Kapitel 5: Streuung, Schiefe, Wölbung

1. Streuungsbegriff und Eigenschaften von Streuungsmaßen

a) Begriff der Streuung (Dispersion)

Streuungsmaße sind einmal beschreibende Statistiken von Häufigkeitsverteilungen und zum anderen auch bedeutsam für die Beurteilung statistischer Berechnungen. Sie dienen

1. der Charakterisierung der Variabilität eines Merkmals oder, gleichbedeutend, der Ausbreitung einer Häufigkeitsverteilung und der Homogenität einer statistischen Masse, d.h. der Ähnlichkeit ihrer Einheiten, bzw. der Distanz zwischen ihnen;
2. der Beurteilung der Güte einer Schätzung (z.B. aufgrund einer Stichprobe) oder der Treffsicherheit einer Prognose, sowie der Messung von Konzepten wie Risiko, Zuverlässigkeit und Fehleranfälligkeit.

zu 1:

Streuungsmaße sind wichtige Ergänzungen zu den Mittelwerten, die die zentrale Tendenz einer Verteilung widerspiegeln sollen. Bei geringer Streuung ist ein Mittelwert eher ein typischer Wert einer Verteilung, als bei einer starken Variabilität der Daten.

Für bestimmte Probleme kann die Streuung (Dispersion) einer Häufigkeitsverteilung sogar wichtiger sein als der Mittelwert. Bei zwei Garnsorten A und B kann z.B. das Garn A zwar eine größere mittlere Reißfestigkeit als das Garn B haben, trotzdem kann B dem Garn A vorgezogen werden, weil die Variabilität des Garns A im Vergleich zu Garn B derart groß ist, dass der Anteil der Fadenbrüche aufgrund eines häufigeren Unterschreitens der kritischen Reißfestigkeit beim Garn A nicht akzeptabel ist.

zu 2:

Die Streuung ist auch von Bedeutung für die Beurteilung der Treffsicherheit einer statistischen Prognose, die in der Regel auf einem stochastischen Modell beruht, und sie ist generell von Bedeutung für die Stichprobentheorie.

Wie in der Induktiven Statistik zu zeigen sein wird, ist der für eine gegebene Genauigkeit und Sicherheit der Schätzung erforderliche Stichprobenumfang eine Funktion eines Streuungsmaßes der Grundgesamtheit. Man kann sich dies leicht anhand des folgenden Extremfalles klar machen: Sind alle Einheiten der Grundgesamtheit in bezug auf das Merkmal X gleich (ist also die Streuung der Variablen X in der Grundgesamtheit Null), so genügt ein Stichprobenumfang von n=1, also einer Einheit, um mit Sicherheit und ohne Fehler Aussagen über die Grundgesamtheit machen zu können. Häufig werden, wie hier, die Begriffe Streuung und Dispersion als synonym betrachtet. Bei einigen Autoren (z.B. Ferschl) wird der Begriff Dispersion jedoch eingeschränkt auf relative Streuungsmaße (auch Streuungskoeffizienten genannt), die dimensionslos sind und als Quotient aus einem absoluten Streuungsmaß und einem Lagemaß gebildet werden (vgl. Abschn. 1d).

Beispiel 5.1 soll das Konzept der Streuung anhand verschiedener Häufigkeitsverteilungen veranschaulichen (Abb. 5.1). Dabei wird jeweils die Varianz berechnet, ein Streuungsmaß, das jedoch erst an späterer Stelle (Abschn. 2) definiert wird.

Beispiel 5.1:

Gegeben seien die folgenden drei Häufigkeitsverteilungen mit jeweils gleichem arithmetischen Mittel $\bar{x} = 3$ und zunehmender Streuung (was anhand der Abb. 5.1 leicht zu beurteilen ist, da alle Verteilungen symmetrisch sind):

Verteilung A

x_i	h_i
3	1,0

Verteilung B

x_i	h_i
2	0,2
3	0,6
4	0,2

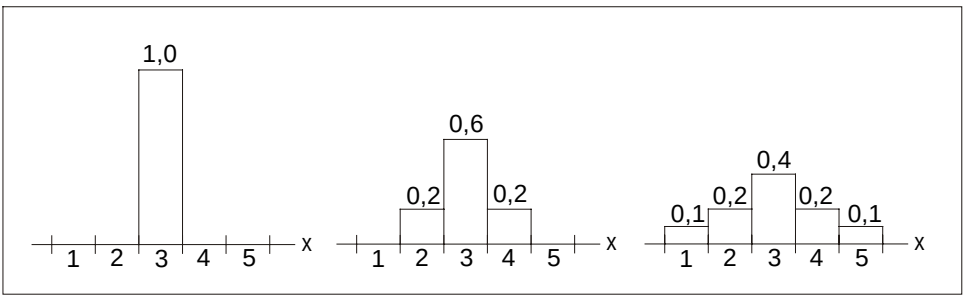
Verteilung C

x_i	h_i
1	0,1
2	0,2
3	0,4
4	0,2
5	0,1

Lösung 5.1:

Es ist ganz offensichtlich, dass in Abb.5.1 die Streuung von links nach rechts zunimmt. Verteilung A ist eine sog. Einpunktverteilung; alle Merkmalswerte sind gleich und die Streuung ist deshalb Null. Die Streuung, gemessen an der Varianz ist bei A: 0, bei B: 0,4 und bei C: 1,2.

Abb. 5.1



b) Konstruktion von Maßen der absoluten Streuung

Es gibt **drei Konstruktionsprinzipien** nach denen die gebräuchlichen Maße der absoluten Streuung gebildet werden, **wenn** das Merkmal X **metrisch skaliert** ist, also das Konzept des Abstands sinnvoll ist[1]. Ein Streuungsmaß kann danach berechnet werden als Maßzahl aus:

1. Abständen der Merkmalswerte von einem Lageparameter, z.B. von einem Mittelwert (nach diesem Prinzip sind die folgenden Streuungsmaße konstruiert: durchschnittliche Abweichung, Medianabweichung, Varianz und Standardabweichung),

[1] Zu Streuungsmaßen für nicht-metrisch skalierte Merkmale vgl. Exkurs am Ende von Abschnitt 3.

2. dem Abstand zweier Ordnungsstatistiken (Beispiele: Spannweite [range] oder mittlerer Quartilsabstand),
3. Abständen der Merkmalswerte untereinander (z.B. Ginis Maß).

Streuungsmaße, die Mittelwerte der Abstände der Beobachtungen von einem Mittelwert darstellen (Konstruktionsprinzip Nr. 1) unterscheiden sich nach

- der Art des Mittelwerts von dem die Abstände gemessen werden und
- nach der Art des Mittelwerts mit dem die Abstände gemittelt werden.

Übersicht 5.1 zeigt diese Zusammenhänge für einige besonders gebräuchliche Streuungsmaße.

Übersicht 5.1: Streuungsmaße nach dem Konstruktionsprinzip Nr. 1

Abweichung vom	Mittel der Abweichungen	Streuungsmaß
arithmet. Mittel	quadratisches Mittel	Standardabweichung
arithmet. Mittel*)	arithmetisches Mittel	Varianz
Median**)	arithmetisches Mittel	durchschn. Abweich.
Median**)	Median	Medianabweichung

*) quadrierte Abweichungen vom arithmetischen Mittel
**) absolute Abweichungen vom Median (Zentralwert)

Man kann sich aufgrund des Schemas der Übers. 5.1 auch weitere Streuungsmaße vorstellen, z.B. bei Verwendung des harmonischen Mittels. Da sich wegen der Schwerpunkteigenschaft des arithmetischen Mittels positive und negative Abweichungen gegenseitig aufheben, so dass stets gilt $\sum(x_v - \bar{x}) = 0$, ist eine arithmetische Mittelung nur bei anders definierten Abständen sinnvoll. Man kann

- absolute Abweichungen $\Sigma |x_v - \bar{x}|$, oder aber
- quadrierte Abweichungen $\Sigma (x_v - \bar{x})^2$ bilden,

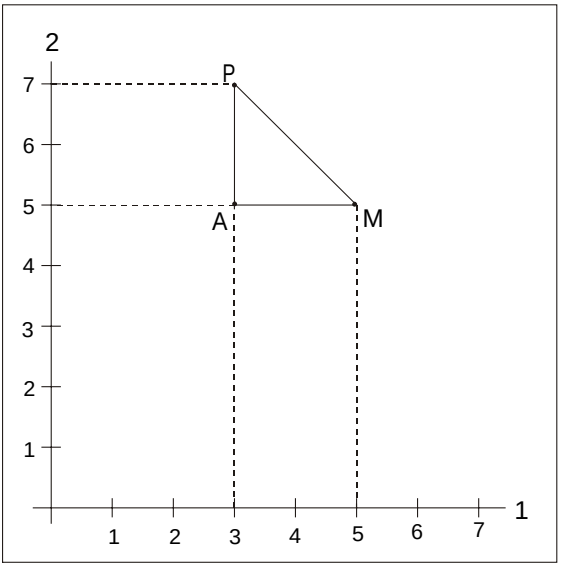
denn in beiden Fällen ist die Summe nichtnegativ und auch nicht notwendig Null (womit sie ohne jede Aussagefähigkeit wäre).

Der erste Weg ist von Laplace vorgeschlagen worden und wird bei der durchschnittlichen Abweichung benutzt, der zweite Weg geht auf Gauß zurück und wird bei der Varianz angewendet.

Beide Prinzipien sind auch bei der Konstruktion von **Distanzmaßen** in der Statistik gebräuchlich: die Summe absoluter Abweichungen wird benutzt bei der sog. City-Block-Distanz und die (bei der Standardabweichung verwendete) Wurzel aus der Summe der quadrierten Abweichungen ist die euklidische Distanz.

Geht man bei n Beobachtungen von einem n-dimensionalen Koordinatensystem aus, etwa zur graphischen Veranschaulichung von n=2 (Abb. 5.2), so sind die Daten darstellbar als ein Punkt (in Abb. 5.2 der Punkt P) und das arithmetische Mittel ebenfalls als ein Punkt (in Abb. 5.2 der Punkt M mit den Koordinaten 5 und 5).

Abb. 5.2: Euklidische- und City-Block-Distanz



In der Abb. 5.2 ist von den Werten $x_1 = 3$ und $x_2 = 7$ ausgegangen worden, so dass $\bar{x} = 5$ und $\Sigma |x_v - \bar{x}| = |3-5| + |7-5| = 4$ (das ist die Gesamtlänge des Weges von P über A nach M) und $\Sigma (x_v - \bar{x})^2 = 8$, was das Quadrat der euklidischen Distanz zwischen M und P ist. Während diese Distanz quasi die Luftlinie zwischen M und P darstellt, basiert die City-Block-Distanz auf der Vorstellung einer schachbrettartig aufgebauten Stadt, in der man von einem Punkt zum anderen durch Abschreiten rechtwinkliger Staßenzüge gelangt. Eine graphische Darstellung ist natürlich auch bei n=3 möglich. Dass Streuungsmessung und Distanzmessung miteinander verwandt sind, ist kein Zufall, denn die Streuung soll ja Ausdruck der Homogenität oder Heterogenität eines Datensatzes sein. Beide Distanzen, die euklidische- und die City-Block-Distanz sind Spezialfälle der Minkowski-Distanz :

$$d_{PM} = \left[\sum_{v=1}^{n} |x_v - \bar{x}|^r \right]^{\frac{1}{r}}$$ nämlich für r=1 und r=2.

Weitere Bemerkungen zu den Konstruktionsprinzipien:

1. Da bei allen diesen Konstruktionsprinzipien eine Abstandsmessung vorliegt, können die üblicherweise verwendeten Streuungsmaße nur für mindestens intervallskalierte Merkmale sinnvoll gebildet werden. Es gibt aber auch Streuungsmaße wie z.B. die Entropie und die im nachfolgenden Exkurs genannten Maße (z.B. Diversität), die nicht in dieses Schema der drei Konstruktionsprinzipien passen und deshalb auch keine metrische Skala voraussetzen.

2. Die Konstruktionsprinzipien hängen untereinander durchaus zusammen. So ist z.B. die Varianz aus quadrierten Abständen der Merkmalswerte vom arithmetischen Mittel gebildet (Konstruktionsprinzip Nr.1). Man kann die Varianz aber auch als ein Vielfaches der Summe der quadrierten Abweichungen der Merkmalswerte untereinander (Prinzip Nr. 3) darstellen (vgl. hierzu Satz 5.1).

3. Das Prinzip Nr. 2 wird auch bei der Messung der Schiefe angewendet.

4. Wenn auch das erste Konstruktionsprinzip am häufigsten angewandt wird, so hat doch das zweite mit der Verbreitung der explorativen Datenanalyse an Bedeutung gewonnen. Eine interessante Verknüpfung der beiden letzten Konstruktionsprinzipien stellt ein sog. Ginilike-Streuungsmaß dar, das resistent gegenüber Ausreißern ist.

c) Axiomatik absoluter Streuungsmaße

Absolute Streuungsmaße (S) sind Verteilungsmaßzahlen, die unter Berücksichtigung des Skalenniveaus die Axiome S1 bis S4 erfüllen.

S1	Ein absolutes Streuungsmaß S soll den Wert Null annehmen, falls $x_1 = x_2 = ... = x_n = \bar{x}$ gilt, d.h. wenn alle Merkmalswerte identisch sind.
S2	Sofern mindestens zwei Merkmalswerte x_i und x_j voneinander verschieden sind, ist $S > 0$ $(i,j = 1,2,...,n)$.
S3	Ersetzt man den Beobachtungswert x_k aus der Folge der Beobachtungen x_v $(v = 1,2,...,n)$ durch den neuen Wert x_p, so dass die Summe der absoluten Abweichungen von x_p von allen übrigen Werten größer ist als die Summe der absoluten Abweichungen von x_k von allen übrigen Werten, so soll das Streuungsmaß S nicht abnehmen.

S4 Invarianz gegenüber Verschiebungen des Nullpunkts (Translationen) aber nicht gegenüber Maßstabsänderungen: Falls S die Maßeinheit der Merkmalswerte $x_1,x_2,...,x_n$ hat, soll für die Streuung S_y der mit $y_v = a+bx_v$ transformierten Variablen X gelten: $S_y = |b|S_x$, $|b| > 0$. Für ein absolutes Streuungsmaß mit der quadrierten Maßeinheit der Merkmalswerte soll dann gelten $S_y = b^2S_x$.

Bemerkungen zu den Axiomen:

1. Ein Streuungsmaß S sollte nichtnegativ sein ($S \geq 0$), denn die "Streuung" ist ein durch ihr Ausmaß, nicht durch Ausmaß und Richtung zu kennzeichnender Tatbestand. Nach den Axiomen S1 und S2 ist ein Streuungsmaß S dann, und nur dann Null, wenn alle beobachteten Werte x_v gleich sind (bzw. [trivialer Fall] bei n = 1). Eine Obergrenze ist für S nicht vorgesehen, d.h. die Streuung ist nichtnegativ aber auch betragsmäßig nicht beschränkt, wenn nicht besondere Einschränkungen bezüglich der X-Variable gemacht werden. Mit den Merkmalswerten $x_1 = x$ und $x_2 = x + n$ ist z.B. die Varianz als Streuungsmaß $n^2 h_1 h_2$, was mit wachsendem n über alle Grenzen zunimmt.

2. Axiom S3 geht von der intuitiven Vorstellung der "Streuung" aus: Je mehr die Beobachtungswerte voneinander differieren, desto größer sollte ein Streuungsmaß sein. Würde man anstelle der obigen Formulierung von S3 auf Abweichungen von einem Mittelwert abstellen, dann entstünde das Problem der Wahl eines geeigneten Mittelwerts. Ein so formuliertes Axiom würde zu stark auf eines der genannten Konstruktionsprinzipien von Streuungsmaßen Bezug nehmen.

3. Axiom S4 bedeutet, dass ein Streuungsmaß invariant sein soll gegenüber

   a) Verschiebungen des Nullpunkts (Translation) mit der Größe a, d.h. es soll "verschiebungsinvariant" sein,

   b) Veränderungen der Skaleneinheit (Maßstabsänderung) in dem Sinne, dass eine Ver-b-fachung ($b \neq 0$) der Beobachtungswerte zu einem b-fachen Streuungsmaß führt, falls S die Maßeinheit der Merkmalswerte hat.

4. Die Axiome S3 und S4 stellen auf mindestens intervallskalierte Merkmale ab. Bei Streuungsmaßen, die für nominal- oder ordinalskalierte Merkmale konzipiert sind, gelten sie nicht.

d) Relative Streuung

Def. 5.1: Relative Streuung

Die Maße der relativen Streuung (S_r) sind definiert als Quotienten eines absoluten Streuungsmaßes S und eines Lokalisationsmaßes M (wenn $M \neq 0$),

$$(5.1) \quad S_r = \frac{S}{M}$$

sofern S die Maßeinheit der Merkmalswerte hat.

Bemerkungen zu Def. 5.1:

1. Dass absolute Streuungsmaße S die Maßeinheit der Merkmalswerte haben sollten, gewährleistet, dass ein relatives Streuungsmaß dimensionslos ist.

2. Relative Streuungsmaße lassen sich sinnvoll nur für mindestens intervallskalierte Merkmale interpretieren. Vergleicht man Häufigkeitsverteilungen, bei denen sich die Größenordnung der Merkmalswerte stark unterscheidet, dann sind sie aussagefähiger als absolute Streuungsmaße. Beim Konzept der relativen Streuung wird die Größenordnung der Merkmalswerte durch Maße der zentralen Tendenz widergespiegelt. Deshalb soll die für absolute Streuungsmaße im Axiom S4 geforderte Verschiebungsinvarianz hier gerade nicht erfüllt sein. Ein relatives Streuungsmaß besitzt demzufolge Eigenschaften, die bei Disparitätsmaßen als Axiome gefordert werden (vgl. Kapitel 6).

