Tabellarische und Graphische Darstellung des Materials

Die *Darstellung* der Daten ist (neben deren Zusammenfassung in den im nächsten Kapitel behandelten statistischen *Kennwerten*) Aufgabe der beschreibenden Statistik. In vielen Handbüchern der Statistik findet man hierzu noch Anweisungen zur händischen Aufbereitung der Daten (Strichlisten etc.), obwohl derartige Aufgaben mittlerweile schon längst von Statistikprogrammen übernommen worden sind. Aus diesem Grunde wird hier als Einstieg zunächst eine Darstellungsform gewählt, die zwar weniger zur Präsentation der Daten geeignet ist, dafür aber den Übergang zu einer späteren EDV-gerechten Erfassung erleichtert.

Daten werden von Statistikprogrammen in der Regel in Form einer *rechteckigen Datenmatrix* eingelesen. Darunter versteht man eine Anordnung der Daten in n Fälle mit je m Variablen. Betrachten wir zur Verdeutlichung nachstehendes Beispiel: Bei einer Erhebung wurden bei 30 Probanden die folgenden Eigenschaften erfasst: „Raucher Ja/Nein“, die Reaktionszeit in ms vor Verabreichung eines Medikaments und die Reaktionszeit nach Verabreichung eines Medikaments. Die drei ausgewählten Eigenschaften werden auch als *Variablen* bezeichnet, wobei jedem der Probanden drei Variablen zuzuordnen sind. Die möglichen Messergebnisse pro Variable werden als deren *Ausprägungen* bezeichnet. Bei der ersten Variablen lassen sich nur zwei Ausprägungen feststellen, nämlich die Antwortmöglichkeiten „JA“ bzw. „Nein“, die durch die Zahlen „1“ und „2“ dargestellt werden. Da keine Zwischenwerte möglich sind, handelt es sich hier um eine *diskrete* Variable. Die letzten beiden Variablen lassen sich dahingegen beliebig fein abstufen (freilich immer im Rahmen der Messgenauigkeit). Mögliche Reaktionszeiten sind etwa: 33,45 ms oder auch 33,45678 ms. Da sich die Variablen beliebig fein abstufen lassen, bilden ihre Ausprägungen ein Kontinuum. Wir sprechen in diesem Falle von *stetigen* Variablen.

Die möglichen Ausprägungen einer Variablen schließen sich bei einem Probanden gegenseitig aus. So kann beispielsweise die Frage 'Rauchen Sie' nicht gleichzeitig mit „Ja“ bzw. „Nein“ beantwortet werden. Gleichermaßen ist bei Messung einer Reaktionszeit immer nur ein bestimmter Wert möglich. Bei drei Variablen erhalten wir daher drei Messwerte pro Untersuchungsperson, beispielsweise:

1 15 30

Diese drei Messwerte bilden eine Datenzeile, deren Werte wir wie folgt interpretieren: Der Proband hat die Frage 'Rauchen Sie' mit „Ja“ beantwortet (wenn Ja mit 1 codiert wurde), hat bei der ersten Messung der Reaktionszeit einen Wert von 15 ms erhalten und bei der zweiten Messung einen Wert von 30 ms.

Schreibt man die Messwerte aller Probanden zeilenweise untereinander, so erhält man ein rechteckiges Gebilde, das etwa folgendermaßen ausschaut:

1 15 30

2 20 45

1 17 33

2 35 42

usw.

Damit das Programm die Daten später richtig einlesen kann, ist es wichtig, eine einmal getroffene Anordnung der Variablen in jeder Zeile beizubehalten. Fehlt bei einem Probanden die Angabe zu einer Variablen, so ist der dafür vorgesehene Spaltenplatz freizulassen. Alle Ausprägungen einer Variablen stehen somit untereinander. Auf diese Weise entsteht ein rechteckiges Gebilde - eben die bereits erwähnte rechteckige Datenmatrix mit n Fällen und m Variablen. Diese Datenmatrix dient, wie bereits gesagt, zur EDV-gerechten Aufbereitung der Daten. Zur tatsächlichen *Präsentation* der Daten lassen sich grob zwei Darstellungsweisen unterscheiden: 1) tabellarische Darstellung 2) graphische Darstellung.

