3 Lage- und Streuungsmaße

- Häufigkeitsverteilungen geben „quantitative Information“ von Merkmalen vollständig wieder. Oft will man charakteristische Aspekte von Verteilungen durch eine einzelne Zahl charakterisieren. (Informationsverdichtung/-reduktion, einfacher Vergleich von Verteilungen)

- Grafische Darstellungen geben einen allgemeinen Eindruck der Verteilung eines Merkmals:
  - Lage und Zentrum der Daten,
  - Streuung der Daten um dieses Zentrum,
  - Schiefe / Symmetrie und Unimodalität / Multimodalität der Daten.
- Jetzt Quantifizierung / Charakterisierung
- Im Folgenden zunächst: Maßzahlen zur Beschreibung von Lage und Streuung durch $eine$ Zahl.
- $Lage$maße sollen die $zentrale\ Tendenz$ (das Zentrum) eines Merkmals beschreiben. Sie beantworten also Fragen über die Häufigkeitsverteilung wie:

  - Wo liegen die meisten Beobachtungen?
  - Wo liegt der „Schwerpunkt“ einer Verteilung?
  - Wo liegt die „Mitte“ der Beobachtungen?
  - Was ist eine „typische“ Beobachtung?
- Streuungsmaße beschreiben die $Variabilität$ eines Merkmals.

3.1 Arithmetisches Mittel und Varianz

3.1.1 Arithmetisches Mittel: Grundlegendes

Beachte: Es gibt nicht das Lagemaß schlechthin. Die unterschiedlichen Lagemaße sind je nach Situation unterschiedlich geeignet. Die Eignung ist insbesondere abhängig von der Datensituation und dem Skalenniveau.

Definition 3.1.

Sei $x_1, \ldots, x_n$ die Urliste eines (mindestens) intervallskalierten Merkmals X. Dann heißt

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$$

das *arithmetische Mittel* der Beobachtungen $x_1, \ldots, x_n$.

Bemerkungen:

- Das arithmetische Mittel ist also das Lagemaß, das typischerweise als Mittelwert oder Durchschnitt bezeichnet wird.
- Das arithmetische Mittel muss nicht mit einer der beobachteten Ausprägungen zusammenfallen.

Beispiel: Anzahl von Statistikbüchern, die ein Student besitzt (fiktiv).

Person	Anzahl
1	0
2	2
3	1
4	2
5	2
6	3
7	0
8	12
9	1
10	2

$$\bar{x} \quad =$$

Alternative Berechnung basierend auf Häufigkeiten:

Hat das Merkmal X die Ausprägungen $a_1, \ldots, a_k$ und die (relative) Häufigkeitsverteilung $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$, so gilt

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} a_j h_j = \sum_{j=1}^{k} a_j f_j.$$

Im Beispiel: Häufigkeitstabelle:

0	1	2	3	4	5	6	7	8	9	10	11	12

bzw.

Alte Berechnung:

$$\bar{x} =$$

Neue Berechnung:

$$\bar{x} \quad =$$

Weitergehend kann man Daten charakterisieren, indem man spezifische Subgruppen bildet und die artithmetischen Mittel in diesen Subgruppen tabellarisch gegenüberstellt.

Beispiel: Einfacher Tabellenmietspiegel

durchschnittliche Nettomiete in Euro/qm (Fallzahlen)				
	Wohnfläche			
Baujahr	bis 50 qm	51 bis 80 qm	81 qm und mehr	
bis 1918	9.00 (45)	7.88 (164)	7.52 (200)	7.83 (409)
1919 bis 48	6.90 (42)	6.87 (94)	6.50 (52)	6.78 (188)
1949 bis 65	9.04 (129)	7.84 (237)	7.95 (70)	8.21 (436)
1966 bis 80	10.05 (173)	7.97 (313)	7.80 (156)	8.49 (642)
1981 bis 95	10.59 (45)	9.53 (162)	9.72 (63)	9.75 (270)
1996 bis 2001	10.60 (15)	10.28 (58)	9.69 (35)	10.14 (108)
	9.43 (449)	8.20 (1028)	7.93 (576)	8.39 (2053)