2. Varianz und Standardabweichung

a) Berechnung und Eigenschaften

Varianz und Standardabweichung sind die bekanntesten und am häufigsten benutzten Streuungsmaße. Sie sind aus quadrierten Abständen der Merkmalswerte vom arithmetischen Mittel gebildet (oben Konstruktionsprinzip Nr.1 genannt). Man kann die Varianz aber auch in ein Vielfaches

der Summe der quadrierten Abweichungen der Merkmalswerte untereinander umformen (vgl. Satz 5.1).

Def. 5.2: Varianz und Standardabweichung

a) Die Varianz s^2 eines mindestens intervallskalierten Merkmals X ist, wenn sie aus den einzelnen Merkmalswerten $x_1,x_2,...,x_n$ berechnet wird (ungewogener Ansatz), gegeben durch

$$(5.2) \quad s^2 = \frac{1}{n} \Sigma (x_v - \bar{x})^2 \qquad v = 1,2,...,n$$

und wenn sie aus einer Häufigkeitsverteilung (nicht aber bei klassierter Verteilung), d.h. aus den Merkmalsausprägungen $x_1,x_2,...,x_m$ berechnet wird (gewogener Ansatz), gilt

$$(5.3) \quad s^2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2 n_i = \Sigma (x_i - \bar{x})^2 h_i . \qquad i = 1,2,...,m$$

b) Die positive Quadratwurzel aus der Varianz heißt Standardabweichung s

$$(5.4) \quad s = + \sqrt{s^2} .$$

Bemerkungen zur Def. 5.2:

1. Die Varianz s^2 und die Standardabweichung s erfüllen die Axiome S1 und S2. Gilt für alle v (v=1,2,...,n) $x_v = \bar{x}$, so folgt $s^2 = s = 0$. Falls es auch nur einen Wert x_v gibt, der nicht identisch mit $\bar{x}$ ist, so folgt hieraus $s^2 > 0$.

2. Die Gültigkeit des Axioms S3 ergibt sich unmittelbar aus Satz 5.1, nach dem die Varianz s^2 als die $(1/n^2)$-fache Summe der Abweichungsquadrate $(x_i-x_j)^2$, i<j, dargestellt werden kann.

3. Mit $y_v = a + bx_v$ für alle v und $b \neq 0$ ist die Varianz s^2_y des zum Merkmal (zur Variablen) Y transformierten Merkmals X durch

$$s_y^2 = \frac{1}{n} \Sigma (y_v - \bar{y})^2 = \frac{1}{n} \Sigma [a+bx_v - (a+b\bar{x})]^2 = b^2 s_x^2$$

und die Standardabweichung s_y durch $s_y = |b| s_x$ gegeben. Mithin ist das Axiom S4 erfüllt.

4. Die Standardabweichung s ist das quadratische Mittel der Abweichungen $(x_v - \bar{x})$ der Merkmalswerte vom arithmetischen Mittel. Sie ist

besonders anschaulich im Falle [annähernd] normalverteilter Variablen (Lage der Wendepunkte) [Induktive Statistik].

5. Nach dem Verschiebungssatz (Satz 5.2) kann die Varianz bei Einzelbeobachtungen auch wie folgt geschrieben werden:

$$(5.5) \quad s^2 = \frac{1}{n} \Sigma\, x_v^2 - \bar{x}^2$$

bzw. bei einer Häufigkeitstabelle:

$$(5.6) \quad s^2 = \frac{1}{n} \Sigma\, x_i^2\, n_i - \bar{x}^2 = \Sigma\, x_i^2\, h_i - \bar{x}^2$$

Um die Auswirkungen von Rundungsfehlern zu begrenzen, empfiehlt sich insbesondere bei der Entwicklung von Computerprogrammen s^2 nach Gl. 5.5 bzw. 5.6 zu berechnen.

6. Der Verschiebungssatz gem. Bem. Nr. 5 ist ein Spezialfall des Steinerschen Verschiebungssatzes. Die Varianz s^2 läßt sich danach auch mittels der folgenden Beziehung berechnen:

$$(5.7) \quad s^2 = \frac{1}{n} \Sigma\, (x_i - c)^2\, n_i - (\bar{x} - c)^2 .$$

Hierbei ist c eine beliebige reelle Zahl. Der erste Summand auf der rechten Seite von Gl. 5.7 ist die um c berechnete Varianz, die man mit $s^2{}_c$ bezeichnen kann. Zwischen s^2 (oder $s_{\bar{x}}^2$) und s_c^2 besteht nach Gl.5.7 die folgende Beziehung:

$$(5.7a) \quad s_{\bar{x}}^2 = s_c^2 - (\bar{x} - c)^2 \qquad \text{(Steinerscher Verschiebungssatz).}$$

Aus Gl. 5.7a ist unmittelbar zu erkennen:

- die Minimumeigenschaft des arithmetischen Mittels ($s_{\bar{x}}^2$ ist minimal für $\bar{x} = c$) und
- mit c = 0 erhält man Gl. 5.5 als Spezialfall des Verschiebungssatzes von Steiner.

7. Eine weitere Darstellungsart der Varianz wird im Satz 5.3 angegeben.

8. Ersetzt man das arithmetische Mittel $\bar{x}$ in s^2 durch einen anderen Mittelwert M, so spricht man von der **mittleren quadratischen Abwei-**

chung $s_M^2 = n^{-1} \Sigma (x_v - M)^2$. Aus der Minimumeigenschaft des arithmetischen Mittels folgt, dass $s^2 \leq s_M^2$.

9. In Satz 5.4 wird der Einfluß einer zusätzlichen Beobachtung auf die Varianz untersucht. Danach reduziert die neue Beobachtung x_{n+1} die Varianz, falls ihr Abstand vom arithmetischen Mittel das $\sqrt{(n+1)/n}$-fache der ursprünglichen Standardabweichung unterschreitet. Ein Ausreißer kann dagegen die Varianz über alle Grenzen hinaus wachsen lassen, was zeigt, dass s^2 keine resistente Maßzahl der Streuung ist.

10. Eine wichtige Eigenschaft der Varianz ist die im Satz 5.5 gezeigte **Streuungszerlegung**. Danach läßt sich die Varianz s^2 der Variablen X für die aus r Teilgesamtheiten mit den Umfängen $n_1,n_2,...,n_r$ zusammengesetzten Gesamtheit zerlegen in eine externe Varianz s_{ext}^2 und eine interne Varianz s_{int}^2, so dass gilt

(5.8) $s^2 = s_{ext}^2 + s_{int}^2$.

Die externe und die interne Varianz sind jeweils gewogene Mittelwerte. Und zwar ist die externe Varianz

(5.9) $s_{ext}^2 = \Sigma h_k(\bar{x}_k - \bar{x})^2 \quad \text{mit } h_k = \frac{n_k}{n}$

ein gewogenes Mittel der quadrierten Abstände zwischen den r Mittelwerten der Teilgesamtheiten und dem Gesamtmittelwert $\bar{x}$.
Die interne Varianz ist demgegenüber das gewogene Mittel der Varianz s_k^2 der Teilgesamtheiten

(5.10) $s_{int}^2 = \Sigma h_k s_k^2$

mit den relativen Häufigkeiten h_k als Gewichte.

Die in Satz 5.5 dargestellte Varianzzerlegung ist eine Verallgemeinerung der Berechnung der Varianz aus einer Häufigkeitsverteilung nach Gl. 5.3. Setzt man die Subskripte k und i gleich, so dass $\bar{x}_k = x_i$ und r=m, so geht (5.8) in (5.3) über, da dann annahmegemäß alle n_i Einheiten mit der Merkmalsausprägung x_i gleich $\bar{x}_k$ sind (mit $n_i \geq 0$) und dann die Varianzen s_k^2 verschwinden und damit auch s_{int}^2).
Die Bedeutung des Satzes 5.5 ergibt sich ferner daraus, dass mit ihm das Verhalten der Varianz im Falle von Zerlegung und Aggregation (vgl. auch Bem. 11) ersichtlich wird.

- Bei der **Zerlegung** wird versucht, zu einer Kausalinterpretation zu gelangen, indem man die Beiträge verschiedener "Variationsquellen" zur Gesamtvarianz ermittelt (vgl. z.B. das Bestimmtheitsmaß in Kap. 7).
- Bei der **Aggregation** geht es nicht nur darum, die Varianz für die Gesamtheit zu berechnen, sondern auch zu zeigen, in welchem Maße ihr Wert von der Struktur der Gesamtmasse (d.h. den Anteilen der Teilmassen an der Gesamtmasse) abhängt.

Zwar hat die Varianz nicht die Aggregationseigenschaft, die in Kapitel 2 gefordert wurde, da sie zusätzlich noch den Summanden s_{ext}^2 enthält. Jedoch läßt die Zerlegung in eine externe und interne Varianz Interpretationsmöglichkeiten zu, wie sie bei einer entsprechenden Zerlegung anderer Maßzahlen der Streuung nicht gegeben wären, weshalb die Varianz auch anderen, einfach konstruierten und anschaulicher zu interpretierenden Streuungsmaßen (z.B. der durchschnittlichen Abweichung) vorgezogen wird.

11. Klassierte Daten

Liegen die Daten als klassierte Verteilung mit den Größenklassen k=1,2,...,r vor, so ergibt sich die Gesamtvarianz aufgrund der Streuungszerlegung mit

$$(5.11) \quad s^2 = \Sigma\, h_k(\bar{x}_k - \bar{x})^2 + \Sigma\, h_k s_k^2 = s_{ext}^2 + s_{int}^2$$

wobei s_k^2 die Varianz innerhalb der k-ten Klasse ist.
Häufig sind aber bei einer klassierten Verteilung die Einzelwerte nicht mehr bekannt. Während sich die (wahren) arithmetischen Mittelwerte $\bar{x}_k$ der Klassen durch die Klassenmitten m_k approximieren lassen, hat man für die Varianzen s_k^2 keine unmittelbar naheliegenden Näherungsgrößen. Sofern die Streuung innerhalb der Klassen im Vergleich zur Streuung zwischen den Klassen vernachlässigbar gering ist, kann man

$$(5.11a) \quad s_m^2 = \Sigma\, (m_k - m)^2\, h_k$$

(mit m als wahrem oder geschätzten Gesamtmittelwert) als Näherung für die Varianz s^2 verwenden. Im allgemeinen wird man sich in der Praxis mit der Näherung (5.11a) zur Abschätzung der Varianz einer klassierten Verteilung zufrieden geben müssen. Spezielle Korrekturen sehen Informationen über die Verteilung der Merkmalswerte innerhalb der Klassen vor.

Die beiden bekanntesten Korrekturmöglichkeiten sind:

a) Korrektur bei Gleichverteilung innerhalb der Klassen

Sofern die Merkmalswerte innerhalb der Klassen gleichverteilt sind, wird die Varianz s^2 durch s_m^2 unterschätzt. In diesem Fall ist die Varianz durch

(5.13) $s^2 = \Sigma\, n_k \left[(m_k - m)^2 + \frac{b_k(n_k^2 - 1)}{12n_k^2} \right]$

mit der Klassenbreite b_k der k-ten Klasse (k = 1,2,...,r) gegeben.

b) Sheppard-Korrektur (Dreiecksverteilung innerhalb der Klassen)

Falls sich die Merkmalswerte innerhalb der Klassen auf die Klassenmitten konzentrieren oder um die Klassenmitte dreiecksverteilt sind, wird bei gleichen Klassenbreiten b die Korrektur von Sheppard empfohlen. Danach ist

(5.14) $SK = \frac{b^2}{12}$ (SK = Sheppard-Korrektur)

von s_m^2 subtrahiert.

12. Häufig verwendet man auch die Formel

$$\hat{\sigma}^2 = \frac{1}{n-1} \Sigma (x_v - \bar{x})^2 \qquad (v = 1,2,...,n)$$

als Varianz anstelle der Gl. 5.2

$$s^2 = \frac{1}{n} \Sigma (x_v - \bar{x})^2$$

Die Formel für $\hat{\sigma}^2$ ist der (erwartungstreue) Schätzwert für die Varianz der Grundgesamtheit (die σ^2 genannt werden soll) aufgrund der Daten einer Stichprobe, während Gl. 5.2 die Varianz der Daten der Stichprobe (oder nach dem gleichen Muster gerechnet, die Varianz der Grundgesamtheit) wiedergibt.

Beispiel 5.2:

Man berechne Varianz und Standardabweichung aus der folgenden Tabelle der durchschnittlichen Bruttomonatsverdienste männlicher Angestellter in Industrie und Handel nach ("alten") Bundesländern.

Bundesland	Verdienst	Bundesland	Verdienst
Schleswig Holstein	3986	Berlin (West)	4348
Niedersachsen	4081	Nordrhein Westfalen	4408
Saarland	4158	Hessen	4428
Bayern	4246	Baden Württemberg	4509
Bremen	4254	Hamburg	4766
Rheinland Pfalz	4285		

(Quelle: Statist. Jahrb. 1989, S.490)

Lösung 5.2:

Bundesland	x_v	$x_v - \bar{x}$	$(x_v - \bar{x})^2$	x_v^2
Schleswig Holstein	3.986	-329	108.241	15.888.196
Niedersachsen	4.081	-234	54.756	16.654.561
Saarland	4.158	-157	24.649	17.288.964
Bayern	4.246	-69	4.761	18.028.516
Bremen	4.254	-61	3.721	18.096.516
Rheinland-Pfalz	4.285	-30	900	18.361.225
Berlin(West)	4.348	33	1.089	18.905.104
Nordrhein-Westfalen	4.408	93	8.649	19.430.464
Hessen	4.428	113	12.769	19.697.184
Baden Württemberg	4.509	194	37.636	20.331.081
Hamburg	4.766	451	203.401	22.714.756
Summe	47.469	4*	460.572	205.306.567

* wegen Rundungsfehler 4 statt 0 (denn der wahre Mittelwert ist 4315,3636 und nicht 4315, wie in den Spalten $x_v - \bar{x}$ und $(x_v - \bar{x})^2$ gerechnet wurde).

$n = 11$

$\bar{x} = 47469/11 = 4315$

$s^2 = \frac{1}{n} \Sigma(x_v - \bar{x})^2 = 460572/11 = 41870{,}182$ und $s = 204{,}62$DM

andere Berechnungsweise (Gl. 5.6):

$s^2 = \frac{1}{n} \Sigma x_v^2 - \bar{x}^2 = 205306567/11 - (4315{,}3636)^2 = 41870{,}363$ DM^2,

(Variationskoeffizient [Gl. 5.51] V = 0,047 also 4,7%)

Beispiel 5.3:

Beispiel für eine Lineartransformation

Die 200 Beschäftigten einer Arbeitsstätte erhalten einen monatlichen Durchschnittslohn von 2.200 DM mit einer Standardabweichung von s = 800 DM. Aufgrund einer Lohnverhandlung soll das Monatsgehalt jedes Beschäftigten um 10% angehoben werden, und es soll in Zukunft jedes Jahr jedem Beschäftigten ein Urlaubsgeld in Höhe von 960 DM gewährt werden. Wie ändern sich Mittelwert, Standardabweichung und Varianz der Gehälter der Beschäftigten?

Lösung 5.3:

Es liegt eine Lineartransformation vor: die bisherigen Gehälter x_v (v=1,2,...,200) werden zu den "neuen" Gehältern y_v transformiert nach Maßgabe der Transformation $y_v = a + bx_v$ mit a = 960/12 = 80 und b = 1,1.

Man erhält damit das arithmetische Mittel $\bar{y} = 2500$ die Varianz $s_y^2 = (1,1)^2(800)^2 = (880)^2 = 774400$ und die Standardabweichung $s_y = 880$.

b) Sätze über die Varianz

Satz 5.1: Abweichungsquadrate

Die Varianz s^2 läßt sich als die durch n^2 geteilte Summe der Abstandsquadrate aller Merkmalswerte untereinander darstellen:

$$(5.15) \quad s^2 = \frac{1}{n^2} \sum_{i<j} (x_i - x_j)^2 \qquad (\text{mit } i,j = 1,2,...,n)$$

Beweis:

Ausgehend von Gl. 5.5 nehmen wir folgende Umformung vor

$$s^2 = n^{-1} \sum x_i^2 - (n^2)^{-1} (\sum x_i)^2 = (n^2)^{-1} [\, \sum n x_i^2 - (\sum x_i)^2 \,]$$

$$= \left[\sum (n-1) x_i^2 - \sum_{i} \sum_{\substack{j \\ i \neq j}} x_i x_j \right] / n^2 = \frac{1}{n^2} \left[\sum (n-1) x_i^2 - \sum_{i<j} 2 x_i x_j \right],$$

woraus (5.15) unter Verwendung der binomischen Formel für die Summenglieder direkt folgt.

Die Gleichung 5.15 ist gleichbedeutend mit

$$(5.16) \quad s^2 = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2$$

Die obige Doppelsumme stellt die Summe aller Elemente der folgenden Matrix dar

$$\begin{bmatrix} (x_1 - x_1)^2 & (x_1 - x_2)^2 & \dots & (x_1 - x_n)^2 \\ (x_2 - x_1)^2 & (x_2 - x_2)^2 & \dots & (x_2 - x_n)^2 \\ \vdots & \vdots & \vdots & \vdots \\ (x_n - x_1)^2 & (x_n - x_2)^2 & \dots & (x_n - x_n)^2 \end{bmatrix}$$

Satz 5.2: Verschiebungssatz

Bei Einzelwerten (v=1,2,...,n) kann man die Varianz s^2 in der Form

$$(5.5) \quad s^2 = \frac{1}{n} \sum x_v^2 - \bar{x}^2$$

darstellen bzw. im Falle einer Häufigkeitsverteilung

(5.6) $s^2 = \Sigma\, x_i^2 h_i - \bar{x}^2$.