Zur tabellarischen Darstellung des Materials

Die einfachste Darstellungsform ist die Angabe der Häufigkeiten, mit denen die einzelnen Ausprägungen einer Variablen vorkommen. Bei der *diskreten* Variable „Rauchen Sie?“ kann beispielsweise ausgezählt werden, wie viele von den insgesamt 30 erhobenen Personen Nichtraucher bzw. Raucher sind. Weiters lässt sich der prozentuelle Anteil der Nichtraucher bzw. Raucher in einer Erhebung feststellen. Betrachten wir hierzu das folgende Beispiel:

Rauchen Sie?

Value Label	Value	Frequency	Percent
JA	1	14	46.7
Nein	2	16	53.3
		-------	-------
	Total	30	100.0

Erläuterungen: In dieser von SPSS erzeugten Häufigkeitstabelle befinden sich in der Spalte unter ´Value Label´ die als Text angegebenen Ausprägungen der Variable ´Rauchen Sie´ , unter ´Value´ deren numerische Verschlüsselungen, unter ´Frequency´ die Häufigkeiten und unter ´Percent´ steht der jeweilige prozentuelle Anteil.

Weitere Darstellungsmöglichkeiten von Häufigkeitstabellen erkennt man am besten am Beispiel der folgenden, zusätzlich eingeführten Variable ´Fakultätszugehörigkeit´. Letztere Variable hat mehr als nur zwei Ausprägungen, nämlich (nachstehende Fakultäten sind freilich nur eine Auswahl) die Fakultäten ´Theologie´, ´Jus´, ´Medizin´, ´Geisteswissenschaften´ und ´Naturwissenschaften´, kodiert - in der Reihenfolge ihrer Aufzählung - mit den Zahlen von 1 bis 5 (´1´ bedeute ´Theologie´ usw.). Der Einfachheit halber gehen wir davon aus, dass keine Doppelinskriptionen vorkommen, dass also jede Person nur jeweils *einer* Fakultät angehören kann. Für diese Variable können wir beispielsweise folgende Häufigkeitstabelle erstellen:

```
Fakultaet


Value Label                Value  Frequency  Percent   Cum Percent

Theol                          1          5     16.7      16.7

Jurid.                         2          5     16.7      33.3

Medizin                        3          6     20.0      53.3

Geiwi                          4          7     23.3      76.7

Natwi                          5          7     23.3     100.0

                                    -------  -------

                           Total        30    100.0
```

Erläuterungen: Die Spalten ´Value Label´, ´Value´, ´Frequency´ und ´Percent´ entsprechen in ihrer Bedeutung dem oben angeführten Beispiel. Neu ist hier hingegen die Angabe unter ´Cum Percent´. Gemeint sind hiermit die *kumulierten* Prozentangaben (Beispiel: 53.3 % umfasst den prozentuellen Anteil aller Studierenden, die Theologie oder Jus oder Medizin - das ist die Summe von 16.7 + 16.7 + 20.0 - studieren). Derartige kumulierte Prozentangaben verschaffen einen raschen Überblick über Häufigkeitsangaben bei diskreten Variablen mit mehr als zwei Ausprägungen. (Scheinbare Rechenungenauigkeiten sind nur darauf zurückzuführen, dass das Rechenprogramm bei der Auflistung der Ergebnisse Nachkommastellen abschneidet, intern aber mit größerer Genauigkeit rechnet.)

Häufigkeitstabellen, wie sie in den beiden Beispielen dargestellt wurden, lassen sich indes sinnvoll nur bei diskreten Variablen erstellen. Bei *kontinuierlichen* Variablen sind die einzelnen Ausprägungen viel zu zahlreich, um sich in Form von Häufigkeitstabellen darstellen zu lassen - möglicherweise hat jede Versuchsperson eine andere Ausprägung. Verdeutlichen wir uns diesen Umstand am folgenden Beispiel:

Reaktionszeit vorher

Value	Frequency	Percent	Cum Percent
9	1	3.3	3.3
11	1	3.3	6.7
13	1	3.3	10.0
15	3	10.0	20.0
17	2	6.7	26.7
18	1	3.3	30.0
19	1	3.3	33.3
20	1	3.3	36.7
21	3	10.0	46.7
22	2	6.7	53.3
23	1	3.3	56.7
24	3	10.0	66.7
25	2	6.7	73.3
26	1	3.3	76.7
27	1	3.3	80.0
28	1	3.3	83.3
31	2	6.7	90.0
33	1	3.3	93.3
35	1	3.3	96.7
41	1	3.3	100.0
Total	30	100.0	