Beispiel: Augenfarbe

	h_j
0: grün	2
1: grau	2
2: rot	0
3: blau	6

$$\bar{x} =$$

Bemerkungen:

- Das arithmetische Mittel setzt zwingend ein intervallskaliertes Merkmal voraus. Auf einem niedrigerem Skalenniveau ist die Addition nicht erlaubt, und daher sind die entsprechenden Mittelwertsbildungen sinnlos und nicht interpretierbar (auch wenn sie ein Software-Paket ohne zu zögern ausspuckt).
- Einzige Ausnahme: Binäre Merkmale (mit nur zwei Ausprägungen), deren Ausprägungen als 0/1 (nur so!) kodiert werden. In diesem Fall kann das arithmetische Mittel als Anteil von Beobachtungen mit Ausprägung 1 interpretiert werden.

Weitere Eigenschaften des arithmetischen Mittels:

- $\bar{x}$ ist derjenige Wert, den jede Beobachtungseinheit erhielte, würde man die Gesamtsumme der Merkmalsausprägungen gleichmäßig auf alle Einheiten verteilen.

- $\bar{x}$ ist der Schwerpunkt der $x_1, \ldots, x_n$, d.h. es gilt:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

  Vorstellung: Für jede Beobachtung i im Punkt x_i Gewicht mit 1 kg hinlegen.

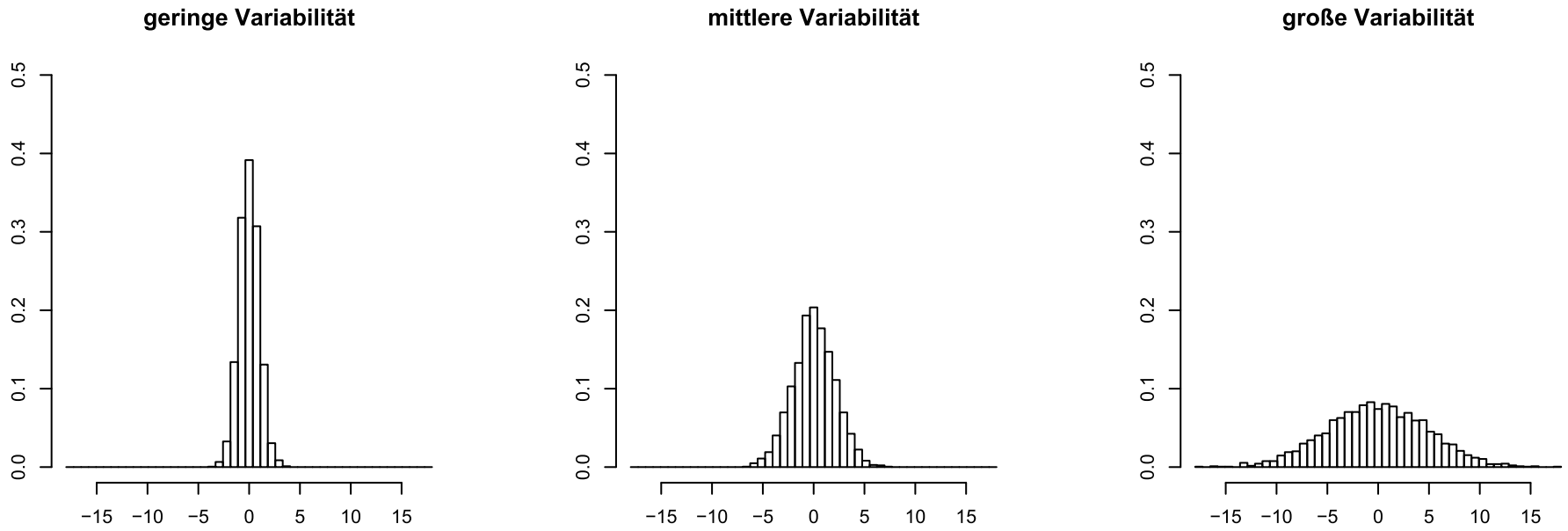
- Die Schwerpunktseigenschaft macht auch deutlich: außerordentliche Hebelwirkung extrem großer und kleiner Werte: (lässt man die Beobachtung 12 im Beispiel weg, dann gilt: $\bar{x} = \frac{13}{9} = 1.44$). Man muss den Effekt kennen, ob er gewünscht ist, oder nicht, ist eine inhaltliche Frage.