Beweis:

Ausgehend von Gl. 5.3 erhält man für Einzelwerte

$$s^2 = \frac{1}{n}\,\Sigma\,(x_v^2 - 2x_v\bar{x} + \bar{x}^2) = \frac{1}{n}\,[\,\Sigma x_v^2 - 2\bar{x}\Sigma\, x_v + n\bar{x}^2]$$

woraus wegen $\bar{x} = (Sx_v)/n$ Gl. 5.5 folgt. Gl. 5.6 ist analog zu beweisen.

Satz 5.3

Die Varianz ist auch darstellbar als

(5.17) $s^2 = \frac{1}{n}\sum (x_v - \bar{x})x_v$ bzw.

(5.18) $s^2 = \frac{1}{n}\sum (x_i - \bar{x})x_i h_i$.

(Eine ähnliche Beziehung gilt auch für die Kovarianz; vgl. Gl. 7.17)

Beweis:

Aus Gl. 5.5 folgt mit $\bar{x} = (\Sigma x_v)/n$

$$s^2 = \frac{1}{n}\left[\Sigma x_v^2 - \bar{x}\,\Sigma\, x_v\right] = \frac{1}{n}\,\Sigma\, x_v(x_v - \bar{x})$$ also Gl.5.17.

In gleicher Weise zeigt man Gl. 5.18.

Satz 5.4: Einfluss einer zusätzlichen Beobachtung

Sei s_{n+1}^2 die Varianz der um die Beobachtung x_{n+1} erweiterten Daten und x_{n+1} das arithmetische Mittel der Werte $x_1, x_2, \ldots, x_n, x_{n+1}$ ($s_n^2 = s^2$ ist die Varianz und $\bar{x} = \bar{x}_n$ das arithmetische Mittel der ursprünglichen Zahlenfolge $x_1, x_2, \ldots, x_n$), dann ist:

(5.19) $s_{n+1}^2 = \frac{1}{n+1}\sum_{v=1}^{n+1}(x_v - \bar{x}_{n+1})^2$.

Die sog. ***Sensitivitätsfunktion*** SF der Varianz s^2, die mit

(5.20) $SF(\bar{x}_{n+1}, s_n^2) = (n+1)(s_{n+1}^2 - s_n^2)$

definiert ist und die (n+1)-fache Veränderung (nicht notwendig Zunahme) der Varianz darstellt, ist durch den Ausdruck

$$(5.21) \quad SF = \frac{n(x_{n+1} - \bar{x})^2}{(n+1)} - s_n^2$$

gegeben.

Beweis:

Aus der Definition von s_{n+1}^2

$$s_{n+1}^2 = \frac{1}{n+1}\left[(x_{n+1} - \bar{x}_{n+1})^2 + \Sigma\,(x_v - \bar{x}_{n+1})^2)\right]$$

erhält man

$$(n+1)s_{n+1}^2 = (x_{n+1} - \bar{x}_{n+1})^2 + \Sigma[(x_v - \bar{x}) + (\bar{x} - \bar{x}_{n+1})]^2$$

(v = 1,2,...,n) und wegen der Schwerpunkteigenschaft des arithmetischen Mittels

$$(n+1)s_{n+1}^2 = (x_{n+1} - \bar{x}_{n+1})^2 + \Sigma(x_v - \bar{x})^2 + \Sigma(\bar{x} - \bar{x}_{n+1})^2$$

Daraus erhält man nach einigen Umformungen

$$(n+1)s_{n+1}^2 = n(x_{n+1} - \bar{x})^2/(n+1) + ns^2{}_n \text{ und damit Gl. 5.21.}$$

Interpretation:

Die Varianz $s^2 = s_n^2$ verändert sich umso mehr durch einen hinzukommenden Wert x_{n+1}, je stärker x_{n+1} vom bisherigen Mittelwert abweicht. Man sieht ferner, dass die Varianz durch einen hinzukommenden Wert nicht notwendig größer werden muss. Dies sei anhand des folgenden Beispiels demonstriert.

Beispiel 5.4:

Der Zusammenhang der Gl. 5.21 soll anhand der folgenden Merkmalswerte verifiziert werden:

ursprüngliche Beobachtungen (n=4): 2,3,5,6;
(damit ist $\bar{x}_n = 4$ und $s^2{}_n = s^2{}_4 = 10/4 = 2{,}5$)
der neu hinzukommende Wert sei nun

a) $x_{n+1} = x_5 = 4$ bzw.
b) $x_{n+1} = x_5 = 9$.

Lösung 5.4:

Nach Gl. 5.21 gilt

$$(n+1)(s_{n+1}^2 - s_n^2) = \frac{n}{n+1}(x_{n+1} - \bar{x})^2 - s_n^2$$

Für das Zahlenbeispiel erhält man im Fall a): $s_{n+1}^2 = 2$ und damit für SF nach Gl. 5.20: $(n+1)(s_{n+1}^2 - s_n^2) = 5(2-2,5) = -2,5$
und für SF nach Gl. 5.21

$$\frac{n(x_{n+1} - \bar{x})^2}{n+1} - s_n^2 = 4(4-4)^2/5 - 2,5 = -2,5$$

also das gleiche Ergebnis, wie nach Satz 5.4 ja auch zu erwarten war und was übrigens auch demonstriert, dass die Varianz durch Hinzutreten eines weiteren Merkmalswerts nicht notwendig größer werden muss. Dies ist jedoch im Fall b) der Fall. Dort gilt

$s_{n+1}^2 = 30/5 = 6$ und für Gl. 5.21 erhält man die eingesetzten Zahlen:
$5(6-2,5) = (4/5)(9-4)^2-2,5.$

Das bestätigt auch Bem. 9 zu Def. 5.2, wonach es darauf ankommt ob $x_{n+1} - \bar{x}$ größer oder kleiner ist als $S_n \sqrt{(n+1)/n}$. Im Falle a ist die Differenz kleiner als $\sqrt{2,5} \cdot \sqrt{5/4} = 1,77$ nämlich Null (die Varianz verringert sich) und im Fall b ist sie 9 - 4 = 5 und damit größer als 1,77 und die Varianz vergrößert sich. Träte als fünfter Wert der Wert $x_{n+1} = 5,77$ hinzu, so würde die Varianz gleich bleiben: $s_{n+1}^2 = s_n^2 = 2,5$. Ist x_{n+1} kleiner (etwa 4) so wird sie kleiner, ist x_{n+1} größer (etwa 9) so wird sie größer.

Satz 5.5: Streuungszerlegung

Gegeben seien r Teilgesamtheiten $G_1, G_2, ..., G_r$ der Gesamtheit G, die eine Zerlegung bilden (r = 1,2,..,r). Die Umfänge der Teilgesamtheiten seien n_k mit $\Sigma n_k = n$ (so dass $h_k = n_k/n$). Die Beobachtung x_i sei jeweils ein Element der k-ten Teilgesamtheit ($x_i \in G_k$). Die Teilgesamtheiten haben die arithmetischen Mittel $\bar{x}_k$ und die Varianzen

$$(5.22) \quad s_k^2 = \frac{1}{n_k} \sum_{x_i \in G_k} (x_i - \bar{x}_k)^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_i - \bar{x}_k)^2$$

Dann gilt für die Varianz s^2 der Gesamtheit G

$$(5.12) \quad s^2 = \Sigma h_k s_k^2 + \Sigma h_k (\bar{x}_k - \bar{x})^2 .$$

Beweis:

Die Varianz s^2 läßt sich in der Form

$s^2 = \frac{1}{n} \sum_{k}^{r} \sum_{i=1}^{n_k} (x_i - \bar{x})^2$ schreiben, was sich durch Nullergänzung zu

$s^2 = \frac{1}{n}\sum_k \sum_i [(x_i - \bar{x}_k)^2 + (\bar{x}_k - \bar{x})]^2$ umformen läßt.

Ausmultiplizieren und Summieren ergibt dann wegen der Schwerpunkteigenschaft des arithmetischen Mittels

$$s^2 = \frac{1}{n}\sum_k \sum_i (x_i - \bar{x}_k)^2 - \frac{2}{n}\sum_k \left[(\bar{x}_k - \bar{x})\sum_i (x_i - \bar{x}_k)\right] + \frac{1}{n}\sum_k \sum_i (\bar{x}_k - \bar{x})^2$$

Der erste Summand ergibt die interne Varianz, der zweite verschwindet wegen der Schwerpunkteigenschaft von $\bar{x}_k$ und der dritte ist die externe Varianz.

Beispiel 5.5:

Streuungszerlegung

Die Firma B – GmbH & Co KG hatte Probleme mit ihrer Belegschaft und ließ eine Untersuchung durch einen BWL-Professor und einen Statistiker durchführen:

1. Der BWL-Prof. stellte nach längerer intensiver Forschungsarbeit fest, dass die Unternehmung (nicht: das Unternehmen) B ein komplexes soziotechnisches System mit einer Triade von Subsystemen ist, in welchem eine nicht näher bestimmte Anzahl von Wirtschaftssubjekten mit einer signifikant differierenden Wirkungsintensität dergestalt operieren, dass emotionale Dysfunktionalitäten virulent waren, über deren Ausmaß aber ohne weitere hochkomplexe Evaluation kaum genauere Aussagen möglich sind und über die strategisch und situativ zu entscheiden ist.

2. Der Statistiker stellte fest, dass in den drei Betrieben des Unternehmens die insgesamt 2000 Beschäftigten sehr unterschiedlich verdienten, so dass in der Belegschaft Unfrieden herrschte. Er ermittelte die folgenden Zahlen:

Betrieb	Anzahl der Beschäftigten	Durchschnitts-verdienst ($\bar{x}$)	Standardab-weichung
1	500	2400	400
2	800	3000	600
3	700	2800	500

Man bestimme Mittelwert und Varianz der Verdienstverteilung des gesamten Unternehmens!

Lösung 5.5:

Mittelwert $\bar{x} = 2400 \cdot 0{,}25 + 3000 \cdot 0{,}4 + 2800 \cdot 0{,}35 = 2780$
interne Varianz s_{int}^2 :
$\Sigma h_j \cdot s_j^2 = 0{,}25 \cdot (400)^2 + 0{,}4 \cdot (600)^2 + 0{,}35 \cdot (500)^2 = 271500$

(Standardabweichung $s_{int.}$ 521,06; sie liegt zwischen 400 und 600)
externe Varianz s^2_{ext}:
$0{,}25\cdot(2400-2780)^2 + 0{,}4\cdot(3000-2780)^2 + 0{,}35\cdot(2800-2780)^2 = 55600$.
Gesamtvarianz 271500 + 55600 = 327100 (Standardabw. 571,93).

3. Andere Maße der absoluten Streuung

a) Durchschnittliche Abweichung und Medianabweichung

Def. 5.3: durchschnittliche- und Medianabweichung

a) Mit $a_1,a_2,...,a_n$ seien die absoluten Abweichungen der Merkmals**werte** $x_1,x_2,...,x_n$ eines mindestens intervallskalierten Merkmals X vom Median $\tilde{x}_{0,5}$ bezeichnet

(5.23) $a_v = |x_v - \tilde{x}_{0,5}| \quad v=1,2,...,n$

und $a_1,a_2,...,a_m$ seien die entsprechenden absoluten Abweichungen der Merkmals**ausprägungen** $x_1,x_2,...,x_m$

(5.24) $a_i = |x_i - \tilde{x}_{0,5}| \quad i=1,2,...,m.$

Dann ist das arithmetische Mittel der absoluten Abweichungen vom Median

(5.25) $d_x = \frac{1}{n} \Sigma a_v$ bei Einzelwerten bzw.

(5.26) $d_x = \Sigma a_i h_i$ bei Häufigkeitsverteilungen

die **durchschnittliche Abweichung** (vom Median). Üblich ist auch die Bezeichnung mittlere - oder **mittlere absolute Abweichung** (mean absolute deviation) .

b) Der Median (Zentralwert) der n absoluten Abweichungen a_v heißt **Medianabweichung** m_x. Bei Einzelwerten ist m_x der (n+1)/2 - te Wert, bzw. der Mittelwert aus dem n/2 - ten und dem folgenden Wert in einer der Größe nach geordneten Folge der absoluten Abweichungen a_v:

(5.27) $m_x = \begin{cases} a_{(n+1)/2} & \text{falls n ungerade} \\ ½[a_{(n/2)} + a_{(n/2+1)}] & \text{falls n gerade} \end{cases}$

c) Ein selteneres, in erster Linie in der Technik angewandtes Streuungsmaß ist a_{max}, die **maximale** absolute **Abweichung** a_v. Da das Maximum ein Grenzfall des

Potenzmittels ist, kann man auch die maximale Abweichung als Streuungsmaß nach dem Konstruktionsprinzip Nr. 1 auffassen.

Bemerkungen zu Def. 5.3:

1. Ebenso wie die Varianz und die Standardabweichung sind die mittlere absolute Abweichung und die Medianabweichung Maßzahlen, die nach dem ersten Konstruktionsprinzip, also unter Verwendung der Abstände der Beobachtungswerte von einem Lagemaß gebildet werden.

2. Da in d_x und m_x die absoluten Abweichungen der Merkmalswerte bzw. -ausprägungen vom Median einbezogen werden, sind beide Maßzahlen nichtnegativ.
   Gilt $x_1 = x_2 = ... = x_n$, so ist $\tilde{x}_{0,5} = x_v$ für alle v=1,2,...,n und somit $d_x = m_x = 0$, so dass d_x und m_x das Axiom S1 erfüllen.

3. Sobald x_i und x_j ungleich sind, so dass $|x_i - \tilde{x}_{0,5}| > 0$ oder/und $|x_j - \tilde{x}_{0,5}| > 0$ muss auch $d_x > 0$ sein.
   Mithin gilt das Axiom S2 für d_x. Allerdings sind entartete Fälle denkbar, in denen m_x Axiom S2 nicht erfüllt (Beispiel 5.7 am Ende dieser Bemerkungen).

4. Verhalten von m_x und d_x bei Lineartransformationen $y_v = a+bx_v$:
   $|y_v - \tilde{y}_{0,5}| = |a + bx_v - (a + b\tilde{x}_{0,5})| = |bx_v - b\tilde{x}_{0,5}| = |b|\, a_v$.

   Daraus folgt $d_y = |b|\, d_x$ und $m_y = |b|\, m_x$, womit die Gültigkeit des Axioms S4 gezeigt ist.

5. Verschiedentlich wird auch anstelle von d_x die weniger übliche mittlere absolute Abweichung um $\bar{x}$ verwendet, die wir d_x^* nennen wollen:

$$(5.28) \qquad d_x^* = \begin{cases} 1/n \sum |x_i - \bar{x}| & \text{bei Einzelwerten} \\ \sum |x_i - \bar{x}| h_i & \text{bei Häufigkeitsverteilungen} \end{cases}$$

   Aus der Minimumeigenschaft von $\tilde{x}_{0,5}$ folgt (5.29) $d_x \le d^*,_x$.

6. Im Vergleich zur Standardabweichung gilt (5.30) $d_x \le d^*,_x \le s$.

7. Der Vorteil der durchschnittlichen Abweichung ist ihre besondere Anschaulichkeit: d_x ist die durchschnittliche Entfernung einer Beobachtung vom Median. Die Standardabweichung ist zwar auch eine mitt-

lere Abweichung, aber das quadratische Mittel ist weniger allgemeinverständlich.

8. Als Faustregel gilt bei eingipfligen, der Normalverteilung ähnlichen Verteilungen $d_x = 0{,}8s$. Gegenüber der Varianz und der Standardabweichung spielt die mittlere absolute Abweichung wegen der analytischen Unhandlichkeit absoluter Abweichungen nur eine untergeordnete Rolle. Insbesondere in der induktiven Statistik dominieren Varianz und Standardabweichung. In die Diskussion gekommen ist die mittlere absolute Abweichung jedoch durch eine Untersuchung von Tukey (1960). Danach ist d_x bei "Verunreinigungen" durch "schlechte" Daten der Standardabweichung überlegen.

9. Die Medianabweichung m_x wird in der explorativen Datenanalyse (EDA) als "Hilfsskalenschätzer" verwendet (In der angloamerikanischen Literatur wird "scale" in diesem Zusammenhang für den Begriff Streuung benutzt). Als Analogon zum Median bei den Lokalisationsmaßen gilt m_x als besonders robust.

Beispiel 5.6:

Es sei die folgende Altersverteilung von n = 9 Personen gegeben (in Einzelwerten): 21, 25, 34, 39, 43, 52, 64, 72, 80. Das arithmetische Mittel $\bar{x}$ dieser Verteilung beträgt 47,78. Berechnen Sie die mittlere absolute Abweichung um den Median und um das arithmetische Mittel sowie die Medianabweichung!

Lösung 5.6:

Für den Median erhält man $\tilde{x}_{0.5} = 43$, woraus sich eine mittlere absolute Abweichung (vom Median) von $d_x = 16{,}56$ ergibt, denn die absoluten Abweichungen betragen:

$|21 - \tilde{x}_{0.5}| = |21 - 43| = 22;$
$|25 - 43| = 18;\ |34 - 43| = 9;\ |39 - 43| = 4;\ |43 - 43| = 0;$
$|52 - 43| = 9;\ |64 - 43| = 21;\ |72 - 43| = 29;\ |80 - 43| = 37$

und die Summe dieser (zur Berechnung von m_x bereits der Größe nach geordneten) absoluten Abweichungen beträgt 0 + 4 + 9 + 9 + 18 + 21 + 22 + 29 + 37 = 149. Die durchschnittliche (mittlere) absolute Abweichung beträgt somit 149/9 = 16,556. Daraus ergibt sich auch, dass die Medianabweichung $m_x = 18$ ist.
Die mittlere absolute Abweichung um $\bar{x}$ beträgt $d^*_{,x} = 17{,}09$.

Beispiel 5.7:

Gegeben seien die Merkmalswerte $x_1 = x_2 = x_3 = 0$ und $x_4 = 1$.
Liegt eine Streuung vor? Wie groß ist die Medianabweichung?