Wie das Beispiel drastisch belegt, hat nahezu jede Versuchsperson eine andere Reaktionszeit - die meisten Häufigkeitsangaben belaufen sich auf einen Fall. Um bei kontinuierlichen Variablen eine sinnvolle (d.h.überschaubare) Häufigkeitsdarstellung zu erhalten, sind derartige Variablen zunächst in Klassen (Kategorien) einzuteilen. Betrachten wir hierzu zunächst das folgende Beispiel:

Reaktionszeit vorher

Value Label	Value	Frequency	%	Valid %	Cum %
5 - 9,999	1	1	3.3	3.3	3.3
10 - 14,999	2	2	6.7	6.7	10.0
15 - 19,999	3	7	23.3	23.3	33.3
20 - 24,999	4	10	33.3	33.3	66.7
25 - 29,999	5	5	16.7	16.7	83.3
30 - 34,999	6	3	10.0	10.0	93.3
35 - 39,999	7	1	3.3	3.3	96.7
40 - 44,999	8	1	3.3	3.3	100.0
		30	100.0	100.0	

Erläuterungen: In dem angegebenen Beispiel wurde die Reaktionszeit in Fünferschritten zusammengefasst. Dadurch entstehen 8 Klassen von Reaktionszeiten. Die in diesem Beispiel verwendete Schrittweite von 5 ms wird als *Kategorienbreite* bezeichnet. Sie berechnet sich aus dem Abstand zwischen den *Kategoriengrenzen*: 5 - 9,999; 10 - 14,999 usw. Genau gerechnet ergäbe dieser Abstand nur einen Wert von 4,999 und nicht exakt 5 ms. Die *mathematisch exakte Kategorienbreite* beträgt aber 4,999... Da aber unsere Messungen immer nur eine Genauigkeit von wenigen Nachkommstellen haben, genügt es, für eine zweifelsfreie Zuordnung der Reaktionszeiten zu den Kategorien sich auf die innerhalb der Messgenauigkeit gelegenen Kategoriengrenzen zu beschränken. Der unterste Wert, von dem ausgehend wir mit der Kategorieneinteilung beginnen - im angegebenen Beispiel der Wert 5 -, wird als *Reduktionslage* bezeichnet. Sie entspricht der unteren Grenze der ersten Kategorie und kann frei gewählt werden (Bei einer Reduktionslage von 7 bekämen wir beispielsweise die Kategorien von 7 - 11,999; 12 - 16, 999 usw.). In dem oben angeführten Beispiel werden unter ´Value Label´ die Kategoriengrenzen der einzelnen Kategorien angegeben, unter ´Value´ die durchnummerierten Kategorien. Alle anderen Spaltenbeschriftungen entsprechen den beiden weiter oben zu diskreten Variablen angeführten Beispielen.

Allgemein lässt sich zur Kategorienbildung folgendes festhalten: Je kleiner die Kategorienbreite gewählt wird, umso größer ist die Anzahl der Kategorien. Nimmt deren Anzahl zu, so werden die Häufigkeitsangaben pro Kategorie abnehmen und unsere Aufgliederung wird dafür umso detaillierter. Wie aber in der Praxis die Anzahl der Kategorien festzulegen ist, lässt sich nicht pauschal beantworten. So werden bei weniger Messungen natürlich auch weniger Kategorien benötigt. Welche Kategorieneinteilung zu wählen ist, hängt in erster Linie vom Zweck der jeweiligen Fragestellung ab. (Man denke beispielsweise an die Bestimmung von Altersklassen. Ist man daran interessiert, den Anteil der Junioren und Senioren in einem Kollektiv zu ermitteln, so wird sich aus dieser Fragestellung eine ganz bestimmte Klasseneinteilung ergeben.)

Zur graphischen Darstellung des Materials

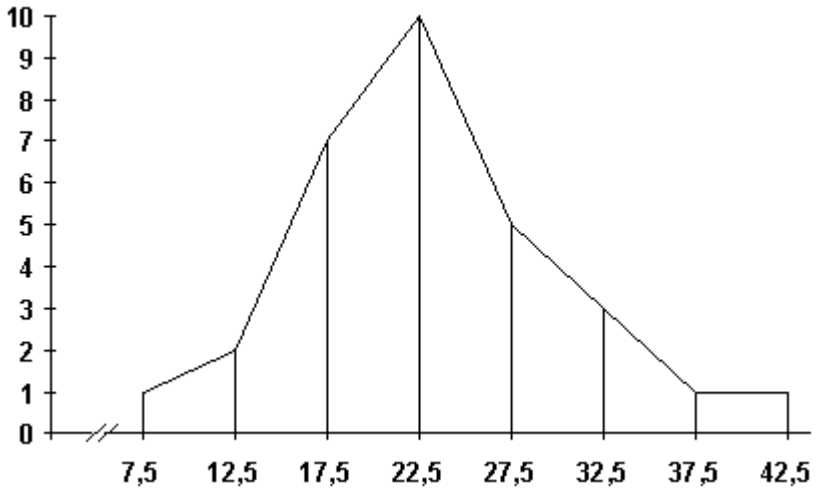
1) Graphische Darstellungen von kontinuierlichen Variablen