- Insbesondere ist damit das arithmetische Mittel sehr $ausreißeranfällig$, d.h. ein falsch gemessener Wert kann „den ganzen Mittelwert zerstören“.
- Wenn es tatsächlich um das in den Ort fließende Einkommen geht, dann ist das der korrekte Wert; wenn man sich für die Lebensverhältnisse einer typischen Person interessiert, dann wird man eher andere Maße verwenden.
- Befürchtet man Ausreißer, so weicht man gelegentlich auf das sogenannte α-$getrimmte\ Mittel$ aus, bei dem man die $\alpha\%$ größten und kleinsten Werte (z.B. α=5) weglässt. Alternativ verwendet man oft den Median (s.u.).

3.1.2 Varianz und Standardabweichung: Grundlegendes

Eine Verteilung ist durch die Angabe von einem oder mehreren Lagemaßen nur unzureichend beschrieben.

Beispiel: Häufigkeitsverteilungen mit gleicher zentraler Tendenz:



Streuungsmaße beantworten Fragen wie

- Wie groß ist die durchschnittliche Abweichung vom Mittelwert?
- Über welchen Bereich erstrecken sich die Beobachtungen?
- Wie stark schwanken die Beobachtungen?

Bemerkung: Von Streuung im eigentlichen Sinne kann man nur bei mindestens intervallskalierten Daten sprechen, da nur dort Abstände interpretierbar sind. (Es gibt verschiedene Versuche, ein analoges Konzept für ordinal skalierte Daten zu definieren, aber bisher hat sich keine dieser Definitionen durchgesetzt.

Varianz: Sei $x_1, \ldots, x_n$ die Urliste eines intervallskalierten Merkmals X. Dann heißen

$$\tilde{s}_X^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

die *(empirische) Varianz* oder *Stichprobenvarianz* und

$$\tilde{s}_X := \sqrt{\tilde{s}_X^2}$$

die *(empirische) Streuung, Stichprobenstreuung* oder *(empirische) Standardabweichung von* X.

Bemerkungen:

- Die Varianz misst die durchschnittliche quadratische Abweichung vom Mittelwert.
- Vorsicht: Der Begriff Streuung wird in einem doppelten Sinne gebraucht: Allgemein als Phänomen generell („wir suchen nach Maßzahlen zur Beschreibung der Streuung der Daten“), andererseits als eine bestimmte Maßzahl für das Problem.
- Durch das Quadrieren tragen negative und positive Abweichungen vom Mittelwert gleichermaßen zur Varianz bei.
- Die Varianz besitzt im Vergleich zum Merkmal X die quadrierte Einheit. Sie ist daher unanschaulicher zu interpretieren, besitzt aber andererseits viele mathematische Vorzüge. Die Standardabweichung dagegen wird in der gleichen Einheit gemessen wie X.

- Sind die Ausprägungen $a_1, \ldots, a_k$ mit (relativer) Häufigkeitsverteilung $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$ gegeben, so gilt

$$\begin{aligned} \tilde{s}_X^2 &= \frac{1}{n} \sum_{j=1}^{k} h_j (a_j - \bar{x})^2 = \\ &= \sum_{j=1}^{k} f_j (a_j - \bar{x})^2. \end{aligned}$$

- Ist aus dem Kontext klar ersichtlich welches Merkmal betrachtet wird, so lässt man das X in der Notation auch häufig weg, schreibt also einfach $\tilde{s}^2$ und $\tilde{s}$.

Beispiel: Statistikbücher

Ausprägungen	h_j
0	2
1	2
2	4
3	1
12	1
$\sum$	10

Berechnung der Varianz über die ursprüngliche Formel:

$$\tilde{s}^2 \quad =$$

Berechnung über die Häufigkeitsverteilung:

$$\tilde{s}^2 \quad =$$

Standardabweichung:

$$\tilde{s} =$$

Verschiebungssatz: Es gilt

$$\tilde{s}_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \overline{x^2} - (\bar{x})^2.$$

Achtung (sehr häufige Fehlerquelle):

Der Verschiebungssatz ist sehr bequem zum Berechnen der Varianz, es können aber beim Verwenden von Taschenrechnern bei sehr großen Ausprägungen starke Rundungsfehler auftreten, die das Ergebnis eventuell verfälschen. Für Aufgaben von Klausurlänge ist es aber meist geschickt, den Verschiebungssatz verwenden!

Beispiel: Statistikbücher.

Berechne die empirische Varianz mit Hilfe des Verschiebungssatzes.

Person i	Anzahl Bücher: X x_i	
1	0	
2	2	
3	1	
4	2	
5	2	
6	3	
7	0	
8	12	
9	1	
10	2	

$$\tilde{s}_X^2 = \qquad \tilde{s}_X =$$

Korrigierte empirische Varianz:

Neben der empirischen Varianz existiert noch eine alternative Definition der Varianz, die sog. *korrigierte (empirische) Varianz*.

Sei $x_1, \ldots, x_n$ die Urliste eines intervallskalierten Merkmals X. Dann heißt

$$s_X^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

die *korrigierte empirische Varianz* oder *korrigierte Stichprobenvarianz von* X.

- Der Sinn des Vorfaktors $\frac{1}{n-1}$, also der Begriff „korrigierte (empirische) Varianz“ wird erst in Statistik II deutlich: s_X^2 hat inferenz-theoretisch schönere Eigenschaften als $\tilde{s}_X^2$.
- Für großen Stichprobenumfang n nähern sich s_X^2 und $\tilde{s}_X^2$ an, weil dann $n - 1 \approx n$.

3.1.3 Arithmetisches Mittel und Varianz unter (linearen) Transformationen

Die Intervallskala erlaubt lineare Transformationen der Form $a + bX$, die Ratioskala Transformationen der Form $b \cdot X$, wobei a und b feste Konstanten sind, so dass man aus der Urliste $x_1, x_2, \ldots, x_n$ eine neue Urliste $y_1, y_2, \ldots, y_n$ erhält, mit $y_i = ax_i + b$, $i = 1, \ldots, n$. Wie verändert sich das arithmetische Mittel bei diesen oder allgemeineren Transformationen?

Beispiele:

- Lineare Transformation $Y = a \cdot X + b$

- Nichtlineare Transformation

Satz 3.2. [Arithmetisches Mittel und lineare Transformationen.]

Gegeben sei die Urliste $x_1, \ldots, x_n$ eines (mindestens) intervallskalierten Merkmals X mit arithmetischem Mittel $\bar{x}$. Betrachtet wird für reelle Konstanten a, b das (linear transformierte) Merkmal $Y = a \cdot X + b$ und die zugehörigen Ausprägungen $y_1, \ldots, y_n$. Dann gilt für das arithmetische Mittel $\bar{y}$ von Y:

$$\bar{y} = a \cdot \bar{x} + b.$$

Beweis:

Bemerkungen:

- Vorsicht: Ist X verhältnisskaliert, so geht für $b \neq 0$ der natürliche Nullpunkt für Y verloren.
- Der Satz gilt im Allgemeinen nur, falls die Transformation von X auf Y linear ist. Z.B. ist bei $Y = X^2$ im Allgemeinen $\bar{y} \neq (\bar{x})^2$ (wie im Beispiel gezeigt).

Varianz unter Transformationen: Wie ändert sich die Varianz bei (linearer) Transformation eines Merkmals?

Satz 3.3.

Sei $x_1, \ldots, x_n$ die Urliste eines mindestens intervallskalierten Merkmals X und $y_1, \ldots, y_n$ die zugehörige Urliste des Merkmals $Y = a \cdot X + b$. Dann gilt

Bemerkungen:

- Eine spezielle Transformation, die sogenannte $Standardisierung$, ist der Übergang zum Merkmal Z mit

$$z_i := \frac{x_i - \bar{x}}{\tilde{s}_X}.$$

Z besitzt arithmetisches Mittel 0 und (empirische) Varianz 1. Man erzeugt damit in gewisser Weise eine natürlich Skala.