Lösung 5.7:

Es liegt offensichtlich eine Streuung vor, da nicht alle vier Beobachtungen identisch sind. Der Median ist $\tilde{x}_{0,5} = 0$ und man erhält die Abweichungen $a_1 = a_2 = a_3 = 0$ und $a_4 = 1$. Die Medianabweichung ist damit Null, also $m_x = 0$, obgleich eine gewisse Streuung vorliegt.

b) Spannweite, Quartilsabstand und Quantilsabstände

Die im folgendenen definierten Streuungsmaße sind weniger gebräuchlich. Es sind Streuungsmaße, die nach dem zweiten Konstruktionsprinzip aus dem Abstand zweier Ordnungsstatistiken gebildet worden sind.

Def. 5.4: Spannweite, Quartilsabstand, Quantilsabstände

a) Die Differenz zwischen dem größten Beobachtungswert $x_{(n)}$ und dem kleinsten $x_{(1)}$ heißt **Spannweite** R (range, Wertebereich, Variationsbreite):

$$(5.31) \quad R = x_{(n)} - x_{(1)}$$

(Die Berechnung von R ist nur bei Einzelwerten, nicht bei Häufigkeitsverteilungen üblich).

b) Der **Quartilsabstand** $Q_{0,25}$ (Interquartilsabstand IQR) ist die Differenz zwischen dem dritten und ersten Quartil (Gl. 5.32) und der **mittlere Quartilsabstand** $\bar{Q}_{0,25}$ (Semiquartilsabstand) ist durch Gl. 5.33 gegeben:

$$(5.32) \quad Q_{0,25} = Q_3 - Q_1 \quad \text{und} \quad (5.33) \quad \bar{Q}_{0,25} = ½\, Q_{0,25}$$

c) Der **Quantilsabstand** (Interquantilsabstand) Q_p ist die Differenz zwischen dem (1-p)-Quantil $\tilde{x}_{1-p}$ und dem p-Quantil $\tilde{x}_p$,

$$(5.34) \quad Q_p = \tilde{x}_{1-p} - \tilde{x}_p \quad \text{mit } 0 < p < 0{,}5 \,,$$

Analog zu Gl.5.36 heißt dann die Maßzahl (5.35) $\bar{Q}_p = ½Q_p$ **mittlerer Quantilsabstand** (Semiquantilsabstand).

Beispiel 5.8:

Man berechne die Spannweite und den mittleren Quartilsabstand für das Beispiel 5.6!

Lösung 5.8:

R = 80 - 21 = 59. Die Quartile sind (mit Interpolation) $Q_1 = 29{,}5$ (der 2,5-te Wert), $Q_2 = 43$ (= Median) und $Q_3 = 68$ (der 7,5-te Wert). Folglich ist der Interquartilsabstand 68 - 29,5 = 38,5 und der mittlere Quartilsabstand 38,5/2 = 19,25.

Bemerkungen zu Def. 5.4:

1. Aufgrund der Differenzenbildung ist unmittelbar einsichtig, dass alle Maßzahlen das Axiom S1 erfüllen. Axiom S2 wird ganz offensichtlich von der Spannweite erfüllt, nicht aber notwendig auch von Q_p und damit den anderen obigen Maßzahlen (vgl. hierzu Bsp. 5.9). Auch Axiom S3 muss nicht notwendig vom mittleren Quartilsabstand erfüllt sein (vgl. für ein konstruiertes, extremes Beispiel Bsp. 5.10).

2. Da die Spannweite nur von den beiden extremen Werten eines Datensatzes abhängt, nutzt sie die in den Daten enthaltene Information unzureichend aus und reagiert äußerst empfindlich auf Ausreißer. Daher wird R als Streuungsmaß kaum verwendet. Allerdings hat die Spannweite eine gewisse Bedeutung bei Ausreißertests und in der statistischen Qualitätskontrolle.

3. Der Quartilsabstand gibt den Bereich an, in den 50% der mittleren Beobachtungswerte fallen. Einerseits wird dadurch ein beträchtlicher Teil der Informationen eines Datensatzes "verschenkt". Gerade deshalb ist aber diese Maßzahl sehr robust, weshalb sie in der explorativen Datenanalyse (vgl. Exkurs über Boxplots) verwendet wird.

4. In der Form $Q_{0,25} = (Q_3-Q_2)+(Q_2-Q_1)$ wird der Quartilsabstand in eine "rechtsseitige" Streuung $(Q_3 - Q_2)$ rechts vom Median $(\tilde{x}_{0,5} = Q_2)$ und eine "linksseitige" Streuung (Q_2-Q_1) aufgespalten, so dass $\bar{Q}_{0,25}$ als Mittelwert der beiden Abstände interpretiert werden kann. Diese rechts- und linksseitige Streuung wird auch zur Konstruktion eines Schiefemaßes herangezogen.

5. Bei normalverteilten Daten ist der Quartilsabstand gleich dem 0,6745-fachen Wert der Standardabweichung.

6. Da die folgenden Streuungsmaße jeweils Spezialfälle des Potenzmittels von Abweichungen darstellen gilt

(5.36) $d_x \le s \le R.$

7. Offensichtlich ist der (mittlere) Quantilsabstand eine Verallgemeinerung des (mittleren) Quartilsabstands. Neben dem Quartilsabstand bieten sich somit eine Reihe weiterer spezieller Streuungsmaße an:

Bekanntlich ist es möglich, Verteilungen in beliebig viele gleichhäufig besetzte Abschnitte (Quantile) zu zerlegen. Bei einer Vier-Teilung spricht man von einer Zerlegung in drei Quartile (Q_1, Q_2, Q_3, wobei Q_2 dem Median entspricht). Hieraus ist der Interquartilsabstand (IQR) abzuleiten. Mit der gleichen Betrachtungsweise könnte man ein Streuungsmaß auf der Basis der vier Quintile, die wir Q_1^*, Q_2^*, Q_3^*, Q_4^*, welche die Verteilung in fünf Abschnitte zerlegen, konstruieren. Die Differenz zwischen dem vierten und dem ersten Quintil wäre dann der **Quintilsabstand**. Näherungswerte für verschiedene Q_p stellen die in der explorativen Datenanalyse verwendeten "spreads" (spr) dar.

Beispiel 5.9:

Gegeben seien die sieben Beobachtungswerte 2,3,3,3,3,3,4. Man bestimme die Spannweite und den mittleren Quartilsabstand.

Lösung 5.9:

Die Spannweite ist R = 4 - 2 = 2. Aber der Quartilsabstand ist in diesem (sehr konstruierten) Beispiel Null (und damit auch der mittlere Quartilsabstand), da $Q_3 = Q_1 = 3$; offenbar erfüllt also dieses Streuungsmaß nicht notwendig die Forderung S2.

Beispiel 5.10:

Gegeben seien die folgenden sieben Beobachtungen 2,3,4,5,6,7,7. Der Wert 3 werde durch den Wert 10 ersetzt. Man bestimme die Spannweite und den mittleren Quartilsabstand für die Reihen:

a) 2,3,4,5,6,7,7 und
b) 2,4,5,6,7,7,10.

Erfüllt der mittlere Quartilsabstand das Axiom S3 ?

Lösung 5.10:

Die Spannweite beträgt

im Fall a) R = 7 - 2 = 5
im Fall b) R = 10 - 2 = 8

Die Voraussetzung des Axioms S3 ist, dass der Ersatz des Wertes 3 durch den Wert 10 zu einer größeren Summe der absoluten Abweichungen der

einzelnen Beobachtungswerte untereinander führt. Das ist in diesem Beispiel gegeben. Die genannte Summe beträgt
im Fall a) 50 und im Fall b) 64.
Die Spannweite wird, wie es Axiom S3 verlangt, nicht kleiner, sondern sogar größer. Der mittlere Quartilsabstand ist aber

im Fall a) $\bar{Q}_{0,25} = ½(7-3) = 2$ und
im Fall b) $\bar{Q}_{0,25} = ½(7-4) = 1,5$.

Exkurs: Boxplot

Ein Boxplot ist eine komprimierte graphische Darstellung eines Datensatzes, die von Tukey (1977) eingeführt worden ist. Eine Box, die durch das erste und dritte Quartil begrenzt und durch den Median geteilt wird, vermittelt einen Überblick über die mittleren 50% der Beobachtungen eines Datensatzes (die Breite der Box ist beliebig). Um die Verhältnisse eines Datensatzes an seinen äußeren Enden einschätzen zu können, werden sog. Zäune (whiskers, adjacent values) festgelegt, deren Abstand von den Quartilen Q_1 und Q_3 durch gestrichelte senkrechte Linien dargestellt wird. Und zwar sind die Zäune durch diejenigen Beobachtungen festgelegt, die gerade noch innerhalb des durch die Quartile und den Quartilsabstand (IQR,$Q_{0,25}$) definierten Intervalls $[w_e,w_u]$ mit

(5.37) $w_e = Q_1 - ½\, Q_{0,25}$ und (5.38) $w_u = Q_3 + ½\, Q_{0,25}$

liegen. Darüber hinaus liegende Beobachtungen sind als mögliche Ausreißer "verdächtig". Sie können einzeln durch ein * als outside values (outlayers) gesondert gekennzeichnet werden. Aus einem Boxplot lassen sich rasch Informationen über die Lokalisation (Lage des Median), Streuung (Höhe der Box) und die Schiefe (Vergleich der beiden Hälften der Box oder der Längen der gestrichelten Linien) eines Datensatzes sowie über evtl. vorliegende Ausreißer ("deviante Beobachtungen") gewinnen.

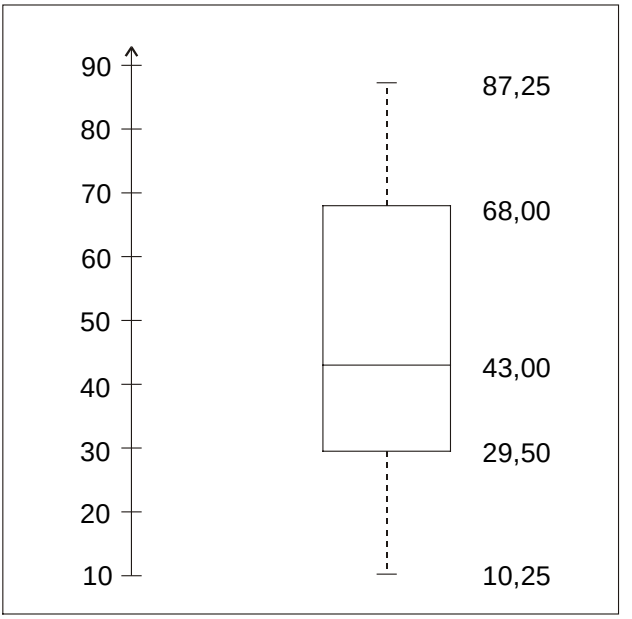
Beispiel 5.11:

Man zeichne den Boxplot für das Beispiel 5.6 (bzw. 5.8).

Lösung 5.11:

Im Beispiel 5.6 (bzw. 5.8) erhält man $Q_1 = 29,5$ $Q_2 = 43$ und $Q_3 = 68$. Ferner ist $Q_{0,25} = 38,5$ und die Zäune liegen bei $w_e = 29,5 - 19,25 = 10,25$ und $w_u = 68 + 19,25 = 87,25$. Es gibt also in diesem Beispiel keine ausreißerverdächtige Werte, weil der kleinste Wert 21 und der größte 80 ist. Ein Bild des Boxplots ist Abb. 5.3.

Abb. 5.3: Boxplot (Beispiel 5.11)



c) Ginis Dispersionsmaß (Ginis mittlere Differenz)

Ginis Dispersionsmaß (nicht zu verwechseln mit dem Disparitätsmaß von Giordano Gini [vgl. Kap. 6]) basiert auf dem dritten der drei Konstruktionsprinzipien von Streuungsmaßen. Es wird also aus den Abständen aller Beobachtungswerte untereinander gebildet.

Def. 5.5: Ginis Streuungsmaß

Für die Merkmalswerte $x_1,x_2,...,x_n$ eines metrisch skalierten Merkmals X ist Ginis Dispersionsmaß (auch mittlere Differenz genannt) gegeben durch

$$(5.39) \quad S_G = \frac{2}{n(n-1)} \sum_{v<w} |x_v - x_w|$$

(bei Einzelwerten v,w = 1,2,...,n) und bei einer Häufigkeitsverteilung durch

$$(5.40) \quad S_G = \frac{2}{n(n-1)} \sum_{i<j} |x_i - x_j| n_{ij} \; .$$

Beispiel 5.12:

Man berechne S_G nach Gl. 5.40 und S^*_G (vgl. Bem. 3 zu Def. 5.5) gem. Gl. 5.42 für die Daten

a) des Beispiels 5.10,
b) des Beispiels 5.6, bzw.5.8.

Lösung 5.12:

zu a) Die Summe der absoluten Abweichungen ist im Beispiel 5.10 bei 7 Werten bereits mit 50 bzw. 64 angegeben worden. Das ist die Doppelsumme

$$\sum_{v<w}|x_v - x_w| = \frac{1}{2}\sum_v\sum_w|x_v - x_w| \quad (\text{mit } v < w)$$

(die Berechnung dieser Summe wird in Teil b demonstriert)
Folglich ist bei 21 Paarvergleichen $S_G = 50/21 = 2{,}381$
bzw. $S_G = 64/21 = 3{,}048$ und

$S^*_G = 100/49 = 2{,}041$ bzw. $S^*_G = 128/49 = 2{,}612$ (weil $49 = 7^2$).

zu b) Datensatz von Beispiel 5.6: 21,25,34,39,43,52,64,72,80. Die folgende Matrix enthält die absoluten Differenzen:

	21	25	34	39	43	52	64	72	80	Summe
21	-	4	13	18	22	31	43	51	59	241
25		-	9	14	18	27	39	47	55	209
34			-	5	9	18	30	38	46	146
39				-	4	13	25	33	41	116
43					-	9	21	29	37	96
52						-	12	20	28	60
64							-	8	16	24
72								-	8	8
80									-	-
Summe										900

Die vollständige Matrix ist symmetrisch mit Nullen in der Hauptdiagonalen. Sie hat insgesamt $9^2 = 81$ Elemente, darunter $9(9-1)/2 = 36$ oberhalb der Hauptdiagonalen. Folglich ist $S_G = 900/36 = 25$ und $S^*_G = 1800/81 = 22{,}22 = (8/9)S_G$.

Bemerkungen zu Def. 5.5:

1. Offenbar erfüllt Ginis Dispersionsmaß das Axiom S1. Dass S_G auch Axiom S2 erfüllt, ergibt sich aus der analogen Betrachtung bei der Varianz (Satz 5.1). Auch Axiom S3 ist unmittelbar einsichtig.

2. Der Faktor 2/n(n-1) ist der reziproke Binomialkoeffizient $\binom{n}{2}$. Dies ist die Anzahl der Paarvergleiche ohne Berücksichtigung der Anordnung [Reihenfolge] von zwei verschiedenen Beobachtungswerten x_v und x_w (bzw. x_i und x_j). Mithin ist S_G das arithmetische Mittel der absoluten Abweichungen der Merkmalswerte untereinander, wobei jede Differenz einmal gerechnet wird.

3. Man kann auch analog zur Varianz eine mittlere Differenz aus allen n^2 möglichen Differenzen zwischen zwei Beobachtungen x_v und x_w, bzw. x_i und x_j bilden:

$$(5.41) \quad S_G^* = \frac{1}{n^2} \sum_{v=1}^{n} \sum_{w=1}^{n} |x_v - x_w|.$$

   Wegen der zusätzlichen Mittelung über die n Null-Differenzen für v = w in S_G^* gilt

$$(5.42) \quad S_G = \frac{n}{n-1} S_G^*.$$

4. Aus der Darstellung der Varianz in Abhängigkeit von den Abständen der Beobachtungswerte (Satz 5.1) untereinander erkennt man, dass

$$\sum_{v<w} (x_v - x_w) = \sum_{v=1}^{n} \sum_{w=1}^{n} (x_v - x_w) = 0$$

   ist, weshalb sich die Wahl der **absoluten** Werte der Abweichungen als sinnvoll erweist.

5. Während die Varianz von extremen Abweichungen aufgrund der Quadrierung sehr stark beeinflusst wird, zeichnet sich Ginis Dispersionsmaß durch eine größere Resistenz gegenüber Ausreißern aus. Deshalb hat S_G in neuerer Zeit wieder eine gewisse Beachtung als ein Konzept zur Konstruktion von "Gini-like" Lokalisations- und Dispersionsmaßen gefunden.

6. Auf einen Zusammenhang zwischen S_G^* und dem Disparitätsmaß von Gini (D_G) wird in Kap. 6 eingegangen.

d) Entropie

Die im folgenden dargestellte Entropie (E) kann als Streuungsmaß für kategoriale (qualitative) Daten aufgefaßt werden, weil sie einige Eigenschaften besitzt, die für ein Streuungsmaß zu fordern sind. Sie eignet sich für nominalskalierte Merkmale, weil sie nur von den relativen Häufigkeiten, nicht aber von den Merkmalswerten abhängig ist. Gerade deshalb

reagiert E aber nicht auf Transformationen der Merkmalswerte und Veränderungen des Wertebereichs von X, was aber bei quantitativen Merkmalen im Widerspruch zur anschaulichen Vorstellung der "Streuung" steht. Die Entropie E spielt auch eine Rolle bei der Disparitätsmessung.

Def. 5.6: Entropie

Für die Häufigkeitsverteilung (x_j,h_j) eines Merkmals X ist die Maßzahl E als Entropie von X definiert ($h_j > 0$):

$$(5.43) \quad E = \Sigma\, h_j\, ld(1/h_j)$$

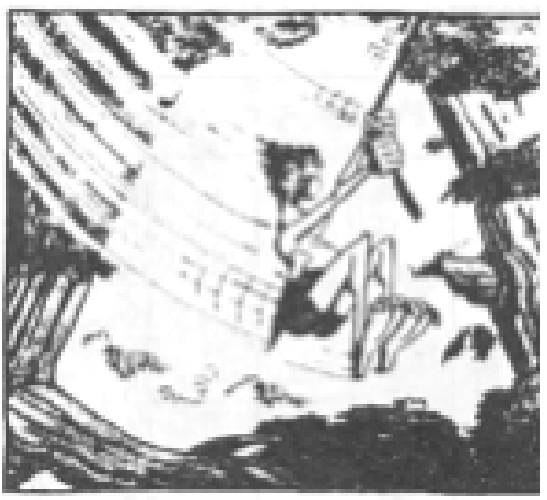
mit ld(x) als Logarithmus zur Basis 2 (logarithmus dualis); insbesondere gilt $ld(h_j) = \log(h_j)/\log(2) = 3{,}3219 \cdot \log(h_j)$. Die folgende Berechnungsformel ist äquivalent zu (5.43):

$$(5.44) \quad E = -\Sigma\, h_j\, ld(h_j).$$

Vor einer Darstellung der Interpretation und Eigenschaften der Entropie soll die Berechnung von E an einem Beispiel demonstriert werden. Das Konzept der Entropie stammt aus der Nachrichtentechnik, in der es den Gehalt einer Information quantifizieren soll.

Beispiel 5.13:

Schauspieler S (Rollenfach: jugendlicher Naturbursche) ist auf Tarzan-Filme spezialisiert. Gelegentlich spielt er aber auch in Krimis, Heimat- und Actionfilmen mit. Sein Produzent führte folgende Statistik über die Filme, an denen S mitgewirkt hat:



Art des Films	Anz. d. Filme mit S 1989	1990
Tarzan	4	6
Krimis	2	2
Heimat	1	1
Action	1	3
Summe	8	12

a) Hat die Unterschiedlichkeit der vielfältigen (bzw. vierfältigen) schauspielerischen Betätigung von S zu- oder abgenommen?

b) Man berechne die Entropien für die beiden Jahre!

c) Wie groß wären die Entropien, wenn S an jeweils gleich vielen Filmen jedes Typs mitgewirkt hätte?

Lösung 5.13:

a) Die Beantwortung dieser Frage dürfte ohne Berechnung eines für Nominalskalen geeigneten Streuungsmaßes schwierig sein.

b) Entropie 1989
$E_{1989} = (1/2)ld(2) + (1/4)ld(4) + (1/8)ld(8) + (1/8)ld(8) =$
$= ½\cdot 1 + ¼\cdot 2 + ¼\cdot 3 = 1{,}75.$

Entropie 1990
$E_{1990} = (1/2)ld(2) + (1/6)ld(6) + (1/12)ld(12) + (1/4)ld(4) =$
$= ½\cdot 1 + (1/6)\cdot 2{,}58496 + (1/12)\cdot 3{,}58496 + ¼\cdot 2 = 1{,}7296.$

c) Bei einer Gleichverteilung hätte S 1989 an zwei Filmen und 1990 an drei Filmen jedes Typs mitwirken müssen. Die Entropien wären dann gewesen $E^*_{1989} = E^*_{1990} = 4\cdot[¼\cdot ld(4)] = ld(4) = 2$.
Das Streuungsmaß E ist also in beiden Jahren durch die Anzahl der Ausprägungen des nominalskalierten Merkmals "Art des Films" nach oben begrenzt mit ld(4) = 2.

Bemerkungen zu Def. 5.6:

1. Im Falle der Einpunktverteilung (d.h. alle Merkmalswerte sind gleich) ist $E = 1\cdot ld(1) = 0$. Falls $m > 1$ ist $E > 0$, so dass beide Teile des Axioms S1 erfüllt sind. Sobald auch nur zwei unterschiedliche Beobachtungen auftreten ist $E > 0$, denn $½ld(2) = 1/2$. Somit erfüllt die Entropie auch das Axiom S2.

2. Im Unterschied zu allen bisher behandelten Maßzahlen führt bei E die Berechnung aus einer Häufigkeitsverteilung ("gewogene" Berechnung) nicht zum gleichen Ergebnis, wie die Berechnung aus Einzelwerten ("ungewogene" Berechnung), sondern notwendig zu einem kleineren Zahlenergebnis. Dies wird im Beispiel 5.14 demonstriert.
Wegen $h_j \geq 1/n$ gilt auch $ld(1/h_j) \leq ld(n)$, so dass $h_j\cdot ld(1/h_j) \leq h_j\cdot ld(n)$ für $n_j > 1$. Damit kann $E = \Sigma\ h_j\cdot ld(1/h_j)$ (gewogene Berechnung) nicht größer sein als $E_o = \Sigma(1/n)\cdot ld(n) = ld(n)$ (ungewogene Berechnung).

3. Satz 5.6 zeigt, dass $E_o = ld(n)$ auch die Obergrenze der Entropie darstellt. Mithin gilt

(5.45) $\quad 0 \leq E \leq ld(n)$

Nach Satz 5.6 nimmt E bei gegebener Anzahl m von Merkmalsausprägungen ein Maximum an, wenn jede Merkmalsausprägung gleich häufig ist.

4. E erfüllt die Axiome S3 und S4 nicht, da die Häufigkeiten nicht von einer Transformation der Merkmalswerte berührt werden und E nur von den Häufigkeiten, nicht aber von den Merkmalswerten abhängt.

5. Ersetzt man die relativen Häufigkeiten h_j durch die Merkmalsanteile q_j, so läßt sich E auch als Konzentrationsmaß verwenden (vgl. Kap. 6).

6. Besonders vorteilhaft ist das Verhalten der Entropie bei Aggregation und Zerlegung: Sind bei einer klassierten Verteilung jeweils alle n_k Beobachtungen innerhalb der Klasse k unterschiedlich (hat jede also die Häufigkeit $1/n_k$), so erhält man die Entropie E_k innerhalb der k-ten Klasse mit

$$(5.46) \qquad E_k = \sum_{l=1}^{n_k} \frac{1}{n_k} \operatorname{ld}(n_k) = \operatorname{ld}(n_k) \qquad (l = 1,2,...,n_k) \ .$$

Für die Gesamtentropie (auf der Grundlage von i=1,2,...,n Einzelwerten) gilt dann

$$(5.47) \qquad E_{ges} = \sum_{i=1}^{n} \frac{1}{n} \operatorname{ld}(n) = \operatorname{ld}(n).$$

Damit erhält man die folgende Streuungszerlegung nach Art der Varianzzerlegung (nach Satz 5.5)

(5.48)	$E_{ges} = E_{ext} + E_{int}$	mit
(5.48a)	$E_{ext} = \Sigma\, h_k \cdot \operatorname{ld}(1/h_k)$	und
(5.48b)	$E_{int} = \Sigma\, h_k\, E_k$.	

Die in Bsp. 5.15 verifizierte Zerlegung der Entropie gem. Gl. 5.48 ist wie folgt zu beweisen: Die Summe von E_{ext} und E_{int} gem. Gl. 5.49 und 5.50 beträgt

$\Sigma\, h_k \cdot \operatorname{ld}(n/n_k) + \Sigma h_k \cdot \operatorname{ld}(n_k) = \Sigma\, h_k\, [\operatorname{ld}(n/n_k) + \operatorname{ld}(n_k)] =$
$\Sigma\, h_k\, [\operatorname{ld}(n) - \operatorname{ld}(n_k) + \operatorname{ld}(n_k)] = \operatorname{ld}(n)\, \Sigma\, h_k = \operatorname{ld}(n) = E_{ges}$.

Beispiel 5.14:

Berechnen Sie die Entropie E für die

a) folgenden Einzelwerte : 2,3,3,4,4,4
b) folgende Häufigkeitsverteilung (vgl. auch Bsp. 6.3):

x_j	2	3	4
h_j	1/6	1/3	½

Es ist unschwer zu erkennen, dass es sich in beiden Fällen um die gleichen Daten handelt.

Lösung 5.14

a) Auf die Zahlenwerte für die Beobachtung kommt es nicht an. Es liegen sechs Beobachtungen vor, so dass sich E gem. Gl. 5.43 wie folgt errechnet:

$E = (1/6)ld(6) + ... + (1/6)ld(6) = ld(6) = log(6)/log(2) = 0{,}77815/0{,}3010 = 2{,}585.$

b) $E = (1/6)ld(6) + (1/3)ld(3) + (1/2)ld(2) = 2{,}585/6 + 1{,}585/3 + 1/2 = 1{,}459.$ Das ist, wie in Bem. Nr. 2 dargelegt, kleiner als die Berechnung aus Einzelwerten, die zu 2,585 führte.

Beispiel 5.15:

Gegeben sei die folgende Häufigkeitsverteilung (vgl. Bsp. 5.14):

Klasse	Beobachtungen	n_k	h_k
1	$x_1 = 2$	1	1/6
2	$x_2=3$, $x_3=4$	2	1/3
3	$x_4=3$, $x_5=3$, $x_6=3$	3	1/2

Man verifiziere anhand dieses Beispiels den Satz über die Aggregation bzw. Zerlegung der Entropie (vgl. oben Bem. Nr. 6)

Lösung 5.15:

Die externe Entropie ist im Bsp. 5.14 bereits berechnet worden mit dem Ergebnis $E_{ext} = (1/6)ld(6) + (1/3)ld(3) + (1/2)ld(2) = 1{,}459.$

Um die interne Entropie zu errechnen sind zunächst die Entropien innerhalb der drei Klassen zu berechnen. Man erhält:

$E_1 = ld(1) = 0$
$E_2 = ld(2) = 1$ und
$E_3 = ld(3) = log(3)/log(2) = 1{,}58496.$

Die interne Entropie ist dann

$E_{int} = \Sigma h_j E_j = 1/3 + (1/2)ld(3) = 1{,}9183.$

Die Summe aus externer und interner Entropie ist dann

$E_{ges} = (1/6)ld(6) + (1/3)ld(3) + 1/2 + 1/3 + (1/2)ld(3) = 2{,}58496 = ld(n) = ld(6).$

Satz 5.6:

Die Entropie E nimmt bei Gleichverteilung ($h_j = 1/m$, j= 1,2,...,m) der m Merkmalswerte ihren maximalen Wert ld(m) an.

Beweis:

Die Maximierung von E unter der Nebenbedingung $Sh_j = 1$ (mit j = 1, 2,...,m) läßt sich mittels der Lagrange-Funktion

$L = - \Sigma h_j \, ld(h_j) - \lambda (\Sigma h_j - 1)$ durchführen. Man erhält dann

(*) $\partial L/\partial h_j = -ld(h_j) - ld(e)/h_j - \lambda = 0$ (j=1,2,...,m) und

(**) $\partial L/\partial \lambda = \Sigma h_j - 1 = 0$

Gemäß den m Gleichungen (*) muss für λ stets gelten: $\lambda = -ld(h_j) - ld(e)/h_j$, was nur möglich ist, wenn alle h_j gleich sind. In Verbindung mit (**) folgt daraus, dass für alle j gelten muss $h_j = 1/m$. Aus den Bedingungen zweiter Ordnung folgt, dass an der Stelle ($h_1 = 1/m$, $h_2 = 1/m$,... , $h_m = 1/m$) in der Tat das Maximum von E liegt. Die Entropie nimmt
dann den Wert $-ld(1/m) = ld(m)$ an.

Exkurs: Dispersionsindex und Diversität[2]

Die Anzahl möglicher Paarvergleiche einer Einheit mit $x = x_i$ mit jeweils einer Einheit mit $x =,/ x_i$ ist $n_i(n-n_i)$. Folglich ist die Anzahl der Paare bei denen sich beide Beobachtungswerte unterscheiden $\Sigma n_i(n-n_i)$ und infolge von Satz 5.6 ist dieser Ausdruck dann maximal, wenn für alle i gilt:

$$n_i = \frac{n}{m} = k \qquad \text{und} \qquad \Sigma k = mk = n^2 \frac{m-1}{m},$$

also eine Rechteckverteilung vorliegt, so dass man mit

$$(5.49) \quad S_D = \frac{m}{(m-1)n^2} \sum_{i=1}^{m} n_i (n-n_i) = \frac{m}{m-1} (1- \Sigma h_i^2)$$

den **Dispersionsindex** von Hammond und Householder (1962) erhält, der lediglich eine **Nominalskala** voraussetzt. Offenbar ist $S_D = 0$, wenn eine Einpunktverteilung vorliegt und $S_D = 1$, wenn alle Häufigkeiten n_i gleich sind, so dass $0 \le S_D \le 1$ gilt.

Man beachte, dass hier ein Konzept der Variabilität vorliegt, dass keinen Abstandsbegriff voraussetzt, sondern sich, ähnlich wie die Entropie, nur daran orientiert, in welchem Maße bestimmte Merkmalsausprägungen gehäuft auftreten. S_D nimmt aber auch nicht Bezug auf den Modus. Das Maß S_D (oder D) ist geeignet um **Strukturen** (Gliederungen nach einem qualitativen Merkmal) zu beschreiben hinsichtlich der Abweichung von einer Rechteckverteilung, z.B. die Arbeitsteilung und deren Veränderung oder die unterschiedliche Struktur der Warenkörbe bei Preisindizes.

---

[2] Hinweise auf die in diesem Abschnitt behandelten Streuungsmaße verdanke ich Herrn Prof. Dr. Piesch.

Im Falle einer **Ordinalskala** sind ebenfalls keine Abstände definiert, auf die ein Streuungsmaß Bezug nehmen könnte, wohl aber die Summenhäufigkeit. Darauf beruht die Diversität.

Für die lediglich eine **Ordinalskala** voraussetzende **Diversität** S_{D^*} oder D^* gilt:

$$(5.50) \quad S_{D^*} = D^* = \frac{4}{m\text{-}1} \sum_{i=1}^{m-1} H_i\,(1\text{-}H_i)$$

wobei H_i die Summenhäufigkeit darstellt.

Bei einer Rechteckverteilung ($n_i = n/m$ für alle $i = 1,2,...,m$) erreicht der Ausdruck $2\sum_{i=1}^{m-1} H_i\,(1\text{-}H_i)$ mit $H_i = \frac{i}{m}$ und $H_m = 1$ den Wert $\frac{m^2\text{-}1}{3m}$. Bei einer Nominalskala ist dies der maximale Wert. Im Falle einer Ordinalskala kann man allerdings von einer extremeren Situation mit $h_1 = 1/2$, $h_2 = h_3 = ... = h_{m-1} = 0$ und $h_m = 1/2$ ausgehen. Man erhält dann bei geradzahligem m für $2\sum_{i=1}^{m-1} H_i\,(1\text{-}H_i)$ den Wert $\frac{m\text{-}1}{2}$, was größer ist als $\frac{m^2\text{-}1}{3m}$ $= \frac{m\text{-}1}{2} \cdot \frac{2(m+1)}{3m}$ sobald $m > 2$. So erklärt sich Gl. 5.50

4. Maße der relativen Streuung

Mit Def. 5.1 ist die relative Streuung definiert als Relation mit einem Maß der absoluten Streuung im Zähler und ein (hierzu passender) Mittelwert im Nenner. Der Vorteil dieses Verhältnisses ist vor allem die Maßstabsunabhängigkeit der so gemessenen Streuung, so dass damit auch Streuungen verschiedener Häufigkeitsverteilungen vergleichbar sind. Das bei weitem bekannteste Maß der relativen Streuung ist der Variationskoeffizient (Def. 5.7), der auf der Basis der Standardabweichung konstruiert ist. Man kann auch andere Streuungsmaße zur Bildung von Maßen der relativen Streuung heranziehen, z.B. den mittleren Quartilsabstand beim Quartilsdispersionskoeffizienten (Def. 5.8).

Def. 5.7: Variationskoeffizient

Der folgende durch Gl. 5.51 definierte Ausdruck V ist bekannt als Variationskoeffizient:

$$(5.51) \quad V = \frac{s}{\bar{x}} \quad .$$

Bemerkungen zu Def. 5.7:

1. Der mit 100 multiplizierte Wert von V drückt die Streuung (gemessen als Standardabweichung) in Prozent des Durchschnitts (arithmetischen Mittels) aus. Die Standardabweichung (s) mißt zwar eine Art durchschnittlichen Abstand zwischen den Beobachtungswerten und dem arithmetischen Mittel $\bar{x}$, sie wird aber nicht ins Verhältnis zu $\bar{x}$ gesetzt, was dazu führt, dass die Streuungen von zwei verschiedenen Beobachtungsreihen nicht miteinander verglichen werden können. Dies folgt auch daraus, dass die Standardabweichung s nicht maßstabsunabhängig ist (nach Axiom S4). Zur Herstellung einer Vergleichbarkeit wird die Standardabweichung s in Beziehung zu $\bar{x}$ gesetzt.

   Eine Standardabweichung von 5 bei einem $\bar{x}$ von 500 erscheint nicht sehr hoch. Ist $\bar{x}$ allerdings nur 10, dann ist die Standardabweichung von 5 sehr hoch. Deswegen ist es bei einem Vergleich von Streuungen verschiedener Beobachtungsreihen sinnvoll s auf $\bar{x}$ zu beziehen bzw. den Variationskoeffizienten zu berechnen, der im ersten Fall 1% und im zweiten 50% beträgt.

2. Der Variationskoeffizient lässt sich sinnvoll nur für verhältnisskalierte (ratioskalierte) Merkmale, mit einem positiven Mittelwert interpretieren.

3. Die Dimensionslosigkeit des Variationskoeffizienten ermöglicht auch einen Vergleich von Verteilungen mit verschiedenen Maßeinheiten (km, DM, Jahr).

4. Der Variationskoeffizient kann auch als Maß der Ungleichheit (Disparität) aufgefaßt werden (Kap. 6). Das folgende Beispiel zeigt auch, wie sich der Variationskoeffizient durch eine Lineartransformation verändert.