Die Häufigkeitsverteilung einer in Kategorien zusammengefassten kontinuierlichen Variablen lässt sich graphisch auf verschiedene Weisen darstellen, nämlich als Polygonzug, als Summenpolygon und als Histogramm.

a) *Polygonzug*

Vorgehensweise: Auf der x-Achse werden die *Kategorienmitten* der verschiedenen Kategorien aufgetragen und auf der y-Achse die Häufigkeiten. Die Kategorienmitten werden aus dem Mittelwert von exakter unterer und exakter oberer Kategoriengrenze pro Kategorie berechnet (Beispiel: (5 + 10)/2 = 7,5). Von den Kategorienmitten zieht man einen senkrechten Strich, dessen Länge der jeweiligen Kategorienhäufigkeit entspricht (so kommt beispielsweise in der Kategorie mit der Mitte 7,5 ein Fall vor, in der anschließenden Kategorie mit der Mitte 12,5 zwei Fälle usw.).

Verdeutlichen wir uns einen Polygonzug am Beispiel der oben für die Kategorien der Variable ´Reaktionszeit vorher´ angegebenen Häufigkeitsverteilung:



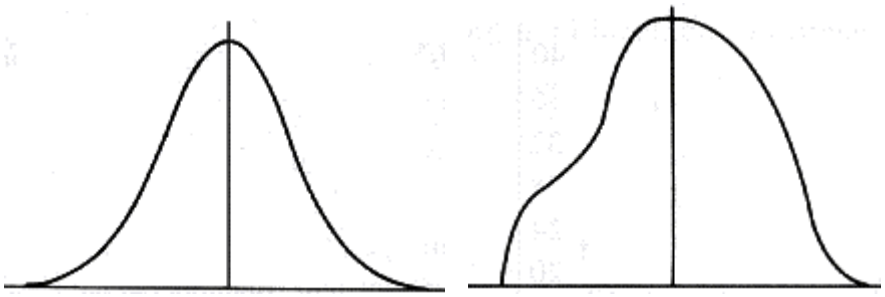
An dieser graphischen Darstellung lässt sich auf anschauliche Weise die *Verteilung* der Reaktionszeiten im untersuchten Kollektiv verdeutlichen. Nehmen wir an, die Berechnung der durchschnittlichen Reaktionszeit in unserem Kollektiv hätte einen Wert von 22,43 ergeben. Wir können an dem Polygonzug nun ablesen, dass in der Nähe dieses Wertes - nämlich im Intervall von 20 bis 24,999 - auch die meisten Werte vorkommen (10 Messungen bei der Kategorienmitte 22,5). Je weiter die Reaktionszeit von diesem durchschnittlichen Wert nach unten bzw. nach oben abweicht, umso geringer wird die Kategorienhäufigkeit. Die durchbrochene Linie auf der x-Achse soll darauf hinweisen, dass die dargestellte Verteilung

der Messwerte nur im Wertebereich der angegebenen Kategorien definiert ist. Es gibt daher auch keinen absoluten Nullpunkt.

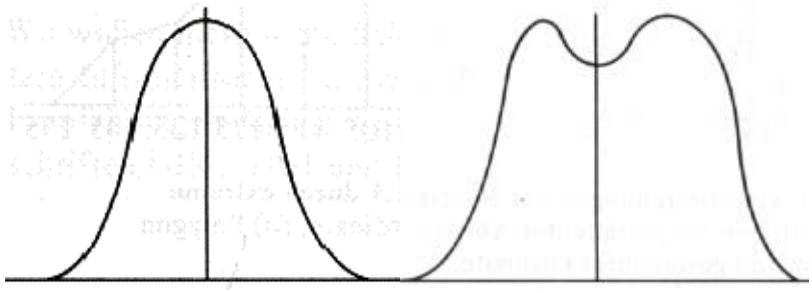
Wir haben diese Darstellungsform gewonnen, indem wir die stetige Variable ´Reaktionszeit vorher´ durch Klassenbildung in eine diskrete Variable umgewandelt haben. Je kleiner diese Klassengrenzen gewählt werden, desto mehr nähert sich der Polygonzug - allerdings nur bei einem entsprechend groß gewählten Kollektiv - einer Kurve. Die Gesamtfläche unter der Kurve ergibt 100 % aller Fälle der jeweiligen Untersuchung.