Begründung:

3.1.4 Das arithemtische Mittel bei gruppierten Daten

Häufig hat man die Daten nur in gruppierter Form vorliegen.
Wie lässt sich in diesem Fall ein sinnvoller Mittelwert definieren?

Typisches Beispiel: Einkommensverteilung

	Anzahl h'_l	
$0 \leq x < 750$	3	
$750 \leq x < 1250$	8	
$1250 \leq x < 1750$	6	
$1750 \leq x < 2250$	2	
$2250 \leq x < 3250$	1	
$\sum$	20	

Definition 3.4.

Sei X ein intervallskaliertes Merkmal, das in gruppierter Form mit k Klassen $[c_0, c_1), [c_1, c_2), \ldots, [c_{k-1}, c_k]$ erhoben wurde. Mit $h'_l,\ l = 1, \ldots k$, als absoluter Häufigkeit der $l-$ten Klasse, f'_l als zugehöriger relativer Häufigkeit und $m_l := \frac{c_l + c_{l-1}}{2}$ als der jeweiligen Klassenmitte definiert man als *arithmetisches Mittel für gruppierte Daten*

$$\bar{x}_{grupp} := \frac{1}{n} \sum_{l=1}^{k} h'_l m_l = \sum_{l=1}^{k} f'_l m_l.$$

Im Beispiel:

Bemerkungen:

- Bei nach oben offener letzter Kategorie (Einkommen größer als 2250), wäre die Klassenmitte nicht definiert.
- Im Allgemeinen gilt $\bar{x} \neq \bar{x}_{grupp}$; nur in Extremfällen, z.B. wenn das Merkmal in jeder Gruppe gleichmäßig verteilt ist, erhält man die Gleichheit.
- $\bar{x}_{grupp}$ hängt von der Gruppenmitte und damit von der gewählten Gruppierung ab: Fasst man z.B. die ersten drei Gruppen und die letzten beiden jeweils zusammen, so erhält man

	h'_l	m_l
$0 \leq x < 1750$	17	
$1750 \leq x < 3250$	3	

und

$$\bar{x}_{grupp} = \frac{1}{n} \sum_{l=1}^{k} h'_l m_l$$

- Im Allgemeinen ist $\bar{x}_{gruppt}$ natürlich nur eine grobe Approximation an den „echten“, d.h. auf ungruppierten Daten beruhenden, Mittelwert. Deshalb extreme Vorsicht bei Vergleichen zweier Gesamtheiten.

* Eigentlich kann man nur mit Sicherheit folgende Abschätzung geben: Jeder in der l-ten Gruppe verdient mindestens c_{l-1} und höchstens c_l. Damit ergibt sich als Abschätzung für das arithmetische Mittel

$$\bar{x}_{unten} := \frac{1}{n}\sum_{l=1}^{k} h_l c_{l-1} \leq \bar{x} \leq \frac{1}{n}\sum_{l=1}^{k} h_l c_l =: \bar{x}_{oben}$$

Diese Abschätzung ist oft relativ grob. Andererseits ist sie aber oft das Beste, was man ohne unüberprüfbare Zusatzannahmen aus den Daten herausholen kann.
Ein gesicherter Vergleich zweier Gesamteheiten ist dann und nur dann möglich, wenn $\bar{x}_{unten}$ einer Gesamtheit kleiner ist, als $\bar{x}_{oben}$ einer anderen Gesamtheit.

- Sind die ungruppierten Daten erhältlich, so ist $\bar{x}$ vorzuziehen, da jede Gruppierung Informationsverlust mit sich bringt.

- Andererseits sind gruppierte Daten leichter (und oft wahrheitsgetreuer) erhebbar.

3.1.5 Arithmetisches Mittel und Varianz unter geschichteten Daten

Insbesondere bei Tertiäranalysen hat man häufig nicht die Urliste zur Verfügung, sondern nur Mittelwerte $\bar{x}_l$ in einzelnen Schichten $l = 1, \ldots, z$, in die die Grundgesamtheit zerlegt ist. (Man denke ferner an geschichtete Stichproben, wie sie in Kapitel 1 erwähnt werden.)

Beachte: hier wird nicht das Merkmal, sondern die Grundgesamtheit in Gruppen eingeteilt.

Satz 3.5.

Betrachtet werde ein (mindestens) intervallskaliertes Merkmal X mit Urliste $x_1, \ldots, x_n$, die schichtweise zusammengefasst sei. Sei $x_1^{(l)}, \ldots, x_{n^{(l)}}^{(l)}$, Urliste in Schicht l, $l = 1, \ldots, z$, und $\bar{x}^{(l)} = \frac{1}{n^{(l)}} \sum_{i=1}^{n^{(l)}} x_i^{(l)}$ der Mittelwert in der l-ten Schicht, $l = 1, \ldots, z$, so gilt für das arithmetische Mittel $\bar{x}$ von X:

$$\bar{x} = \frac{1}{n} \sum_{l=1}^{z} n^{(l)} \bar{x}^{(l)}$$

Beweis:

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

$$= \frac{1}{n} \sum_{l=1}^{z} n^{(l)} \bar{x}_l$$

Varianzzerlegung / Streuungszerlegung: Varianz bei geschichteten Daten.

Satz 3.6.

Schicht	$1, \ldots, l, \ldots, z$	
Besetzungszahlen	$n^{(1)}, \ldots, n^{(l)}, \ldots, n^{(z)};$	$\sum_{l=1}^{z} n^{(l)} = n$
Mittelwerte	$\bar{x}^{(1)}, \ldots, \bar{x}^{(l)}, \ldots, \bar{x}^{(z)}$	
Varianzen	$\tilde{s}^{2(1)}, \ldots, \tilde{s}^{2(l)}, \ldots, \tilde{s}^{2(z)}$	

Für das arithmetische Mittel gilt

$$\bar{x} = \frac{1}{n} \sum_{l=1}^{z} n^{(l)} \bar{x}^{(l)}.$$

Seien nun

$$\tilde{s}^2_{innerhalb} := \frac{1}{n} \sum_{l=1}^{z} n^{(l)} \tilde{s}^{2^{(l)}}$$

sowie

$$\tilde{s}^2_{zwischen} := \frac{1}{n} \sum_{l=1}^{z} n^{(l)} (\bar{x}^{(l)} - \bar{x})^2$$

Varianzzerlegung Es gilt

$$\begin{aligned} \text{Gesamtvarianz} &= \\ \tilde{s}^2 &= \end{aligned}$$

Bemerkungen:

- Im Detail gilt also mit den Urlisten $\{x_1^{(l)}, x_2^{(l)}, \ldots, x_{n^{(l)}}\}$ in Schicht $l, l = 1, \ldots, z$,

$$\frac{1}{n}\sum_{l=1}^{z}(\sum_{i=1}^{n^{(l)}}(x_i^{(l)} - \bar{x})^2) = \frac{1}{n}\sum_{l=1}^{z}\sum_{i=1}^{n^{(l)}}(x_i^{(l)} - \bar{x}^{(l)})^2 + \frac{1}{n}\sum_{l=1}^{z} n^{(l)}(\bar{x}^{(l)} - \bar{x})^2.$$

- Diese Zerlegungsmöglichkeit gilt *nur für Varianzen*, nicht aber für andere Streuungsmaße. Letztendlich ist sie der Grund für die Beliebtheit der Varianz – trotz anderer Unannehmlichkeiten. Deshalb sollte man eher von der Varianzzerlegung als von der Streuungszerlegung sprechen.

- Bei vielen Verfahren werden Streuungszerlegungen betrachtet; dies ist ein ganz grundlegendes Prinzip in der Statistik.
- Interpretation anhand des Beispiels mit den Einkommen der einzelnen Bundesländer: Man kann die Wichtigkeit(Erklärungskraft) der schichtbildenden Variable bewerten: ja größer $\tilde{s}^2_{zwischen}$ im Vergleich zu $\tilde{s}^2$ bzw.$\tilde{s}^2_{innerhalb}$ ist, desto „mehr Variation“ wird durch die Schichtungsvariable „erklärt“.