Beispiel 5.16:

Wie ändert sich in Aufg. 5.3 der Variationskoeffizient? Interpretieren Sie das Ergebnis!

Lösung 5.16:

In Aufgabe 5.3 erhielt man: $\bar{x} = 2200$, $s_x = 800$ so dass $V_x = 0{,}364$ ist, $\bar{y} = 2500$, $s_y = 880$, womit sich V_y errechnet zu $V_y = 0{,}352$.

Der Variationskoeffizient verringert sich also, was allein auf den festen "Sockelbetrag" von 80,- DM für das Urlaubsgeld zurückzuführen ist. Das arithmetische Mittel vergrößert

sich um 13,64%. Die Standardabweichung, auf die allein die "lineare" Erhöhung aller Gehälter um 10%, nicht aber der Sockelbetrag einen Einfluß hat, vergrößert sich dagegen nur um 10%.

Def. 5.8: Quartilsdispersionskoeffizient

Setzt man den mittleren Quartilsabstand $\bar{Q}_{0,25} = ½(Q_3 - Q_1)$ als Maß der absoluten Streuung ins Verhältnis zum Wert $½(Q_1 + Q_3)$, den man als eine Art Mittelwert interpretieren kann, so erhält man QD, den Quartilsdispersionskoeffizient

$$(5.52) \quad QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} .$$

Bemerkungen zu Def. 5.8:

1. Der Quartilsdispersionskoeffizient kann auch mit dem Median ($\tilde{x}_{0,5} = Q_2$) berechnet werden, man erhält dann

   $$(5.52a) \quad QD^* = \frac{Q_3 - Q_1}{Q_2} .$$

2. Auf der Basis des Medians lassen sich auch andere Maße der relativen Streuung konstruieren, etwa

   $$(5.52b) \quad RD = \frac{d_x}{Q_2} = \frac{d_x}{\tilde{x}_{0,5}}$$

   eine relativierte durchschnittliche Abweichung.

3. Die relative Streuung QD ist nicht zu verwechseln mit einem auf Quartile beruhenden Schiefemaß.

5. Momente

Momente sind sehr allgemeine Kennzahlen von Verteilungen. Viele Maßzahlen von Häufigkeitsverteilungen können als spezielle Momente angesehen werden. Übersicht 5.2 stellt ausgehend vom Begriff des Moments um a (vgl. Def. 5.9) die Zusammenhänge zwischen Momenten verschiedenen Typs dar.
Verteilungen lassen sich außer durch Lage- und Streuungsmaße auch durch andere Gestaltparameter, wie Schiefe und Wölbung charakterisieren. Diese Kennzahlen werden i.d.R. aus Momenten abgeleitet.

Def. 5.9: Momente

a) Momente sind Mittelwerte der (k-ten Potenz der lineartransformierten) Größen

$$\left[\frac{x_v - a}{b}\right]^k$$

mit b = s (Standardabweichung) und a = $\bar{x}$ (arithmetisches Mittel) erhält man ***standardisierte Momente***.

b) Mit b = 1 und der beliebigen reellen Konstanten a erhält man das k-te ***Moment um a***:

- bei Einzelwerten (ungewogene Berechnung)

$$(5.53) \quad m_{k(a)} = \frac{1}{n} \Sigma (x_v - a)^k \qquad \text{[k-tes Moment um a] und}$$

- bei Häufigkeitsverteilungen (gewogene Berechnung)

$$(5.54) \quad m_{k(a)} = \frac{1}{n} \Sigma (x_i - a)^k n_i = \Sigma (x_i - a)^k \, h_i \, .$$

c) Spezialfälle: ***Anfangs-Momente*** (oder Momente um Null) und ***zentrale Momente*** sind Spezialfälle des Moments um a (Übers. 5.2).

d) Von geringerer Bedeutung sind absolute Momente: analog Gl. 5.53 ist das k-te ***absolute Moment*** um a definiert als

$$(5.53a) \quad m^*_{k(a)} = \frac{1}{n} \Sigma |x_v - a|^k \qquad \text{[k-tes absolutes Moment um a].}$$

Bei einer geraden Zahl k sind die absoluten Momente gleich den "gewöhnlichen Momenten" [= Momente im Sinne von b) bzw. c)]

e) In der Induktiven Statistik spielen auch ***faktorielle Momente*** eine Rolle; $\frac{1}{n} \Sigma x_v (x_v - 1)(x_v - 2) \ldots$.

Bemerkungen zu Def. 5.9:

1. Man sieht leicht, dass das erste Anfangsmoment m_1 das arithmetische Mittel $\bar{x}$ ist und dass das zweite zentrale Moment z_2 die Varianz s^2 ist.

2. Das dritte zentrale Moment hat einen Zahlenwert von Null, wenn eine symmetrische Verteilung vorliegt. Es hat einen positiven Wert bei einer linkssteilen Verteilung und ist negativ bei einer rechtssteilen Verteilung, weshalb z_3 zur Konstruktion von Maßzahlen der **Schiefe** (Asymmetrie) verwendet wird. Dieser Zusammenhang gilt für alle ungeraden zentralen Momente, bis auf das erste z_1, das wegen der Schwerpunkteigenschaft von $\bar{x}$ immer Null ist.

3. Das vierte zentrale Moment z_4 charakterisiert die **Wölbung** einer Verteilung. Die Werte dieses, sowie aller anderen höheren geraden Momente nehmen nämlich um so mehr zu, je höher die Wölbung einer Verteilung ist.

4. Zentrale Momente und Anfangsmomente verhalten sich bei einer proportionalen Transformation der Variablen X zur Variablen Y mit $y_v = bx_v$ $(b \neq 1)$ wie folgt:

$$(5.56) \quad z_{k(y)} = b^k \cdot z_{k(x)} \; .$$

Das folgt aus $z_{k(y)} = \frac{1}{n} \Sigma (bx_v - b\bar{x})^k = \frac{1}{n} \Sigma b^k (x_v - \bar{x})^k = b^k [\frac{1}{n} \Sigma x_v - \bar{x}^k]$.

5. Momente sind nicht unabhängig vom Ursprung der Skala für die Variable X. Deshalb unterscheiden sich auch Momente um a, Anfangsmomente und zentrale Momente. Anders dagegen die sog. **Kumulanten** k_r: außer k_1 sind sie invariant gegenüber Translationen (Verschiebungen des Nullpunkts) gem. $y_v = a + x_v$. Bei proportionalen Transformationen ($y_v = bx_v$) gilt Gl. 5.56.

6. Unter bestimmten, in der Praxis fast immer gegebenen Bedingungen ist eine Häufigkeitsverteilung durch die Folge ihrer Momente eindeutig bestimmt.

7. Bei zwei und mehr Variablen ist das **Produktmoment** eine Verallgemeinerung. Bei zwei Variablen X und Y ist das zentrale Produktmoment bekannt als Kovarianz und das standardisierte Produktmoment als Korrelationskoeffizient (vgl. Kap. 7).

8. Zum Begriff "Moment": die Analogie zur physikalischen Terminologie ist berechtigt. Die Varianz s^2 ist in der Tat das Trägheitsmoment bei Rotation um das arithmetische Mittel, wobei dieses als Schwerpunkt (in einer Dimension) der Quotient aus Summe der Momente ($\Sigma x_j h_j$, also Kraft h_j mal Hebelarm x_j) und Summe der Gewichte (Σh_j) ist.

Übersicht 5.2: Momente

allgemein: Mittelwerte der Größen

$$\left[\frac{x_v - a}{b}\right]^k$$

Spezialfälle

b = 1 a als beliebige reelle Konstante *Momente um a*	b = s a = $\bar{x}$ *standardisierte Momente* ms_k
(5.53) $m_{k(a)} = \frac{1}{n}\sum_{v=1}^{n}(x_v - a)^k$	$ms_1 = 0$; $ms_2 = 1$
[k-tes Moment um a] *)	(Mittelwert und Varianz einer standardisierten Variabel sind 0 und 1)

Anfangsmoment a = 0	*zentrales Moment* a = $\bar{x}$
(5.54) $m_k = \frac{1}{n}\sum x_v^k$	(5.55) $z_k = \frac{1}{n}\sum(x_v - x,^-)^k$
(5.54a) $m_k = \sum x_i^k h_i$	(5.55a) $z_k = \sum(x_i - x,^-)^k h_i$

weitere Spezialfälle:

$m_0 = 1$	$z_1 = 0$ (Schwerpunkteigenschaft!)
$m_1 = \bar{x}$	$z_2 = s^2$ (Varianz)

*)(ungewogene Berechnung, die gewogene Berechnung erfolgt analog, vgl. Def. 5.9)

Zusammenhänge zwischen Anfangs- und zentralen Momenten

$z_2 = m_2 - (m_1)^2$ (Verschiebungssatz für die Varianz).Analog folgt:

$z_3 = m_3 - 3m_1m_2 + 2(m_1)^3$ $\quad z_4 = m_4 - 4m_1m_3 + 6(m_1)^2m_2 - 3(m_1)^4$ usw.

$$\text{allgemein: } z_k = \sum_{i=0}^{k} \binom{k}{i} m_{k-i}\,(-m_1)^i$$

Daraus folgt: Anfangsmomente sind nicht verschiebungsinvariant

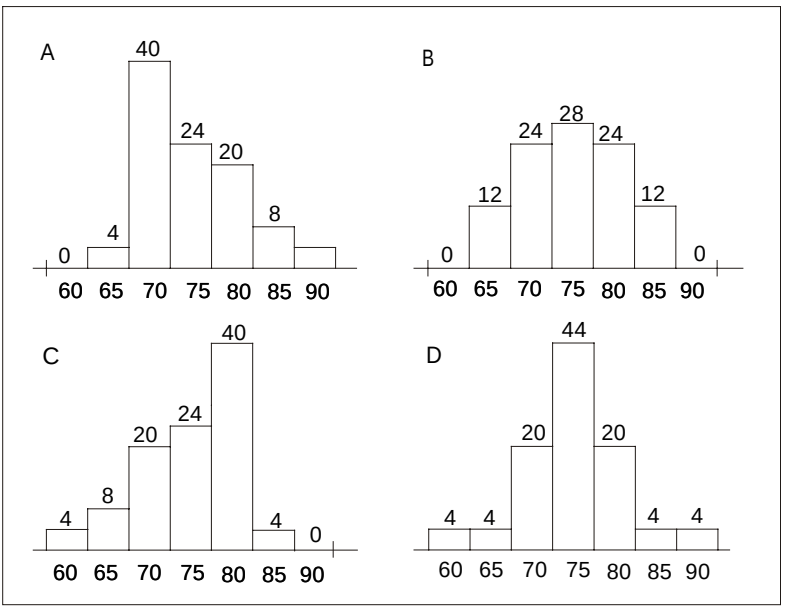
$m_i(c) = \sum (x_v + c)^i \neq m_i = \sum x_v^i$ wohl aber zentrale Momente

$z_i(c) = \sum \left[(x_v + c) - (\bar{x} + c)\right]^i = z_i$, denn $m_1(c) = \bar{x} + c$.

Beispiel 5.17:

Die folgenden vier Häufigkeitsverteilungen (A,B,C,D) haben jeweils das gleiche arithmetische Mittel ($\bar{x}$ = 75) und die gleiche Varianz (s^2 = 36). Gleichwohl ist ihre Gestalt hinsichtlich Schiefe und Wölbung sehr unterschiedlich (vgl. Abb. 5.4). Dies wird auch deutlich, wenn man die dritten und vierten zentralen Momente berechnet (Beispiel entnommen aus K.Stange, Angewandte Statistik, Bd. 1, S.87). Die Gesamthäufigkeit n ist jeweils 100, so dass man die relativen Häufigkeiten leicht ablesen kann.

Abb. 5.4: Häufigkeitsverteilungen des Bsp.5.17



x_j	absolute Häufigkeit			
	A	B	C	D
60	0	0	4	4
65	4	12	8	4
70	40	24	20	20
75	24	28	24	44

x_j	absolute Häufigkeit			
	A	B	C	D
80	20	24	40	20
85	8	12	4	4
90	4	0	0	4

Berechnen Sie die dritten und vierten und zentralen Momente für die vier Häufigkeitsverteilungen!

Lösung 5.17:

Verteilung	A	B	C	D
z_3	150	0	-150	0
z_4	3600	2700	3600	5100

6. Schiefemaße

a) Begriff der Schiefe

Mit der Schiefe (skewness) soll der Grad der Asymmetrie einer Häufigkeitsverteilung gemessen werden. Schiefe ist die Abweichung von der symmetrischen Verteilung eines metrisch skalierten Merkmals. Asymmetrie (= Schiefe) hat zwei Formen: Linkssteilheit und Rechtssteilheit. Die folgenden Begriffe werden synonym verwendet:

linkssteil	=	rechtsschief
rechtssteil	=	linksschief .

Linkssteilheit bedeutet, dass sich die Masse der Merkmalsträger am unteren Ende einer Häufigkeitsverteilung konzentriert. Sie wird auch oft ähnlich interpretiert wie "Ungleichheit" (Disparität), es gibt jedoch Unterschiede zur Disparität im Sinne der Statistik (vgl. Kap. 6).
Die Abb. 5.4 veranschaulicht die Begriffe Symmetrie, Linkssteilheit und Rechtssteilheit. Die Verteilungen B und D sind symmetrisch (sie unterscheiden sich jedoch durch die Wölbung), die Verteilung A ist linkssteil (positive Schiefe) und die Verteilung C ist rechtssteil (negative Schiefe).
Schiefe kann auch für die Datenanalyse als störend empfunden werden: es kann dann z.B. schwierig zu entscheiden sein, welcher Lageparameter das Niveau einer Verteilung (die Größenordnung der Merkmalswerte) ange-

messen beschreibt und welche Beobachtung als Ausreißer anzusehen ist. Deshalb gilt es nicht nur, die Schiefe zu messen, sondern auch gegebenenfalls Methoden anzuwenden, um diese zu beseitigen (Abschn. c).

Viele natürliche Erscheinungen sind symmetrisch verteilt (z.B. Körpergröße, Körpergewicht aber auch z.B. der Intelligenzquotient), während "soziale" Erscheinungen häufig linkssteil verteilt sind. Das gilt besonders für Einkommen und Vermögen. Es gibt zahlreiche Modelle zur Erklärung einer linkssteilen Einkommensverteilung.
Nach dem zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung ist eine Größe X, die sich als Summe sehr vieler stochastisch unabhängiger (zum Begriff der Unabhängigkeit vgl. Kap. 7) Einflußfaktoren u_j darstellen läßt $X = u_1 + u_2 + u_3 + ... + u_n = \Sigma u_j$, asymptotisch (mit wachsendem n) normalverteilt. Ist Z dagegen ein Produkt solcher Einflussfaktoren v_j also $Z = v_1 v_2 v_3 ... v_n$ so ist Z asymptotisch logarithmisch-normalverteilt. Die Normalverteilung ist eine symmetrische Verteilung, die logarithmische Normalverteilung eine linkssteile Verteilung. Man spricht auch vom "Gesetz der proportionalen Effekte", d.h. die Einzeleinflüsse verstärken sich gegenseitig, bzw. schwächen sich gegenseitig ab, in der Art, wie es im Matthäus-Evangelium beschrieben wird:

> *"Denn wer da hat, dem wird gegeben, dass er die Fülle habe, wer aber nicht hat, von dem wird auch genommen, das er hat" (Matth.13, Vers 12).*

Linkssteilheit und Rechtssteilheit sind zwei entgegengerichtete Abweichungen von der Symmetrie. Wenn eine Verteilung nicht symmetrisch ist, dann kann sie nur entweder linkssteil oder rechtssteil sein. Schiefemaße sollen Richtung und Ausgeprägtheit der Abweichung von der Symmetrie messen. Zu einer exakteren Begriffsbildung gelangt man durch die folgende Definition:

Def. 5.10: Achsensymmetrie

Die Häufigkeitsverteilung des metrisch skalierten Merkmals X heißt symmetrisch bezüglich des Medians $\tilde{x}_{0,5}$, falls für **alle** Werte einer reellen Konstante c gilt

$$h(\tilde{x}_{0,5} - c) = h(\tilde{x}_{0,5} + c) \quad c > O \; . \tag{5.57}$$

Dabei ist $h(\tilde{x}_{0,5} - c)$ die relative Häufigkeit der Merkmalsausprägung $x_c = \tilde{x}_{0,5} - c$ und $h(\tilde{x}_{0,5} + c)$ ist entsprechend definiert. Eine Verteilung ist schief oder asymmetrisch, wenn Gl. 5.57 nicht gilt.

Bemerkungen zu Def. 5.10

1. Die Definition ist eine exakte Beschreibung der Bedingungen, unter denen eine Häufigkeitsverteilung achsensymmetrisch um den Median ist. Sie ist nicht notwendig auch operational zur Herleitung eines Schiefemaßes. Nur dann, wenn eine Häufigkeitsverteilung einen Me-

dian dergestalt besitzt, dass die möglichen Ausprägungen der Variable X jeweils links und rechts gleich weit entfernt von $\tilde{x}_{0,5}$ liegen, kann Symmetrie gem. Gl. 5.57 definiert und verifiziert werden (Bsp. 5.18).

2. Man beachte, dass Gl. 5.57 für die kumulierten Häufigkeiten impliziert:

(5.58)	$h(x \le \tilde{x}_{0,5} - c) = h(x \ge \tilde{x}_{0,5} + c) \quad (c \ne 0)$ oder äquivalent: $H(\tilde{x}_{0,5} - c) = 1 - H(\tilde{x}_{0,5} + c)$ für jedes $c \ge 0$.

Die Größe H(x) ist die bis x erreichte kumulierte relative Häufigkeit (Fläche unter der Häufigkeitsverteilung). Beispiel 5.19 zeigt, dass Achsensymmetrie Flächengleichheit links und rechts von bestimmten Punkten unter der Häufigkeitsverteilung impliziert, nicht aber umgekehrt (aus Flächengleichheit [bei ausgewählten Intervallen] folgt nicht Achsensymmetrie).
Gl. 5.58 führt zu der Def. 5.11, die zwar bei einer stetigen Verteilung keine Schwierigkeiten bereiten würde, im diskreten Fall aber nicht anwendbar sein kann (Bsp. 5.19).

3. Im Satz 5.7 wird gezeigt, dass die Definition der Symmetrie mit Def. 5.10 - als Achsensymmetrie um $\tilde{x}_{0,5}$ bzw. um $\bar{x}$ - im Einklang steht mit einer Momentschiefe von Null.

4. Ausgehend von der Symmetrie im Sinne von Gl. 5.57 (Achsensymmetrie) könnte man definieren, dass Linkssteilheit dann entsteht, wenn ein Merkmalsträger mit dem Merkmalswert $\bar{x} + c$ (mit $c > 0$) ausgetauscht wird gegen einen Merkmalsträger mit dem Wert $\bar{x} - c$. Entsprechend wäre Rechtssteilheit zu definieren. Bei dieser Art, eine linkssteile Verteilung zu "erzeugen" verringert sich jedoch das arithmetische Mittel $\bar{x}$ zum neuen Wert $\bar{x} - 2c/n$. Entsprechend vergrößert sich das arithmetische Mittel beim Übergang zu einer rechtssteilen Verteilung zu $\bar{x} + 2c/n$.

Beispiel/Lösung 5.18:

Ein Beispiel für die Anwendbarkeit von Def. 5.10 wäre eine Gesamtheit mit den n = 9 Beobachtungen 10,15,15,20,20,20,25,25,30. Der Median ist 20 und die Häufigkeitsverteilung ist nach Def. 5.10 symmetrisch, da bei c = 5 gilt:

$$h(15) = h(\tilde{x}_{0,5} - 5) = h(\tilde{x}_{0,5} + 5) = h(25) = 2/9$$

und bei c = 10 ist entsprechend

$$h(\tilde{x}_{0,5} - 10) = h(\tilde{x}_{0,5} + 10) = 1/9.$$

Die Definition wäre aber z.B. nicht anwendbar auf eine Häufigkeitsverteilung mit dem Median 20, wenn zwar die Beobachtung x=17 nicht aber x=23 vorkommt.

Beispiel 5.19:

Die Einzelbeobachtungen 10,16,20,20 bilden eine Häufigkeitsverteilung, die nach anschaulicher Vorstellung nicht symmetrisch, sondern rechtssteil ist. Man überprüfe die Symmetriedefinition gem. Gl.5.58 und berechne das dritte zentrale Moment!

Lösung 5.19:

Das arithmetische Mittel beträgt $\bar{x} = 16{,}5$ und der Median (mit Interpolation) $\tilde{x}_{0,5} = 18$. Gilt Gl. 5.58 für **jedes** c, so müßte es auch für c = 0 gelten. Man erhält damit aber die für den Zentralwert stets erfüllte Gleichung: $h(x < \tilde{x}_{0,5}) = h(x > \tilde{x}_{0,5}) = ½$, natürlich auch hier, ohne dass deshalb die Verteilung symmetrisch wäre.
Symmetrie wäre nach Gl. 5.58 auch angezeigt für c = 1, denn

$$h(x \le \tilde{x}_{0,5} - 1) = h(x \le 17) = h(x \ge \tilde{x}_{0,5} + 1) = h(x \ge 19) = ½.$$

Die Bezugnahme auf Flächengleichheit zur Definition der **Symmetrie** (vgl. Def. 5.11) macht nur Sinn, wenn Gl. 5.58 für **jedes** c gilt. In diesem Beispiel gilt Gl. 5.58 z.B. nicht für c = 4. Denn

$$h(x \le \tilde{x}_{0,5} - 4) = h(x \le 14) = 1/4 > h(x \ge \tilde{x}_{0,5} + 4) = h(x \ge 22) = 0,$$

was nach Gl. 5.58b (Def. 5.11) auf Rechtssteilheit hinweist, die [gemessen an der Momentschiefe] auch tatsächlich gegeben ist.
Das dritte zentrale Moment beträgt hier nämlich

$$z_3 = ¼ [(10 - 16{,}5)^3 + (16 - 16{,}5)^3 + 2(20 - 16{,}5)^3] = -189/4 = -47{,}25.$$

Dieses Beispiel macht den Hintergrund der folgenden Def. 5.11 deutlich.

Def. 5.11: Schiefe

Eine Verteilung ist symmetrisch wenn **für alle** c gilt

(5.58) $h(x \le \tilde{x}_{0,5} - c) = h(x \ge \tilde{x}_{0,5} + c)$,
oder äquivalent: $H(\tilde{x}_{0,5} - c) = 1 - H(\tilde{x}_{0,5} + c)$.

und sie ist entsprechend

linkssteil wenn (5.58a)	$h(x \le \tilde{x}_{0,5} - c) < h(x \ge \tilde{x}_{0,5} + c)$
rechtsteil wenn (5.58b)	$h(x \le \tilde{x}_{0,5} - c) > h(x \ge \tilde{x}_{0,5} + c)$

für **mindestens ein c** gilt.

Bemerkungen zu Def.5.11:

Zur Erläuterung vgl. Bsp. 5.19. Diese Definition der Schiefe legt den Gedanken nahe, dass bestimmte Quantile jeweils vom Median gleich weit entfernt sein müssen, wenn Symmetrie herrscht. So ist z.B. für die Abstände von Quartilen zu fordern

bei Symmetrie: $Q_3 - Q_2 = Q_2 - Q_1$
bei Linkssteilheit: $Q_3 - Q_2 > Q_2 - Q_1$ und
bei Rechtssteilheit: $Q_3 - Q_2 < Q_2 - Q_1$.

Auf diesem Gedanken beruhen der Quartils- und [allgemeiner] der Quantilskoeffizient als Schiefemaß (vgl. Def. 5.12, Teil b).

Satz 5.7:

a) Bei einer Häufigkeitsverteilung, die im Sinne der Def. 5.10 symmetrisch ist gilt für das dritte zentrale Moment $z_3 = 0$.

b) Wird ausgehend von der symmetrischen Verteilung
mit $2(h_o+h_1) = 1$ und $z_3 = 0$

x_i	h_i
$\bar{x} - c$	h_1
$\bar{x}$	$2h_0$
$\bar{x} + c$	h_1

die relative Häufigkeit a (mit$0 < a < h_1$)
1. von $\bar{x}$+c auf $\bar{x}$-c
2. von $\bar{x}$-c auf $\bar{x}$+c
"verlagert" so entsteht eine linkssteile (Fall 1) bzw. rechtssteile Verteilung (Fall 2):

Fall 1:

x_i	h_i
$\bar{x} - c$	$h_1 + \alpha$
$\bar{x}$	$2h_o$
$\bar{x} + c$	$h_1 - \alpha$

Fall 2:

x_i	h_i
$\bar{x} - c$	$h_1 - \alpha$
$\bar{x}$	$2h_o$
$\bar{x} + c$	$h_1 + \alpha$

mit den folgenden Momenten
1) Mittelwert
$\bar{x}^* = \bar{x} - 2c\alpha$ (Fall 1) $\qquad \bar{x}^{**} = \bar{x} + 2c\alpha$ (Fall 2)
2) drittes zentrales Moment:
$z_3^* = 2c^3\alpha(6h_1 - 1 - 8\alpha^2) \qquad z_3^{**} = - z_3^*$

$z_3^* > 0$ (linkssteil) $z_3^{**} < 0$ (rechtssteil)

Beweis: Elementar durch Ausmultiplizieren. Die Ausdrücke für die Momente werden im folgenden in einem Beispiel verifiziert.

Beispiel 5.20

Ausgehend von der symmetrischen Verteilung (mit c = 10, $\bar{x}$ = 20, $h_o = h_1 = 1/4$) soll mit
α = 0,05 die im obigen Satz beschriebene "Erzeugung" einer links- bzw. rechtsteilen Verteilung demonstriert werden. Wie hängt die Schiefe vom Parameter α ab?

Lösung 5.20

Ausgangsverteilung

x_i	h_i
$\bar{x}$-c=10	h_1=1/4
$\bar{x}$=20	$2h_o$=1/2
$\bar{x}$+c=30	h_1=1/4

Fall 1

x_i	h_i
10	0,3
20	0,5
30	0,2

Fall 2

x_i	h_i
10	0,2
20	0,5
30	0,3

das arithmetische Mittel und das dritte zentrale Moment z_3 sind dann jeweils:

$\bar{x} = 20$ $\bar{x}^* = 19$ $\bar{x}^{**} = 21$

$z_3 = 0$ $z_3^* = 48$ $z_3^{**} = -48$

in der Tat ist mit $\bar{x}$=20, c=10, α=0,05
$\bar{x}^* = x,^- - 2c\alpha$ und $\bar{x}^{**} = \bar{x} + 2c\alpha$
$z_3^* = 2c^3\alpha(6h_1 - 1 - 8\alpha^2)$ und $z_3^{**} = - z_3,^*$.

Mit c = 10 und h_1 = ¼ gilt: $z_3^* = 2c^3\alpha(6h_1 - 1 - 8\alpha^2) = 2000\alpha(½ - 8\alpha^2)$
Für verschiedene Werte von α erhält man für z_3^*

α	0	0,05	0,10	0,15	0,20	0,25
$z_3,^*$	0	48	84	96	72	0

Die Funktion $z_3^* = 2000\alpha(½ - 8\alpha^2)$ hat ein Maximum an der Stelle
$\alpha = \sqrt{5/24} = 0{,}14434$ und beträgt an dieser Stelle $z_3^* = 96{,}225$.

Fechnersche Lageregel

Bei Häufigkeitsverteilungen gilt in der Regel, wenn sich die Daten über die x-Achse ausreichend dicht verteilen (und eine eingipflige Verteilung vorliegt) die Lageregel von Fechner:

- bei symmetrischer Verteilung gilt: $x = \tilde{x}_{0.5} = \bar{x}_M$
- bei linkssteiler (rechtsschiefer) Verteilung:
  arithmetisches Mittel ($\bar{x}$) > Median ($\tilde{x}_{0.5}$) > Modus ($\bar{x}_M$)
- bei rechtssteiler (linksschiefer) Verteilung:

arithmetisches Mittel ($\bar{x}$) < Median ($\tilde{x}_{0.5}$) < Modus ($\bar{x}_M$)
Also gilt

(5.59) linkssteil: $\bar{x}_M < \tilde{x}_{0.5} < \bar{x}$ rechtssteil: $\bar{x} < \tilde{x}_{0.5} < \bar{x}_M$.

Ausnahmen sind denkbar, insbesondere dann, wenn nur wenige Einzelwerte vorliegen (Bsp. 5.22). Auf der Basis der Fechnerschen Lageregel lassen sich auch Schiefemaße konstruieren.

Beispiel 5.21:

Man verifiziere die Fechnersche Lageregel für das Beispiel 5.17!

Lösung 5.21:

Im Bsp. 5.17 erhält man die folgenden Lageparameter

Verteilung	Modus	Median*)	arithmet. Mittel	Urteil**)
A	70	71,25	75	linkssteil $\bar{x}_M < \tilde{x}_{0.5} < \bar{x}$
B	75	75	75	symmetrisch
C	80	73,75	75	rechtssteil***)
D	75	75	75	symmetrisch

*) mit Interpolation.
**) mit der Fechnerschen Lageregel (Gl. 5.59); eine Beurteilung aufgrund der Momentschiefe erfolgt in Bsp. 5.25.
***) Allerdings ist hier der Median nicht größer, sondern kleiner als das arithmetische Mittel.

Beispiel 5.22:

Gegeben seien die Werte 10,16,17,20,22. Bestimmen Sie den Median und das arithmetische Mittel! Ist die Verteilung symmetrisch?

Lösung 5.22:

Es gilt $\bar{x} = \tilde{x}_{0,5} = 17$. Die Häufigkeitsverteilung ist gemessen an der Gleichheit von $\bar{x}$ und $\tilde{x}_{0,5}$ symmetrisch. Das dritte zentrale Moment beträgt aber $z_3 = -192/5 = -38,4$, so dass die Verteilung danach rechtssteil ist.

b) Schiefemaße

Angesichts der Schwierigkeiten, Schiefe zu definieren überrascht es nicht, dass sich Schiefemaße nicht auf eine anerkannte Axiomatik stützen

können und dass die Messung der Schiefe auf verschiedenen Konzepten beruht. Man kann von Schiefemaßen aber mindestens fordern, dass sie, wie auch andere Gestaltmaße (Formmaße), wie etwa die Streuung oder Wölbung

1. invariant sind gegenüber Translationen
2. die Richtung der Asymmetrie korrekt anzeigt: die Konvention über das Vorzeichen eines Schiefemaßes SK ist
   SK = 0 wenn die Verteilung symmetrisch ist
   SK > 0 wenn sie linkssteil ist (positive Schiefe)
   SK < 0 wenn sie rechtssteil ist (negative Schiefe).

Alle in Def. 5.12 präsentierten Schiefemaße erfüllen diese Forderungen. Schiefemaße sind in der Regel nicht beschränkt, so dass meist gilt:
$-\infty < SK < +\infty$.

Man kann Schiefemaße entwickeln auf der Basis

- der ungeraden zentralen Momente, wobei man zur Rechenvereinfachung von z_3 ausgeht (Konzept der Momentschiefe);
- der Abstände gewisser Lageparameter untereinander (nach dem in Def. 5.11 präsentierten Symmetriebegriff) oder auf der Basis
- der Fechnerschen Lageregel.

Def. 5.12: Schiefemaße

a) Die von Bowley und Fisher eingeführte **Momentschiefe** (Momentkoeffizient der Schiefe) lautet:

$$(5.60) \quad SK_M = \frac{z_3}{s^3} \qquad (\text{zu } z_3 \text{ vgl. Gl. 5.55}).$$

b) Als Quantilskoeffizient der Schiefe wird bezeichnet:

$$(5.61) \quad SK_{Q,\,p} = \frac{(\tilde{x}_{1-p} - Q_2) - (Q_2 - \tilde{x}_p)}{\tilde{x}_{1-p} - \tilde{x}_p} \quad (p < ½)$$

wobei $Q_2 = \tilde{x}_{0.5}$ = Median; der bekannteste spezielle Koeffizient (p = ¼) ist der Quartilskoeffizient der Schiefe (nach Yule und Bowley):

$$(5.62) \quad SK_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \quad .$$

c) Auf der Fechnerschen Lageregel beruhen die folgenden, von (Yule und) Pearson vorgeschlagenen Schiefmaße ($\bar{x}_M$ = Modus):

$$(5.63)\quad SK_{P1} = \frac{\bar{x} - \bar{x}_M}{s} \qquad \text{und} \qquad (5.64)\quad SK_{P2} = \frac{3(\bar{x} - \tilde{x}_{0,5})}{s}$$

Das zweite Pearsonsche Schiefemaße ist meist vorzuziehen, weil oft der Modus schwer zu bestimmen ist; (vgl. Bsp. 5.23).

Bemerkungen zu Def. 5.12

1. Die auf verschiedenen Konzepten beruhenden Schiefemaße sind nicht miteinander vergleichbar, d.h. es ist denkbar, dass eine Verteilung, die gemessen an der Momentschiefe linkssteil ist, nach anderen Schiefemaßen als symmetrisch oder rechtssteil beurteilt wird.

2. Liegt Symmetrie im Sinne der Def. 5.10 bzw. 5.11 vor, so sind alle Schiefemaße Null. Die Umkehrung dieses Satzes ist jedoch nicht zulässig.

3. Die Division durch s bzw. s^3 soll sicherstellen, dass das Schiefemaß dimensionslos ist, also nicht abhängig von der Maßeinheit der Variablen X ist.

4. Es gilt häufig mit guter Näherung $\bar{x} - \bar{x}_M \approx 3\,(\bar{x} - \tilde{x}_{0,5})$ worauf die Pearsonschen
Schiefemaße beruhen.

5. Als Schiefemaße werden in der Literatur auch Maßzahlen präsentiert, wie $(Q_3 + Q_1) / (Q_3 - Q_1)$, was nichts anderes ist als der reziproke Quartilsdispersionskoeffizient, oder die Größe $(Q_3 - Q_2)/(Q_2 - Q_1)$, die hinsichtlich des Vorzeichens offensichtlich die Anforderungen an ein Schiefemaß erfüllt.

6. Eine Variante des Quantilskoeffizienten der Schiefe ist:

$$(5.61a)\quad SK_{Q;0,2} = \frac{(\tilde{x}_{0,8} - Q_2) - (Q_2 - \tilde{x}_{0,2})}{\tilde{x}_{0,8} - \tilde{x}_{0,2}} \qquad (p = 0,2)$$

auf der Basis von Quintilen. Dieser Quintilskoeffizient der Schiefe ist nicht zu verwechseln mit der Quintilenschiefe q, die ein Disparitätsmaß (vgl. Kap. 6) ist. Sie lautet (in der Symbolik von Kap. 6)

$$(5.65)\quad q = \Sigma|h_j - 0,2|$$

und läßt sich mit den Angaben von Beispiel 5.24 errechnen (vgl. dort).

6. Aus der ersten Schreibweise von SK_Q in Gl. 5.62 ist bereits erkennbar, dass die Beziehung $-1 \le SK_{Q,p} \le +1$ gilt. Man kann ferner zeigen, dass gilt $-3 \le SK_{P2} \le +3$.

Wird p in $SK_{Q,p}$ nicht zu klein gewählt, so ist der mittlere Teil und nicht der extrem niedrige oder hohe Abschnitt der Häufigkeitsverteilung schiefebestimmend und $SK_{Q,p}$ ist dann relativ resistent gegenüber Ausreißern.

Beispiel 5.23:

Durch Variation des Beispiels 5.18 sind die folgenden drei Häufigkeitsverteilungen entstanden, die von A nach C [wie eine graphische Darstellung zeigen würde] im anschaulichen Sinne immer linkssteiler werden:

Verteilung A

x_j	n_j
15	3
20	4
25	1
30	1

Verteilung B

x_j	n_j
15	5
20	2
25	1
40	1

Verteilung C

x_j	n_j
15	6
30	1
60	1

Das arithmetische Mittel ist jeweils 20. Man bestimme die Momentschiefen SK_M der drei Häufigkeitsverteilungen!

Lösung 5.23:

Zur Berechnung des zweiten und dritten zentralen Moments bei der Verteilung A wird die folgende Arbeitstabelle aufgestellt:

$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$(x_j - \bar{x})^2 n_j$	$(x_j - \bar{x})^3$	$(x_j - \bar{x})^3 n_j$
15 - 20= - 5	25	75	-125	-375
+5	25	25	125	+125
+10	100	100	1000	+1000
		$\sum$ 200		$\sum$ 750

Daraus ergibt sich:

Verteilung A: $z_2 = s^2 = 200/9$, $z_3 = 750/9$, so dass $SK_M = +0{,}7955$

In entsprechender Weise errechnet man:

Verteilung B: $z_2 = s^2 = 550/9$, $z_3 = 7500/9$ und $SK_M = +1{,}7444$
Verteilung C: $z_2 = s^2 = 1850/8$, $z_3 = 64250/8$ und $SK_M = +2{,}2838$.

Die Schiefekoeffizienten sind im Einklang mit der anschaulichen Vorstellung, dass die Verteilungen von A über B nach C zunehmend linkssteiler werden.

Beispiel 5.24:

Gegeben ist die folgende Verteilung der Haushaltsnettoeinkommen*)

Einkommen		relative Häufigkeit		Anteil**)
von...bis unter...		h_i	H_i	q_i
0	1000	0,1	0,1	
1000	1400	0,1	0,2	0,1
1400	1500	0,05	0,25	
1500	2000	0,15	0,4	0,12
2000	2400	0,1	0,5	
2400	2600	0,1	0,6	0,2
2600	3400	0,15	0,75	
3400	3800	0,05	0,8	0,24
3800	4800	0,1	0,9	
über 4800		0,1	1	0,34

*) Die Zahlenangaben sind so gewählt, dass sich bestimmte Quantile leicht bestimmen lassen. Damit ergibt sich eine stark vereinfachte Darstellung der Einkommensverteilung, wie sie in dieser Form 1988 in der Bundesrepublik bestand.

**) Anteil des ersten,...,fünften Quintils am Gesamteinkommen aller Haushalte.

Berechnen Sie den Quartils- und den Quintilskoeffizient der Schiefe, die Quintilenschiefe sowie die Schiefemaße von Pearson! Das arithmetische Mittel beträgt 2650 DM.

Lösung 5.24:

a) Quartilskoeffizient der Schiefe: Quartile $Q_1 = 1500$, $Q_2 = 2400$ und $Q_3 = 3400$; damit ist der Koeffizient (1000-900)/(1000+900) = +0,053.

b) Quintilskoeffizient der Schiefe:
Quintile: $Q_1^* = 1400$, $Q_4^* = 3800$, so dass man für den Koeffizienten erhält [(3800-2400)-(2400-1400)]/(3800+1400) = +0,167.

c) Quintilenschiefe: Es sind die absoluten Abweichungen der Größen 0,1; 0,12; 0,20; 0,24 und 0,34 von 0,2 zu bilden und zu addieren. Das Ergebnis ist dann q = 0,36 (ein Maß der Disparität [vgl. Kap.6]).

d) Schiefemaße von Pearson: der Modus ist hier schwer zu bestimmen, weil er abhängig ist von der Klasseneinteilung; sinnvoller ist es SK_{P2} zu bestimmen (allerdings ist aus den Angaben auch die Standardabweichung nicht zuverlässig zu bestimmen). Der Zähler von SK_{P2} ist positiv, weil der Zentralwert (Median) mit $\tilde{x}_{0,5} = 2400$ kleiner ist als das arithmetische Mittel (das mit $\bar{x} = 2650$ angegeben ist). Auch nach der Lageregel von Fechner ist die Verteilung linkssteil.

Exkurs: Schiefediagramm

Ein Schiefediagramm ist eine graphische Darstellung zur Beurteilung der Asymmetrie, was evtl. aufschlußreicher sein kann als die Berechnung einer summarischen Maßzahl. Ausgehend vom Median $\tilde{x}_{0,5} = Q_2$ werden die Abstände ausgewählter Quantile vom Median in einem rechtwinkligen Koordinatensystem wie folgt eingetragen:

Abszisse $\tilde{x}_{1-p} - \tilde{x}_{0,5}$ Ordinate $\tilde{x}_{0,5} - \tilde{x}_p$.

Mit dem Beispiel 5.25 soll die Konstruktion des Schiefediagramms demonstriert werden. Dieses Beispiel ist kennzeichnend für eine ausgesprochen rechtssteile Verteilung, bei der die Punkte überwiegend (in Abb. 5.5 sogar alle Punkte) rechts unterhalb der 45º-Linie liegen. Die 45º-Linie ist Ausdruck der Gleichheit $\tilde{x}_{1-p} - \tilde{x}_{0,5} = \tilde{x}_{0,5} - \tilde{x}_p$, also der Symmetrie gem. Def.5.11 (vgl.Bem. 2 zu Def. 5.11). Im Falle einer linkssteilen Verteilung liegen die Punkte dagegen überwiegend oberhalb der 45°-Linie, da für viele p gilt:

$\tilde{x}_{1-p} - \tilde{x}_{0,5} < \tilde{x}_{0,5} - \tilde{x}_p$.

Beispiel 5.25:

Man bestimme das Schiefediagramm für ausgewählte Werte von p für das Beispiel 5.6!

Lösung 5.25:

Die vorgegebenen Beobachtungswerte des Beispiels (21, 25, 34, 39, 43, 52, 64, 72, 80) sowie bestimmte Quantile sind im folgenden untereinander aufgelistet (der Median beträgt 43):

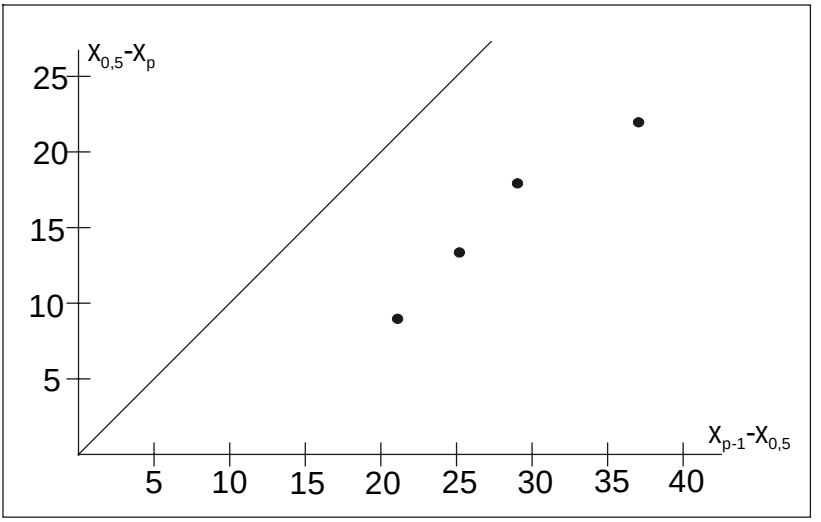
Wert	Quantil	$\tilde{x}_{0,5} - \tilde{x}_p$	Wert	Quantil	$\tilde{x}_{1-p} - \tilde{x}_{0,5}$
21	$=\tilde{x}_{1/9}$	43-21=22	80	$=\tilde{x}_{8/9}$	80-43=37
25	$=\tilde{x}_{0,2}{}^{+}$	43-25=18	72	$=\tilde{x}_{0,8}{}^{+}$	72-43=29
29,5*	$=\tilde{x}_{0,25}=Q_1$	43-29,5=13,5	68*	$=\tilde{x}_{0,75}=Q_3$	68-43=25
34	$=\tilde{x}_{1/3}$	43-34=9	64	$=\tilde{x}_{2/3}$	64-43=21

$^{+}$ erstes, bzw. fünftes Quintil

* interpoliert

Die Punkte mit den Koordinaten (37,22), (29,18), (25,13½), und (21,9) bilden das Schiefediagramm (Abb. 5.5).

Abb. 5.5: Schiefediagramm für das Beispiel 5.25



c) Symmetrisierende Transformationen

Zur Beseitigung einer Asymmetrie in den Daten werden bestimmte Transformationen empfohlen, deren bekannteste die Potenztransformation ist.

Def. 5.13: Potenztransformation

Die Variable X wird in die Variable Y nach Maßgabe einer Potenztransformation transformiert, wenn gilt

$$(5.66) \quad y_v = \begin{cases} (x_v + c)^p & \text{für } p \neq 0 \\ \ln(x_v + c)^p & \text{für } p = 0 \end{cases}$$

Bemerkungen zu Def. 5.13

1. Man spricht auch von einer Leiter der Transformationen (ladder of powers), weil der (nur durch trial and error zu findende) Parameter p beliebige Werte annehmen kann. Wichtige Spezialfälle sind (bei c=0):

   p = -1 $\quad y = 1/x$
   p = 1/2 $\quad y = \sqrt{x}$ (Wurzeltransformation)
   p = 1 $\quad y = x$ (Lineartransformation)

2. Ist p < 1, so werden die größeren Werte von X stärker reduziert (gestaucht) als die kleineren Werte. Diese Transformation eignet sich für linkssteile Verteilungen von X (die Verteilung von Y kann dann [nahezu] symmetrisch sein). Für p > 1 gilt dann das Umgekehrte.

3. Negative Werte von p führen zu einer Vertauschung der Reihenfolge, d.h. ist $x_1 < x_2$ dann ist $y_1 > y_2$.

4. Die Konstante c wird eingeführt (und so bemessen) damit die Variable Y nicht negativ wird.

5. Beispiel 5.26 zeigt für einige Werte von x und p die Wirkung der Transformation.

Beispiel/Lösung 5.26:

Wirkung der Potenztransformation

x	Transformierte Werte y bei					
	p = -½	p = 0	p = ½	p = 0,9	p = 1,5	p = 2
10	0,3162	2,3026	3,1623	7,9433	31,6228	100
20	0,2236	2,9957	4,4721	14,8227	89,4427	400
30	0,1826	3,4012	5,4772	21,3506	164,3168	900
40	0,1581	3,6889	6,3246	27,6601	252,9822	1600

Man erkennt: gleichen Abständen zwischen den x-Werten entsprechen nicht mehr gleiche Abstände zwischen den y-Werten. Letztere werden mit zunehmenden x absolut kleiner bei $|p| < 1$ und größer wenn gilt $|p| > 1$. Ist p negativ, so vertauscht sich die Reihenfolge.

7. Wölbung

Verteilungen können sich danach unterscheiden, wie sehr sich die Merkmalswerte in der Mitte oder an den Enden häufen, bzw. je steiler ihr Verlauf in der Umgebung des Medians ist. Man spricht dann von verschiedenen Arten und Stärken der Wölbung [synonym: Kurtosis, Exzess] einer (meist symmetrischen und eingipfligen) Verteilung. In Abb. 5.4 ist beispielsweise die Verteilung B flacher (schwächer) und die Verteilung D steiler (stärker) gewölbt. Die Wölbung ist ein ähnlicher Aspekt einer Häufigkeitsverteilung wie die Streuung, gleichwohl aber hiervon zu unterscheiden: denn die Verteilungen B und D in Abb.5.4 haben, gemessen an der Standardabweichung s, die gleiche Streuung, aber eine unterschiedliche Wölbung.

Maße der Wölbung W (Kurtosis) sollten als Formmaßzahlen (Gestaltparameter) - wie die Schiefemaße - invariant sein gegenüber linearen Transformationen: bei $y_v = a + b_x$ soll gelten $W_y = W_x$ ($b \neq 0$ [bei der Schiefe ist $b > 0$ zu fordern])

Wie in Abschnitt 5 bereits bemerkt, kann mit Hilfe des vierten zentralen Moments z_4 eine Maßzahl W_M der Wölbung gebildet werden, die angibt, inwieweit sich die Wölbung einer bestimmten Verteilung von der einer Normalverteilung unterscheidet.

Def. 5.14: Wölbungsmaße

a) Beim Wölbungsmaß W_M wird das vierte zentrale Moment durch die quadrierte Varianz (denn $(s^2)^2 = s^4$) geteilt:

$$(5.67)\quad W_M = \frac{z_4}{s^4} - 3 .$$

b) Weniger bekannt sind Wölbungsmaße auf der Basis von Quantilen, etwa ein Quantilskoeffizient W_Q der Wölbung:

$$(5.68)\quad W_Q = 1 - \frac{\tilde{x}_{1-p} - \tilde{x}_p}{\tilde{x}_{1-q} - \tilde{x}_q}$$

mit $0 < q < p < ½$.

Bemerkungen zu Def. 5.14:

1. Man sieht, dass in W_M das vierte zentrale Moment relativiert (und damit maßstabsunabhängig) wird durch s^4. Es läßt sich zeigen, dass der Ausdruck z_4 / s^4 für die Normalverteilung den Wert 3 annimmt. Deshalb gilt:

   - $W_M = 0$ bei der Normalverteilung, bzw. einer Häufigkeitsverteilung die genauso gewölbt ist wie die Normalverteilung (man sagt dann, sie sei ***mesokurtisch***),
   - $W_M > 0$ bei Häufigkeitsverteilungen, die vergleichsweise steiler als die Normalverteilung gewölbt sind (***leptokurtisch*** = hochgewölbt, spitz),
   - $W_M < 0$ bei Häufigkeitsverteilungen, die vergleichsweise flacher als die Normalverteilung gewölbt sind (***platykurtisch*** = flachgewölbt).

   In den Beispielen 5.27 und 5.28 wird die Vorgehensweise zur Berechnung der Wölbung W_M dargestellt.

2. Der Momentkoeffizient der Kurtosis W_M hat, wie gezeigt werden kann, den folgenden Wertebereich: $-2 < W_M < \infty$.

3. Ein Beispiel (Spezialfall) für den Quantilskoeffizient W_Q der Kurtosis wäre

   $$(5.68a)\quad W^*_{,Q} = 1 - \frac{Q_3 - Q_1}{Q_4^* - Q_1^*}$$

   mit $q = 0{,}2$ und $p = 0{,}25$ unter Verwendung des ersten und dritten Quartils (Q_1 und Q_3) sowie des ersten und vierten Quintils Q_1^* und Q_4^*. Im Beispiel 5.29 wird die Berechnung dieses Wölbungsmaßes gezeigt.

Beispiel 5.27:

Man bestimme die Momentkoeffizienten der Schiefe und Wölbung für die vier Verteilungen des Beispiels 5.17. Ändern sich Schiefe und Wölbung, wenn man zu allen Merkmalswerten die Zahl 5 addiert?

Lösung 5.27:

Für die zentralen Momente erhält man

Verteilung	z_3	z_4
A	+150	3600
B	0	2700
C	-150	3600
D	0	5100

Da die Standardabweichung bei allen Verteilungen s = 6 ist, ergibt sich für die:

	Momentschiefe z_3/s^3	Wölbung $(z_4/s^4) - 3$
Verteilung A:	+0,694	-(8/36) = - 0,2222
Verteilung B:	0	-33/36 = - 0,9167
Verteilung C:	-0,694	-0,2222
Verteilung D:	0	+0,9352

Schiefe und Wölbung ändern sich nicht, wenn man zu allen Merkmalswerten die Zahl 5 addiert, weil sich dann auch das arithmetische Mittel um 5 erhöht und nur **zentrale** Momente verwendet werden.

Beispiel 5.28:

Durch Variation des Beispiels 5.18 sind die folgenden Verteilungen A bis D erzeugt worden, die anschaulich im Zentrum zunehmend steiler verlaufen (zunehmende Wölbung):

B

x_j	n_j
10	1
15	2
20	3
25	2
30	1

C

x_j	n_j
10	1
15	1
20	5
25	1
30	1

D

x_j	n_j
10	0
15	1
20	7
25	1
30	0

Die Verteilung A besteht aus den Einzelwerten 0,5,10,15,20,25,30, 35 und 40. Alle Verteilungen haben das gleiche arithmetische Mittel von $\bar{x} = 20$.

Lösung 5.28:

	A	B	C	D
z_2	1500/9	300/9	250/9	50/9
z_4	442500/9	22500/9	21250/9	1250/9
$z_4/s^4 - 3$	-1,23	-0,75	+0,06	+1,5

Beispiel 5.29:

Man berechne und interpretiere die Kurtosis W_Q^* gem. Gl. 5.68a für das Bsp. 5.24!

Lösung 5.29:

Die Quintile sind in diesem Fall $Q_1^* = 1400$ und $Q_4^* = 3800$ und für die Quartile erhält man $Q_1 = 1500$ und $Q_3 = 3400$, so dass Gl.5.68a ergibt $W_Q^* = 1 - (3400-1500) / (3800-1400) = 1- 1900/2400 = 0{,}21$. Es ist stets $(Q_3 - Q_1) \leq (Q_4^* - Q_1^*)$, so dass der von 1 zu subtrahieren-de Quotient nicht größer sein kann als 1. Verabredet man die folgenden Strecken unter der Häufigkeitsverteilung:

$A = Q_3 - Q_1$ $\qquad$ $C = Q_1 - Q_1^*$ und

$B = Q_4^* - Q_3$ $\qquad$ $D = Q_4^* - Q_1^*$

so ist offenbar $D = A + B + C$ und $W_Q^* = (B+C)/D$. Bei zunehmender Wölbung wird die zentrale Strecke A kleiner und die Ausläufer B und C werden (relativ zu A) größer. Dann muss W_Q^* zunehmen.