Die Verteilung einer Variablen lässt sich nun nach bestimmten Eigenschaften klassifizieren. Verteilungen können sein: symmetrisch oder asymmetrisch, eingipflig oder mehrgipflig, breit oder schmal (man bezeichnet dies auch als den ´Exzess´), links- oder rechtssteil.

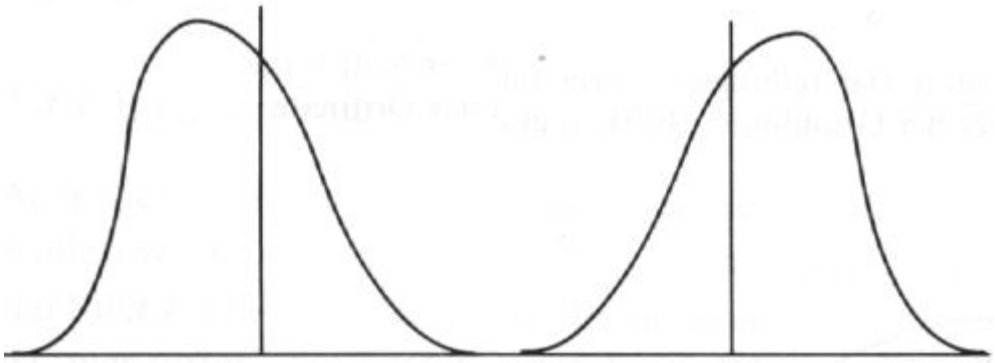
Betrachten wir hierzu die folgenden Beispiele:



