer Wahl zur nächsten die Grundgesamtheit im Sinne der abgegebenen Simmzettel, dann kann die bei Wahlen entscheidende relative Häufigkeit der Nennung einer Partei abnehmen, obwohl die absolute Häufigkeit ihrer Nennung zunahm [2].

Bei ordinalen Merkmalen kann auch die Berechnung der sogenannten absoluten oder relativen kumulativen Häufigkeit sinnvoll sein, die sich bei jedem Schritt aus der Addition der aktuellen zur Summe der bisher abgearbeiteten Merkmalsausprägungen ergibt [3].

Bei Umfragen oder Messreihen gibt die Häufigkeit an, wie oft eine Antwort oder ein bestimmter Messwert in einer Stichprobe auftritt [6, 7]. Die Häufigkeit kann ohne Bezug zur Stichprobengröße als absolute Häufigkeit oder mit Bezug auf die Stichprobengröße, als deren Anteil oder relative Häufigkeit ausgedrückt werden [6].

## 1.8 Klassenbildung und Histogramme

Stellt man die Ausprägungen eines stetigen quantitativen Merkmals einer relativ geringen Anzahl von Merkmalsträgern in einer Häufigkeitstabelle oder einem Stabdiagramm dar, denn erkennt man Häufungen lediglich an einer erhöhten Dichte der Werte, aber man kann nur fast identische Werte nicht als mehrfach vorkommende Merkmalsausprägungen zählen. Dadurch werden Häufigkeitsangaben, Häufigkeitstabellen und Säulendiagramme unübersichtlich und wenig informativ [3, 4]. Manchmal kann man sich dann mit einem sogenannten Stengel-Blatt-Diagramm behelfen [2]. Ansonsten lässt sich dieses Problem überwinden, indem man Wertebereiche zu Klassen zusammenfasst und alle in eine Klasse fallenden Werte (klassierte Werte) zählt [2, 3, 4]. Durch eine Sortierung der Urliste zu einem sortierten Datensatz lässt sich das erleichtern [3]. Von klassierten Werten lassen sich wie gewohnt absolute, relative und kumulierte Häufigkeiten ermitteln. Die Ermittlung von Mittelwerten ist bei klassierten Daten nur näherungsweise möglich, indem man mit den mittleren Werten jeder Klasse rechnet [2]. Die graphische Darstellung der relativen Häufigkeiten klassierter Werte erfolgt in sogenannten Histogrammen wie in Abbildung 1 auf Seite 6.

# 2 Beschreibung verschiedener Diagrammtypen

## 2.1 Säulendiagramme

Zweidimensionale Stab- oder Säulendiagramme bestehen aus gleich breiten, nebeneinander stehenden Rechtecken, deren Höhen proportional zu den jeweiligen absoluten oder relativen Häufigkeiten der von ihnen repräsentierten Merkmalsausprägungen sind [2, 3, 4, 6]. Solche Diagramme werden etwa ab 15 verschiedenen Merkmalsausprägungen zu unübersichtlich. Es gibt auch dreidimensionale Säulendiagramme mit hintereinander stehenden Quadern. Säulendiagramme eignen sich besonders für die Sortierung verschiedener Ausprägungen nach deren Häufigkeit sowie generell für den Vergleich absoluter Häufigkeiten [2, 4]. Nach der Größe der Merkmalsausprägungen sortierte Säulendiagramme nennt man Paretodiagramme [3]. Sollen nicht zu viele Messungen eines quantitativen Merkmals verglichen werden, dann eignen sich hierfür Säulendiagramme, sofern sich die gemessenen Merkmalsausprägungen über der x-Achse sinnvoll sortieren lassen. Wurde hingegen einfach nur ein Merkmal bei sehr vielen Merkmalsträgern untersucht, ist ein Säulendiagramm unpraktisch [4].

Gruppen-Stabdiagramme wie in den Abbildungen 2 auf Seite 7 und 3 auf Seite 8 vergleichen die Verteilungen der Ausprägungen eines Merkmals bei zwei oder mehr Datenerhebungen durch nebeneinander stehende Säulen, die jeweils für eine Datenerhebung stehen [4]. Die Notwendigkeit einer Unterscheidung durch Farben oder Muster erlaubt bei qualitativen Merkmalen allerdings nur den Vergleich einer begrenzten Anzahl verschiedener Datenerhebungen in einem Gruppen-Stabdiagramm.

## 2.2 Balkendiagramme

Von Säulendiagrammen unterscheiden sich Balkendiagramme nur durch liegende Balken statt stehender Säulen, also durch ein Vertauschen von x- und y-Achse.

## 2.3 Block- oder Streifendiagramme

In den auch Streifendiagramme oder Komponenten-Stabdiagramme genannten Blockdiagrammen werden die Merkmalsausprägungen nicht durch einzelne Rechtecke oder Quader, sondern durch Abschnitte oder Blöcke innerhalb eines einzigen die Grundgesamtheit darstellenden Streifens, Balkens oder Quaders dargestellt [2, 4]. Diese Diagrammform eignet sich wie Säulen- und Balkendiagramme zur Darstellung absoluter Zahlenwerte, man kann damit aber auch gut wie in Abbildung 4 auf Seite 8 Anteile eines Ganzen darstellen, welches dann 100% entsprechen muß. Umfrageergebnisse mit der Möglichkeit von Mehrfachnennungen lassen sich mit Blockdiagrammen nur in absoluten Zahlen, nicht aber in Prozentwerten darstellen [2]. Stellt man in einem Blockdiagramm absolute Zahlen dar, dann ergibt sich die Gesamtlänge des Streifens aus der Summe aller Blöcke. Es geht es hingegen um Anteile eines Ganzen, dann darf die Breite des Streifens beliebig gewählt werden. Da man gut mehrere Blockdiagramme aufeinander stapeln kann, eigenen sich diese besonders gut für Vergleiche zwischen den Ergebnissen verschiedener Wahlen oder Umfragen [2, 4]. Sie eignen sich aber nicht so gut wie Kreisdiagramme zum Ablesen von Mehrheitsverhältnissen z.B. nach Wahlen [2], weil sich Strecken nicht ganz so leicht wie Winkel vergleichen lassen. Ein interaktives Lehr-Lernprogramm zur Statistik bezeichnet auch dreidimensionale Säulendiagramme als Blockdiagramme [6].

## 2.4 Kreisdiagramme

Die auch Tortendiagramme genannten Kreisdiagramme eignen sich nur zur Darstellung der relativen Häufigkeiten nicht zu vieler Ausprägungen, von denen jede durch einen Kreissektor repräsentiert wird [2, 4]. Das Verhältnis des Mittelpunktwinkels des Kreissegments zum Vollwinkel von 360° entspricht der relativen Häufigkeit der Merkmalsausprägung [2]. Kreisdiagramme wie in Abbildung 5 auf Seite 9 eignen sich etwas besser als Blockdiagramme zum Vergleich verschiedener Gruppen von Merkmalsausprägungen. Sie eignen sich aber deutlich schlechter für Vergleiche zwischen den Ergebnissen verschiedener Wahlen oder Umfragen. Kreisdiagramme werden außerdem mit zunehmender Anzahl verschiedener Merkmalsausprägungen immer unübersichtlicher [4].

## 2.5 Linien-, Flächen, Netz- und Streudiagramme

Wenn große Mengen ein- oder zweidimensionaler Daten eine Darstellung in Säulen-, Balken-, Block- oder Kreisdiagrammen unmöglich machen und eine Verbindung benachbarter Datenpunkte sinnvoll ist, dann eignen sichzur Darstellung Linien-, Netz- oder Flächendiagramme. Gibt es zwischen sehr vielen Datenpunkten nur insgesamt eine Beziehung, dann kann man sie als unabhängige Punkte in ein zwei- oder dreidimensionales Streudiagramm eintragen.

## 2.6 Stengel-Blatt-Diagramme

Wären beispielsweise einige Autos 51, 55, 59, 61, 62, 64, 65, 65, 66, 66, 66, 67, 68, 68, 70, 73, 78 und 85 Monate alt, dann könnte man diese Alter in Jahre und Monate umrechnen und folgendermaßen als Stengel-Blatt-Diagramm darstellen.

| Jahr | Monate |
|---|---|
| 4 | 3 7 11 |
| 5 | 1 2 4 5 5 6 6 6 7 8 8 10 |
| 6 | 1 6 |
| 7 | 1 |

Tabelle 1: Darstellung einer Altersverteilung als Stengel-Blatt-Diagramm

Diese Darstellung (Stengel = Jahr, Blatt = Monat [2]) ist weder auf den ersten Blick verständlich, noch besonders ansprechend. Sie stellt aber immerhin die noch vollständig vorhandenen Informationen geordneter und übersichtlicher als die reine Aufzählung in der Urliste dar [4]. Die Länge der Zeilen vermittelt einen groben Eindruck von der Verteilung der Daten [4].

## 2.7 Histogramme

In einem einfachen Histogramm könnte man auf die feine Abstufung in Monaten verzichten und Klassenbreiten von einem Jahr wählen, um die Altersverteilung übersichtlicher als das Stengel-Blatt-Diagramm (Tabelle 1) zu machen.

Mit der gesamten Breite jeder Klasse als Grundlinie zeichnet man ohne Lücke Rechtecke über die Klassen, deren Flächen den absoluten oder relativen Häufigkeiten der Klassen

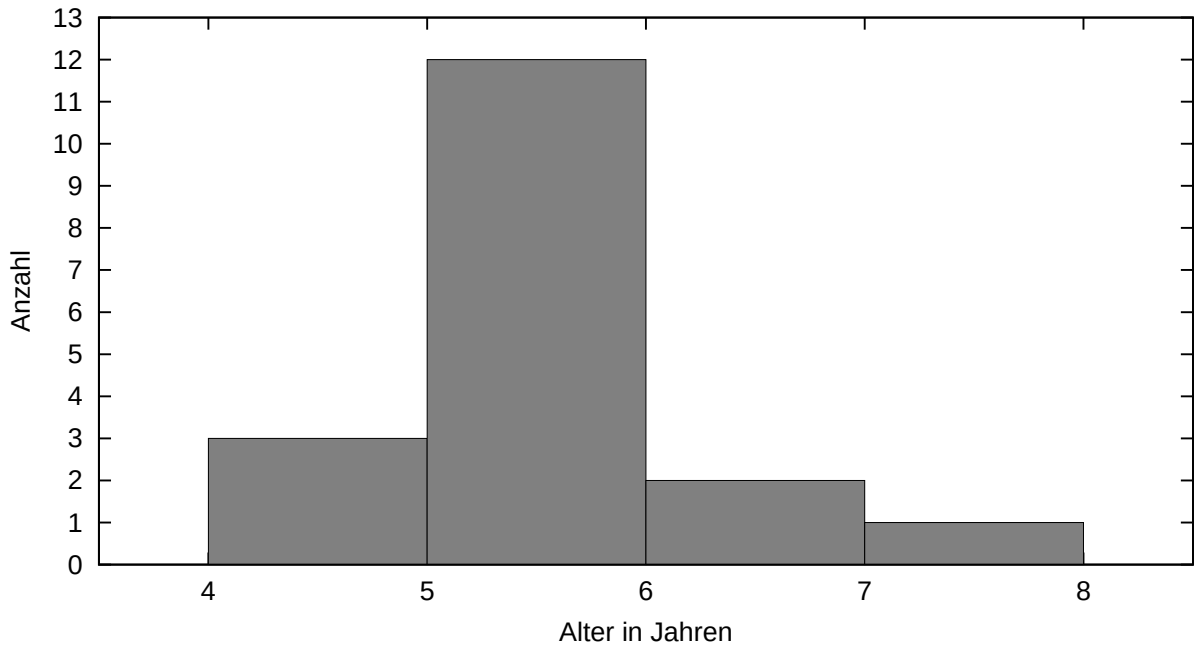


Abbildung 1: Altersverteilung in einem Histogramm

entsprechen. Die Höhen der Rechtecke ergeben sich demnach aus den Quotienten aus den Häufigkeiten und den Breiten der Klassen. Die Gesamtfläche aller Rechtecke eines Histogramms ist gleich 1 [4]. Zur Optimierung von Histogrammen ist ein Kompromiß zu finden zwischen zu gringem Detailreichtum durch zu breite Klassen und dem Überdecken interessanter Entwicklungen und Strukturen durch zufällige Schwankungen bei zu schmalen Klassen. Die Klassen müssen nicht, sollten aber zugunsten der Übersichtlichkeit gleich breit und durch möglichst runde Zahlen begrenzt sein. [2, 3]

# 3 Visualisierung von Daten durch Diagramme

## 3.1 Vergleich zweier Bundestagswahlen in einer Häufigkeitstabelle

| | Zweitstimmen | | % der Wahlberechtigten | |
|---|---|---|---|---|
| | 2005 | 2002 | 2005 | 2002 |
| Wahlberechtigte | 61870711 | 61432868 | 100 | 100 |
| Nichtwähler | 13826577 | 12850107 | 22,3 | 20,9 |
| Wähler | 48044134 | 48582761 | | |
| Ungültige | 756146 | 586281 | | |
| Gültige | 47287988 | 47996480 | | |
| SPD | 16194665 | 18488668 | 26,2 | 30,1 |
| CDU | 13136740 | 14167561 | 21,2 | 23,1 |
| CSU | 3494309 | 4315080 | 5,6 | 7,0 |
| Grüne | 3838326 | 4110355 | 6,2 | 6,7 |
| FDP | 4648144 | 3538815 | 7,5 | 5,8 |
| Die Linke | 4118194 | 1916702 | 6,7 | 3,1 |
| NPD | 748568 | 215232 | 1,2 | 0,4 |

Tabelle 2: Tabelle der Zweitstimmenergebnisse

Man kann nach einer Wahl in eine Tabelle schreiben, wieviele Stimmen die verschiedenen Parteien bei der aktuellen und der vorherigen Wahl erhalten haben. Zusätzlich kann man beispielsweise errechnen, wieviel Prozent der Wahlberechtigten eine bestimmte Partei gewählt haben. Die Tabelle zeigt dann die Häufigkeitsverteilung des untersuchten Merkmals Zweitstimme. Tabelle 2 auf der vorherigen Seite zeigt dies für die Bundestagswahlen der Jahre 2002 und 2005. Solche Tabellen soll man auch als Wertetabellen einer Funktion f(Merkmalsausprägung) = relative Häufigkeit der Merkmalsausprägung auffassen können [2], obwohl zwischen den Parteien und ihren Stimmenanteilen eigentlich kein mathematischer Zusammenhang besteht.

## 3.2 Vergleich zweier Bundestagswahlen in einem Säulendiagramm

Leichter als in einer Tabelle erfassbar werden die Unterschiede zwischen den bekanntesten Parteien und besonders die absoluten Stimmengewinne und -verluste gegenüber der vorherigen Wahl durch eine Darstellung der Zahlen in Form eines Säulendiagrammes.

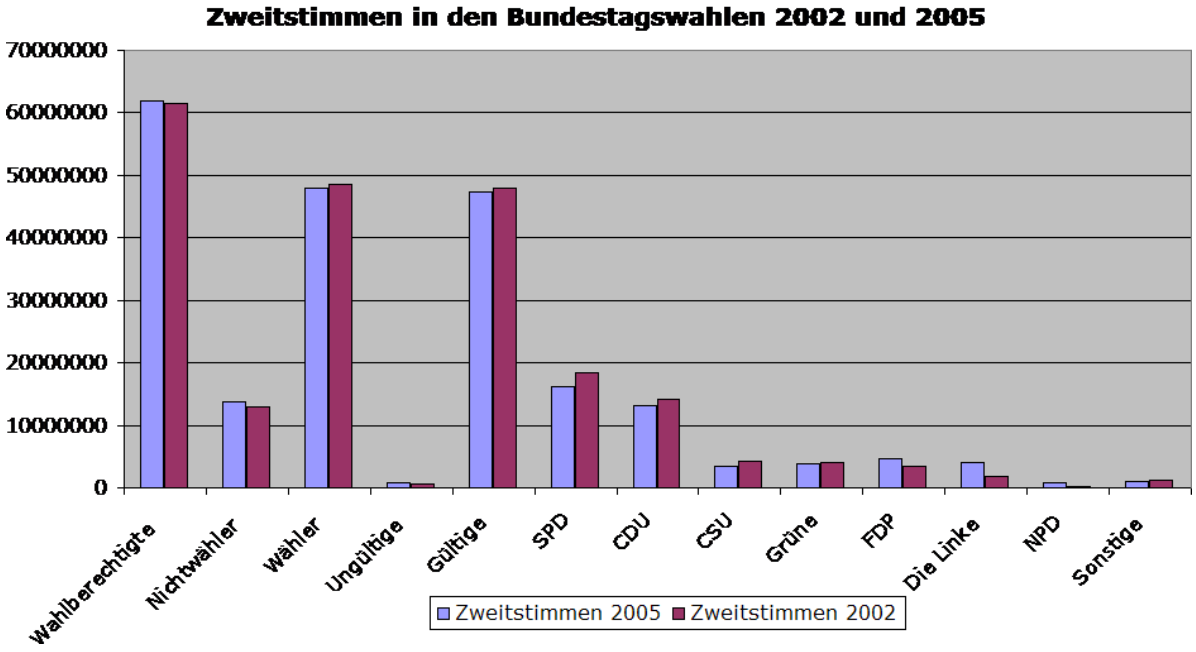


Abbildung 2: Gruppen-Säulendiagramm mit absoluten Stimmenzahlen

Die Abbildung 2 zeigt das zunächst einmal nur für die Häufigkeitsverteilung der absoluten Stimmenzahlen und ignoriert dabei fast alle an der 5% Hürde gescheiterten Parteien, weil das Balkendiagramm sonst zu unübersichtlich geworden wäre.

Entscheidend sind allerdings nach Wahlen nicht die absoluten Wählerstimmenzahlen, sondern die Anteile der Parteien an der Gesamtstimmenzahl. Waren die Wahlbeteiligung und die Zahl der ungültigen Stimmen bei zwei Wahlen sehr unterschiedlich, dann lassen sich Gewinne und Verluste an Stimmenanteilen in einer Tabelle kaum und auch im Balkendiagramm nur schwer erkennen, sofern dieses die absoluten Stimmzahlen anzeigt. Um in einem Balkendiagramm im Vergleich zweier Wahlen die Gewinne und Verluste hinsichtlich der Stimmenanteile gut erkennbar zu machen, kann man die absoluten Stimmenzahlen durch die jeweilige Gesamtzahl der gültigen abgegebenen Stimmen teilen und statt der absoluten Stimmenzahlen die Quotienten anzeigen.

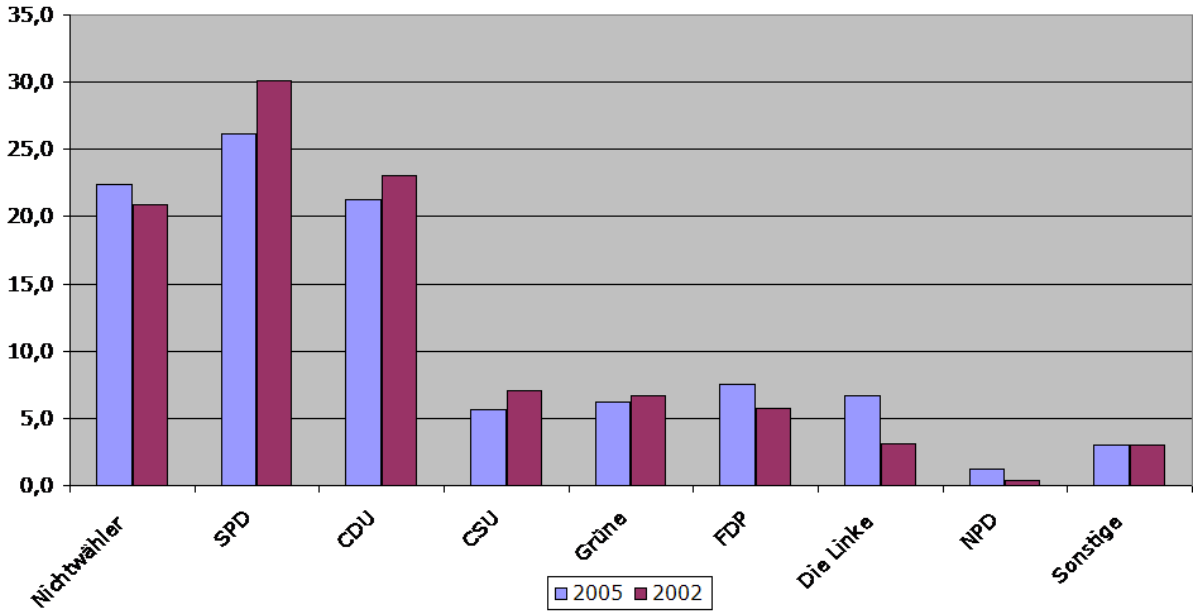


Abbildung 3: Säulendiagramm mit prozentualen Stimmenanteilen bezogen auf alle Wahlberechtigten

Um aber auch den großen Anteil der Nichtwähler zu berücksichtigen, zeigt Abbildung 3 die in den beiden letzten Spalten der Tabelle 2 auf Seite 6 stehenden prozentualen Anteile der Nichtwähler und Wähler der bekanntesten Parteien bezogen auf die Gesamtzahl der Wahlberechtigten. Die Anteile ungültiger Stimmen sowie der nicht gezeigten Parteien sind unter Sonstige zusammengefasst, um insgesamt tatsächlich auf 100% der Wahlberechtigten zu kommen.

## 3.3 Vergleich zweier Bundestagswahlen in einem Blockdiagramm


Abbildung 4: Blockdiagramm mit prozentualen Stimmenanteilen bezogen auf alle Wahlberechtigten

Im Gegensatz zum Säulendiagramm in Abbildung 3 zeigt das Blockdiagramm in Abbildung 4 nicht für jede Partei einen eigenen Balken. Stattdessen liegt hier für jede Wahl ein auf 100% normierter Balken, in dem die prozentualen Stimmenanteile der Parteien und Nichtwähler als farbige Abschnitte erscheinen. Dies macht die Verschiebungen von einer Wahl zur nächsten besser erkennbar. Sortiert man die Abschnitte im Balken nach linken und rechten Parteien und ordnet ihnen charakteristische Farben zu, dann sieht man intuitiv links die Verluste der SPD an die neue Linke und an die Nichtwähler sowie rechts im bürgerlichen Lager die Wählerwanderung von CDU und CSU zu deren erklärtem

Wunschkoalitionspartner FDP. Im Blockdiagramm wird auch deutlich, daß das bürgerliche Lager aus Union und FDP insgesamt deutlich weniger als die rotgrüne Koalition verlor. Augenfällig ist im Blockdiagramm außerdem, wie es für die etablierten Parteien neben Nichtwählern, NPD und der ehemaligen PDS immer enger wird. Kaum erkennbar ist aber die Verhinderung einer Mehrheit von Union und FDP durch die auf Kosten der SPD erstarkte Linke.

### 3.4 Vergleich zweier Bundestagswahlen in Kreisdiagrammen

Prozentuale Anteile werden gerne mit den oft auch Tortendiagrammen genannten Kreisdiagrammen veranschaulicht, weil der Kreis für ein Ganzes oder 100% steht. Anhand des etwas stumpferen Winkels ist hier deutlich erkennbar, daß Grüne, SPD und Linke auch 2005 zusammen mehr Stimmen erhielten als FDP, CDU und CSU. Die absoluten Stimmenzahlen sind aber in Kreisdiagrammen gar nicht darstellbar und man benötigt für den Vergleich zweier Wahlen zwei Kreise. Feine Unterschiede zwischen zwei Wahlen lassen sich daher in Kreisdiagrammen nicht so gut wie in Blockdiagrammen erkennen, aber dafür sehen sie recht nett aus. Abbildung 5 zeigt dies mit den selben Daten wie Abbildung 4 auf der vorherigen Seite.

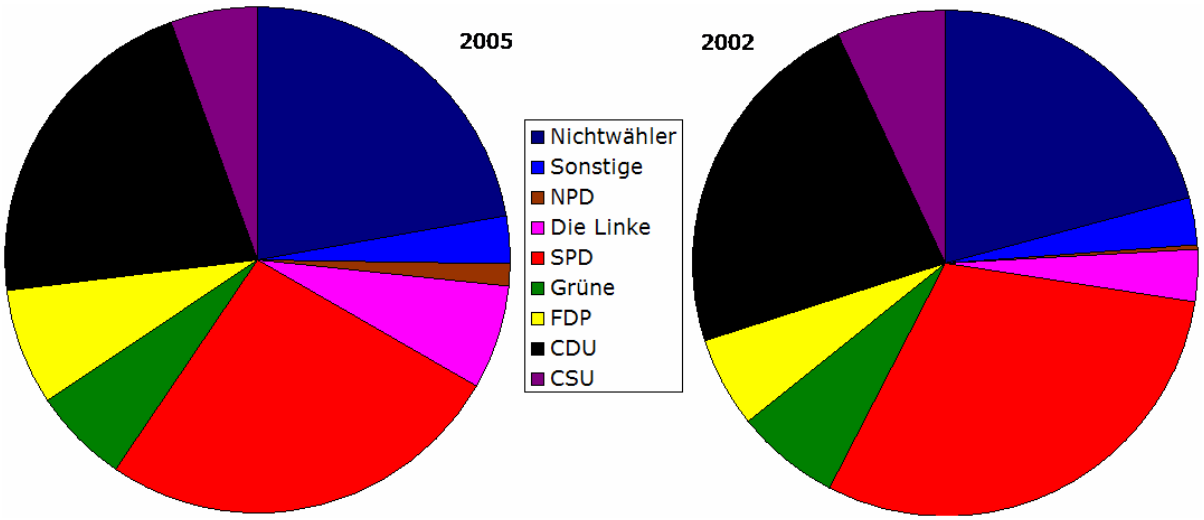


Abbildung 5: Kreisdiagramme mit prozentualen Stimmenanteilen bezogen auf alle Wahlberechtigten

## 4 Lagemaße und Streumaße charakterisieren Häufigkeitsverteilungen

Mit qualitativen Merkmalen kann man normalerweise nicht rechnen, sondern nur die Häufigkeiten verschiedener Ausprägungen darstellen. Ist aber ein Merkmal quantitativ, dann lässt sich die Verteilung seiner Ausprägungen durch Maßzahlen für ihre Lage bzw. das Zentrum der Verteilung sowie für die Streuung um das Zentrum charakterisieren [1, 3]. Kenngrößen der Lage beschreiben Häufigkeitsverteilungen durch Angabe „mittlerer Werte“ [1, 3]. Von solchen Mittelwerten gibt es verschiedene [1, 3], die im folgenden näher beschrieben werden sollen.

# 5 Mittelwerte

Mittelwerte fassen Häufigkeitsverteilungen in einer einzigen Zahl zusammen und machen diese damit handlich und vergleichbar [2, S.112]. Ohne Mittelwerte könnten wir kaum sinnvoll die durchschnittlichen Einkommen in Ost- und Westdeutschland oder die Lebenserwartungen und Längen von Männern und Frauen vergleichen. Dabei ist klar, daß Mittelwerte nichts über den Einzelfall aussagen, der davon stark abweichen kann [2, S.112]. Der Begriff Mittelwert (auch Zentralwert bzw. zentrale Tendenz oder Durchschnittswert genannt) ist aber ein ziemlich unscharfer Begriff, unter dem man recht unterschiedliche Dinge verstehen kann [6, 7]. Die wichtigsten sind das arithmetische Mittel, das geometrische Mittel, das quadratische Mittel, das harmonische Mittel sowie der Median [6]. Normalerweise benutzt man für intervall- und verhältnisskalierte Daten das arithmetische Mittel, für ordinalskalierte Daten den Median und für nominalskalierte Daten den Modalwert [6, 7].

## 5.1 Modalwert

Modalwert nennt man den häufigsten Wert einer Häufigkeitsverteilung [1, 2, 4], sofern es sich nicht um klassierte Werte handelt [4, S.24]. Möglichst sollte außerdem die Verteilung eingipfelig sein und in der Umgebung des Modalwertes sollte eine erkennbare Konzentration der Merkmalswerte vorliegen [4, S.24].

## 5.2 Median

Zur Bestimmung des Median ordnet man eine Zahlenreihe (Urliste) zunächst der Größe nach [1, 2, 4, 6, 7], wobei der Nachfolger nicht unbedingt größer sein muß, sondern auch gleich groß sein darf [2, 4].

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

Dann liegt der (auch Zentralwert, 50%-Quantil oder Medianwert genannte [1, 6, 7]) Median genau in der Mitte dieser Reihe und teilt die Gesamtzahl der Fälle in zwei Hälften [1, 2, 4, 6, 7]. Ist die Anzahl der Merkmalsausprägungen gerade, dann liegt der Median genau zwischen den beiden mittleren Werten [2, 4, 7].

$$\tilde{x} := \begin{cases} x_{\frac{n+1}{2}} & \text{für ungerade n} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n+1}{2}}\right) & \text{für gerade n} \end{cases} \tag{1}$$

Für jeweils mindestens 50% aller Merkmalswerte gilt, daß der Median nicht kleiner oder nicht größer als sie ist [2, 4]. Wenn er nicht mit ihnen identisch ist, dann liegt der Median immer zwischen dem arithmetischen Mittel und dem Modalwert [7]. Er wird weniger empfindlich als das arithmetische Mittel von Ausreißern beeinflußt [2, 119]. Der Median ist außerdem der Wert einer Zahlenreihe, der die kleinste Summe liefert, wenn man ihn von allen übrigen Werte abzieht und die Differenzen addiert [7].

## 5.3 arithmetisches Mittel

Mittelwert im Sinne des arithmetischen Mittels nennt man die durch die Gesamtzahl aller Werte geteilte Summe aller Werte einer Verteilung (Zahlenreihe) [1, 6]. Verfügt man nur über klassierte Daten, rechnet man einfach mit den arithmetischen Mittelwerten zwischen den Unter- und Obergrenzen der Klasen [2, S.113]. Nehmen wir an, ein Experiment, eine Beobachtung oder eine Befragung hinsichtlich des quantitativen Merkmals x hätte zu einer Liste von Merkmalsausprägungen $x_1, x_2, \ldots, x_n$ geführt, deren Mitte nun ermittelt werden soll. Dann ist das arithmetische Mittel $\overline{x}$ folgendermaßen definiert [1, 3, 4]:

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

Bei diskreten Merkmalen sind die Werte $x_1, x_2, \ldots, x_n$ meistens nicht alle unterschiedlich, sondern man sieht in den Häufigkeitstabellen mehrfach vorkommende Merkmalsausprägungen $a_1$ bis $a_k$ mit den absoluten Häufigkeiten $n_1$ bis $n_k$ und den relativen Häufigkeiten $h_1$ bis $h_k$. Daher können alle identischen Merkmalsausprägungen zusammengefasst werden [1, 2, 3].

$$\overline{x} = \frac{1}{n}(n_1a_1 + \ldots + n_ka_k) \Longleftrightarrow \overline{x} = \frac{n_1}{n}a_1 + \ldots + \frac{n_k}{n}a_k \Longleftrightarrow \overline{x} = h_1a_1 + \ldots + h_ka_k \tag{3}$$

Man nennt dies auch gewichtetes arithmetisches Mittel der Ausprägungen [2, S.114]. Der gewichtete arithmetische Mittelwert lässt sich kurz folgendermaßen ausdrücken [2, 3, 4]:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_ia_i = \sum_{i=1}^{k} h_ia_i \tag{4}$$

Aus der Definition des arithmetischen Mittels ergibt sich außerdem [4, S.27]:

$$\sum_{i=1}^{n} x_i - \overline{x} = 0 \tag{5}$$

Die Koordinaten des Schwerpunktes eines Körpers entsprechen den arithmetischen Mittelwerten der x-, y- und z-Koordinaten. Bei einer Fläche ergeben sich die Koordinaten des Schwerpunktes aus den arithmetischen Mittelwerten der x- und y-Koordinaten. Analog kann man den arithmetischen Mittelwert eindimensionaler Daten als deren Schwerpunkt betrachten [2, S.115]. Dies erklärt auch, daß der Einfluß eines Wertes auf den arithmetischen Mittelwert von seinem Abstand vom Mittelwert und von seiner Häufigkeit abhängt [2, S.115].

## 5.4 geometrisches Mittel

Wächst ein Baum jedes Jahr um eine bestimmte Anzahl von Zentimetern, dann kann man beispielsweise die Zuwachsbeträge der letzten 5 Jahre addieren und durch 5 dividieren, um das arithmetische Mittel des jährlichen Wachstums in Zentimetern zu ermitteln. Sind aber hinsichtlich des jährlichen Wachstums von Bäumen, der Wirtschaft oder eines Schuldenberges nur bestimmte Prozentsätze bekannt, dann liefert das arithmetische Mittel der jährlichen Wachstumsraten keinen sinnvollen Mittelwert. Das liegt daran, daß z.B. 10% Wachstum bei einem 10 Meter hohen Baum 1 Meter ergeben, während eine Verdopplung

der Wachstumsrate auf 20% im nächten Jahr nicht 2 Metern, sondern einem Längenzuwachs von 2,20 Metern entspricht. Aus diesem Grund kann man Wachstumsraten nicht einfach addieren, sondern muß sie immer mit dem Produkt der voran gegangenen Multiplikation multiplizieren. Nennen wir A die Ausgangsgröße und $W_1, W_2$ die Wachstumsraten zweier Jahre, dann ist die Endgröße E nach 1 Jahr $E_1 = A + A \cdot W_1$ und nach 2 Jahren $E_2 = A + A \cdot W_1 + (A + A \cdot W_1) \cdot W_2$. Durch Ausklammern erhält man ein endlich wieder übersichtliches Produkt:

$$E_2 = A(1 + W_1) + A(1 + W_1) \cdot W_2 = A(1 + W_1)(1 + 1 \cdot W_2) = A(1 + W_1)(1 + W_2)$$

Man muß also die Ausgangsgröße A mit den jeweils um 1 vergrößerten jährlichen Wachstumsraten multiplizieren, um die Endgröße zu erhalten. Man nennt eine um 1 vergrößerte Wachstumsrate auch Wachstumsfaktor [4, S.27]. Der für Produkte geeignete Mittelwert ist aber nicht das arithmetische, sondern das sogenannte geometrische Mittel [4, S.26]. Das geometrische Mittel $x_g$ für n positive Werte $x_1, x_2, \ldots, x_n$ ist wie folgt definiert [2, 4]:

$$x_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \tag{6}$$

In unserem einfachen Beispiel mit dem Baum erhielten wir den über 2 Jahre gemittelten Wachstumsfaktor durch Einsetzen in diese Formel:

$$x_g = \sqrt[2]{(1 + W_1)\% \cdot (1 + W_2)\%} = \sqrt{110\% \cdot 120\%} = \sqrt{13200\%^2} = 114,89\%$$

Multiplizieren wir die Ausgangshöhe von 10 Metern mit diesem durchschnittlichen Wachstumsfaktor, dann erhalten wir die korrekte Endhöhe von:

$$10m \cdot 114,89\% \cdot 114,89\% = 10m \cdot 1,1489 \cdot 1,1489 = 10m \cdot 1,32 = 13,20m$$

## 5.5 harmonisches Mittel

Ähnlich wie das geometrische Mittel wurde auch das geometrische Mittel nicht einfach erfunden, sondern es ergibt sich aus einem speziellen Aufgabentyp. Fährt ein Auto auf zwei gleich langen Streckenabschnitten $s_1$ und $s_2$ mit unterschiedlichen Geschwindgikeiten $v_1$ und $v_2$, dann benötigt es für die beiden gleich langen Streckenabschnitte unterschiedlich lange Zeiten $t_1$ und $t_2$. Nennen wir zusätzlich die Gesamtstrecke $s$, dann läßt sich die durchschnittliche Geschwindigkeit $\overline{v}_h$ für die gesamte Strecke folgendermaßen berechnen [2, S.125]:

$$\overline{v}_h = \frac{s_1 + s_2}{t_1 + t_2} = \frac{s}{\frac{s_1}{v_1} + \frac{s_2}{v_2}} = \frac{s}{\frac{\frac{1}{2}s}{v_1} + \frac{\frac{1}{2}s}{v_2}} = \frac{s}{\frac{1}{2}s\left(\frac{1}{v_1} + \frac{1}{v_2}\right)} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}}$$

Verallgemeinert man nun diese Berechnung des sogenannten harmonischen Mittelwertes von zwei auf n positive Zahlen $x_1, x_2, \ldots, x_n$, dann erhält man die allgemeine Formel für das harmonische Mittel [2, S.125]:

$$\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} \tag{7}$$

# 6 Streuungsmaße

Streuungsmaße charakterisieren die Verteilung von Werten um einen Mittelwert herum [2, S.126].

## 6.1 Spannweite

Spannweite nennt man die Differenz zwischen dem größten und dem kleinsten Wert einer Datenreihe [2, S.126].

## 6.2 Mittlere lineare Abweichung

Als Streuungsmaß meistens besser als die nur zwei Extremwerte berücksichtigende Spannweite eignet sich zur Charakterisierung der Verteilung von Werten um einen arithmetischen Mittelwert die sogenannte mittlere lineare Abweichung [2, S.126]. Besonders gut passt sie zum Median, weil die mittlere lineare Abweichung den kleinstmöglichen Wert annimmt, wenn man als Mittelwert den Median einsetzt [2, S.131]. Das klingt komplizierter als es ist, denn man ermittelt einfach den Median, addiert die absoluten Beträge der Differenzen zwischen dem Median und jedem einzelnen Wert und teilt die Summe durch die Anzahl der Werte. Übersichtlicher lässt sich dies mathematisch für die Werte $x_1, x_2 \ldots x_n$ darstellen, wenn $\tilde{x}$ der Median und d die mittlere lineare Abweichung von diesem ist [2, S.126]:

$$d = \frac{1}{n}\left(|x_1 - \tilde{x}| + |x_2 - \tilde{x}| + \ldots + |x_n - \tilde{x}|\right) \qquad \text{oder kurz} \qquad d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}| \quad (8)$$

Komplizierter wird die Sache, wenn die Merkmalswerte nicht nur jeweils einmal oder wenigstens alle gleich oft vorkommen und deshalb alle mit einem Faktor h für die relative Häufigkeit versehen sind. Dann werden die Merkmalsausprägungen $x_1, x_2 \ldots x_m$ mit den jeweiligen relativen Häufigkeiten $h_1, h_2 \ldots h_m$ verknüpft und man berechnet z.B. das arithmetische Mittel $\overline{x}$ sowie die mittlere lineare Abweichung d folgendermaßen [2, S.127]:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot h(x_i) \qquad d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| \cdot h(x_i) \quad (9)$$

## 6.3 Mittlere quadratische Abweichung

Quasi eine Weiterentwicklung der mittleren linearen Abweichung ist die sogenannte mittlere quadratische Abweichung [2, S.127]. Sie unterscheidet sich von der Vorgängerin nur dadurch, daß hier die Differenzen zwischen den Werten und dem Mittelwert quadriert werden. So bekommt man automatisch nur positive Werte und der Einfluß der Entfernung einzelner Daten vom Mittelwert wird durch das Quarieren noch größer [2, S.127]. Vor allem soll aber die mittlere quadratische Abweichung aus nicht genannten „innermathematischen“ Gründen gegenüber der mittleren linearen Abweichung vom arithmetischen Mittel bevorzugt werden [2, S.127]. Sie paßt optimal zum arithmetischen Mittel, weil dieses bei gegebenen Werten die kleinst mögliche mittlere quadratische Abweichung liefert. [2, S.131] Mathematisch sieht das in der Kurzform unter Berücksichtigung relativer Häufigkeiten so aus [2, S.127]:

$$\overline{x}^2 = \sum_{i=1}^{m} x_i \cdot h(x_i) \qquad \overline{s}^2 = \sum_{i=1}^{m}(x_i - \overline{x})^2 \cdot h(x_i) \quad (10)$$

Man kann die Berechnung der mittleren quadratischen Abweichung noch vereinfachen [2, S.128]:

$$\overline{s}^2 = \left(x_1^2 \cdot h(x_1) + x_2^2 \cdot h(x_2) + \ldots + x_m^2 \cdot h(x_m)\right) - \overline{x}^2 \quad (11)$$

# Quellenverzeichnis


[1] 4, 5.1, 5.2, 5.3, 5.3
Bibliographisches Institut & F. A. Brockhaus AG: *Der Brockhaus multimedial 2004 premium*. ISBN 3-411-06673-3

[2] 1.1, 1.1.1, 1.3, 1.6, 1.7, 1.8, 2.1, 2.3, 2.4, 2.6, 2.7, 3.1, 5, 5.1, 5.2, 5.2, 5.3, 5.3, 5.3, 5.3, 5.4, 5.5, 6, 6.1, 6.2, 6.2, 6.3, 6.3
Günter Cöster ; Heinz Griesel ; Arnold Hermans ; Horst Jahner ; Andreas Meißner ; Angelika Müller ; Heinz Klaus Strick ; Frierich Suhr ; Rudolf vom Hofe ; Helmut Postel ; Lohar Profke ; Ferdinand Weber: *Elemente der Mathematik 11*. Schroedel Verlag GmbH, Hannover 1999

[3] 1.1, 1.1.1, 1.1.2, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 2.1, 2.7, 4, 5.3, 5.3, 5.3
Dr. Andreas Handl: *Einführung in die Statistik mit R*. http://www.wiwi.uni-bielefeld.de/~frohn/Lehre/Statistik1/Skript/stat12b.pdf

[4] 1.1, 1.1.1, 1.1.2, 1.2, 1.3, 1.5, 1.6, 1.7, 1.8, 2.1, 2.3, 2.4, 2.6, 2.7, 5.1, 5.2, 5.2, 5.3, 5.3, 5.3, 5.4
Dr. Sabine Lauer: *Grundlagen der Statistik*.
http://www.vanille.de/lehre/skript.pdf, 25.1.2006

[5] 1.6
Prof. Dr. Wolfgang Ludwig-Mayerhofer: *ILMES - Internet-Lexikon der Methoden der empirischen Sozialforschung*. http://www.lrz-muenchen.de/~wlm/ein_voll.htm, 2006

[6] 1.1, 1.2, 1.3, 1.4, 1.6, 1.7, 2.1, 2.3, 5, 5.2, 5.3
Verbund Norddeutscher Universitäten: *Methodenlehre-Baukasten, Ein interaktives Lehr-Lernprogramm zur Statistik*. http://www.methodenlehre-baukasten.de, 2005

[7] 1.6, 1.7, 5, 5.2, 5.2
Zentrum für Hochschul- und Weiterbildung (bis 10/2005 Interdisziplinäres Zentrum für Hochschuldidaktik): *LernSTATS, ein interaktives Programm zum Lehren und Lernen der Statistik in der Psychologie, in den Sozial- und Erziehungswissenschaften*. http://www.lernstats.de, 1999