symmetrisch asymmetrisch

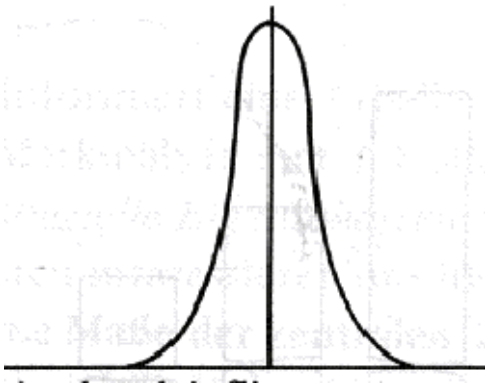


eingipflig mehrgipflig

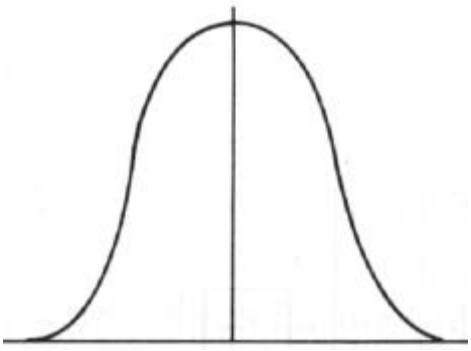


linkssteil

rechtssteil



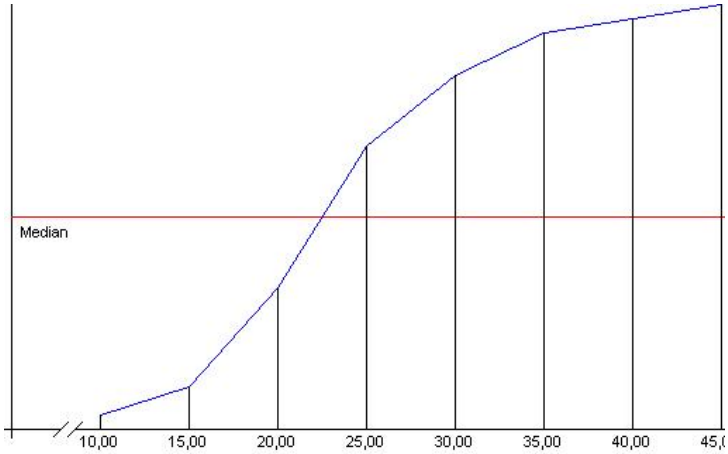
schmalgipflig



breitgipflig

b) *Summenpolygon*

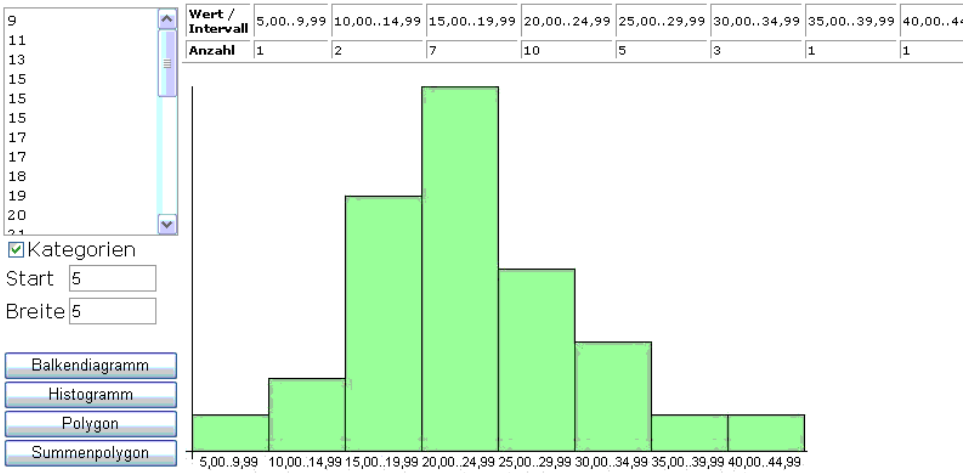
Vorgehensweise: Beim Summenpolygon werden auf der x-Achse die oberen Kategoriengrenzen und auf der y-Achse die kumulierten Prozentwerte aufgetragen:



Aus diesem Diagramm lassen sich nun sehr einfach Median (in der Höhe von 50%) und Quartile (in der Höhe von 25%, 50% und 75%) ablesen (zu diesem Thema mehr im nächsten Kapitel).

c) *Histogramm*

Vorgehensweise: Beim Histogramm werden auf der x-Achse die Kategoriengrenzen und auf der y-Achse die Häufigkeiten aufgetragen. Im Unterschied zum Polygonzug werden die Häufigkeiten der verschiedenen Kategorien durch gleich breite Säulen dargestellt. Betrachten wir hierzu nachstehendes Beispiel:



Wie an diesem Beispiel ersichtlich wird, sind beim Histogramm die einzelnen Säulen *nicht* durch Abstände getrennt. Dies liegt daran, dass bei einer stetigen Variablen die Kategoriengrenzen (gemeint sind die mathematisch exakten Kategoriengrenzen) so gewählt sind, dass sie alle Werte der Verteilung im angegebenen Messbereich (hier von 5 bis 44,999) erfassen.
Die Flächen der aneinander stoßenden Rechtecke sind *proportional* zu den relativen Häufigkeiten. Man nennt Histogramme daher auch flächentreue Darstellungen. Es wird empfohlen, bei Histogrammen eine gleiche Klassenbreite zu wählen. denn nur bei gleichen Klassenbreiten lässt sich an der Höhe der Säulen auch die Häufigkeiten ablesen (da in diesem Falle Größenunterschiede zwischen den verschiedenen Flächen der Rechtecke wegen ihrer gleichen Breite ja nur durch die unterschiedliche Höhe zustande kommen). Unterschiedliche Klassenbreiten werden hier nicht behandelt. Wie im letzteren Falle die Höhen korrigiert werden müssen, um eine flächentreue Darstellung zu erhalten vgl. http://de.wikipedia.org/wiki/Histogramm.

2) Graphische Darstellung von diskreten Variablen (Säulendiagramm)

Da bei einer diskreten Variablen weder Polygon noch Summenpolygon eine sinnvolle Darstellung ergeben, wird hier nur auf die Darstellungsform durch Säulen eingegangen. Im Unterschied zum oben besprochenen Histogramm, das bei in Klassen eingeteilten stetigen Variablen zur Anwendung kommt, bezeichnet man diese Form der Häufigkeitsdarstellung bei diskreten Variablen als *Säulendiagramm*. Die Höhe der Säulen markiert die Häufigkeit der auf der x-Achse aufgetragenen Werte. Die Breite der Säulen kann beliebig gewählt werden. Der Unterschied zum Histogramm besteht darin, dass beim Säulendiagramm die einzelnen Säulen durch einen Abstand voneinander getrennt sind. Diese räumliche Trennung ist erforderlich, um anzuzeigen, dass zwischen den einzelnen auf der x-Achse aufgetragenen Werten bei einer diskreten Variablen keine Zwischenwerte definiert sind. Man betrachte hierzu das folgende Beispiel: