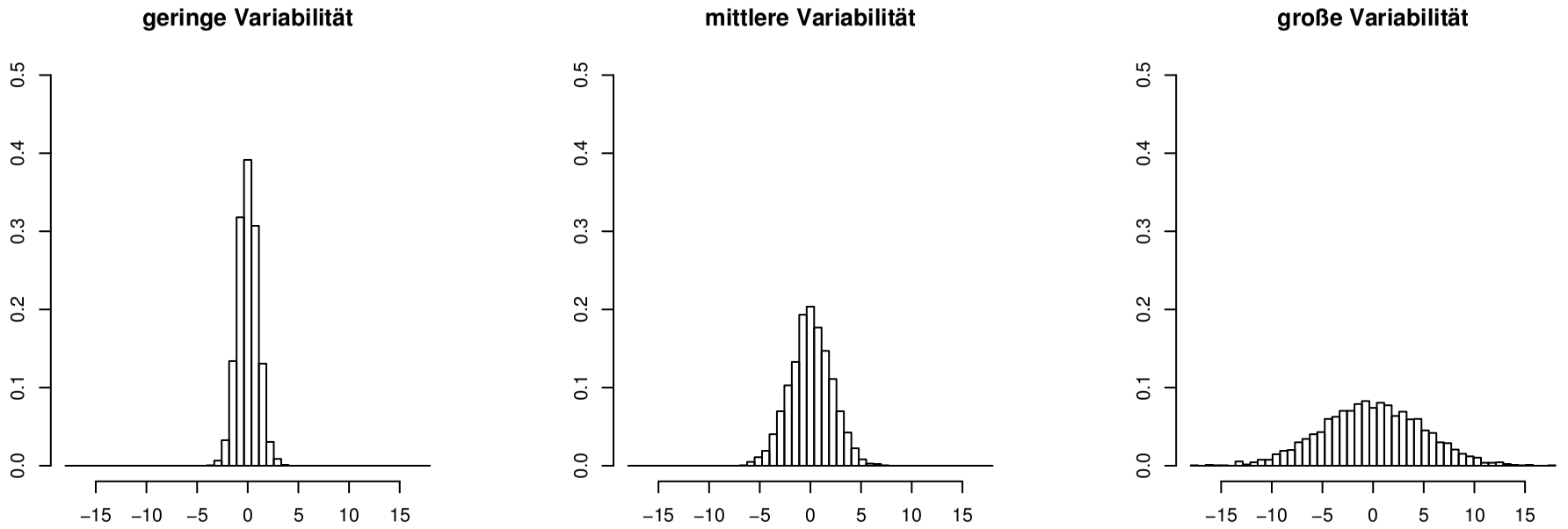
3.2 Streuungsmaße

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.

3.2.1 Varianz und Standardabweichung

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:

$$\tilde{s} =$$

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

Satz 3.7.

Sei $x_1, \ldots, x_n$ die Urliste eines mindestens intervallskalierten Merkmals X mit $\tilde{s}_X > 0$ 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:

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	
	144	

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

Varianzzerlegung / Streuungszerlegung: Varianz bei geschichteten Daten.

Zur Erinnerung: Daten liegen oft in Schichten vor (v.a. bei Sekundär- und Tertiärerhebungen). Beispiel: Daten über Einkommensverteilung geschichtet nach Bundesland. Bei der Berechnung von $\bar{x}$ waren die einzelnen Besetzungszahlen sehr wichtig.

Schicht	$1, \ldots, l, \ldots, z$
Besetzungszahlen	$n_1, \ldots, n_l, \ldots, n_z; \qquad \sum_{l=1}^{z} n_l = n$
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

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

Seien nun

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

sowie

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

- $\tilde{s}^2_{innerhalb}$
- $\tilde{s}^2_{zwischen}$
- $\tilde{s}^2_{zwischen} = 0$

- $\tilde{s}^2_{innerhalb} = 0$

Wie setzt sich die Gesamtvarianz aus den beiden Bestandteilen zusammen?

Varianzzerlegung Es gilt

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

Bemerkungen:

- Im Detail gilt also mit den Urlisten $\{x_1^{(l)}, x_2^{(l)}, \ldots, x_{n_l}^{(l)}\}$ 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“.

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

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

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

Bemerkungen:

- Der Sinn des Vorfaktors $\frac{1}{n-1}$, also der Begriff „korrigierte (empirische) Varianz“ wird erst in Statistik II deutlich: s_X^2 hat 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 $n - 1 \approx n$.
- Auch für die korrigierte Varianz gilt die Aussage zu linearen Transformationen, d.h. ist $x_1, \ldots, x_n$ die Urliste eines mindestens intervallskalierten Merkmals X mit $s_X > 0$ und $y_1, \ldots, y_n$ die zugehörige Urliste des Merkmals $Y = a \cdot X + b$. Dann gilt

$$s_Y^2 = a^2 \cdot s_X^2.$$

3.2.2 Weitere Streuungsmaße

Variationskoeffizient:

Definition 3.8.

Ist $\bar{x} > 0$, so heißt die Größe

$$v_X := \frac{\tilde{s}_X}{\bar{x}}$$

Variationskoeffizient des Merkmals X.

Bemerkungen:

- Gemessen wird hier die Streuung relativ zum Mittelwert. Insbesondere ist v_X dimensionslos.
- Der Variationskoeffizient erlaubt beispielsweise auch den Vergleich der Streuung von Preisen, die in verschiedenen Währungen gemessen wurden.

Inter-Quartils-Abstand: Sind $x_{0.25}$ und $x_{0.75}$ das obere und das untere Quartil eines Merkmals, so heißt

$$d_{QX} := x_{0.75} - x_{0.25}$$

der *Interquartilsabstand*.

Median-Absolute-Deviation: Der Median der Werte $|x_i - x_{med}|$, $i = 1, \ldots, n$ heißt *Median-Absolute-Deviation* von X (MAD_X).

Spannweite: Die Größe

$$R_X := x_{(n)} - x_{(1)}$$

heißt *Spannweite* von X.

Bemerkungen

- Alle betrachteten Streuungsmaße sind nur für (mindestens) intervallskalierte Merkmalse sinnvoll definiert, da sie auf Abständen (typischerweise dem Abstand der Beobachtungen zu einem Lagemaß) beruhen.
- $\tilde{s}^2$, $\tilde{s}$, s^2, s sind die gebräuchlichsten Streuungsmaße.
- $\tilde{s}^2$, $\tilde{s}$, s^2, s sind sehr empfindlich gegenüber Ausreißern! Das Gleiche gilt für die Spannweite R. Die Kennzahlen MAD und d_Q hingegen entstammen der sogenannten robusten Statistik, die sich um ausreißerresistente Methoden bemüht.
- Gilt $x_1 = x_2 = \ldots = x_n$, so weisen alle Streungsmaße den Wert 0 auf. Mit Ausnahme von d_Q gilt auch die Umkehrung: Sind die Steuungsmaße (außer eben d_Q) $= 0$, so sind alle Werte der Urliste gleich.

- Nochmals der Hinweis: Eine häufige Ursache für Verwirrung und Missverständnisse liegt in der Tatsache, dass der Begriff „Streuung“ in der Statistik in einem doppelten Sinn gebraucht wird:

  – in einem allgemeinen Sinn: Streuung als Phänomen („Die Daten streuen stark“).

  – in einem speziellen Sinn: als $eine$ Maßzahl für dieses Phänomen.

Beispiel: Statistikbücher

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

3.3 Box-Plot

Ziele:

- einfache Darstellung von Verteilungen und ihrer Kennzahlen
- Identifikation von potentiellen Ausreißern
  $\Rightarrow$ nicht ausreißeranfällige Meßzahlen verwenden.

Idee:

i) markiere den Median

ii) symbolisiere Lage der „mittleren Werte“ durch eine Box

iii) wie weit reichen „weitere nicht atypische“ Werte

iv) identifiziere potentielle Ausreißer: atypische (ungewöhnlich große, ungewöhnlich kleine) Werte, die genauerer Untersuchung bedürfen

zu ii) wähle mittlere 50%: Box hat Länge $dQ = x_{0.75} - x_{0.25}$

zu iii) als „nicht atypisch“ gelten alle Werte, die nicht weiter als $1.5dQ$ von der Box entfernt sind

Also bestimme:

- $x_{0.25}, \quad x_{0.50}, \quad x_{0.75}$.
- Interquartilsabstand: $d_{QX} = x_{0.75} - x_{0.25}$
- Zäune z_u, z_o, die am kleinsten bzw. größten Datenpunkt im Bereich $x_{0.25} - 1.5 \cdot d_{QX}$; $x_{0.75} + 1.5 \cdot d_{QX}$ liegen.
- Ausserhalb der Zäune werden $alle$ Punkte eingezeichnet; sie sind ausreißerverdächtig.

Vorsicht bei der Anwendung von Software! Vor allem außerhalb der Box sind auch andere Darstellungen üblich (z.B. Zäune immer bis $x_{(1)}$ und $x_{(n)}$). Toutenburg (2002) beispielsweise unterscheidet zwischen Ausreißern ($1.5 \cdot d_{QX}$ bis $3 \cdot d_{QX}$ von Rändern der Box entfernt) und Extremwerten (mehr als $3 \cdot d_{QX}$ vom Rand entfernt). Oft wird der Median durch einen dicken Punkt ausgedrückt.

Der Box-Plot gibt einen kompakten Überblick über die Form der Verteilung (Zentrale Tendenz, Variabilität, Schiefe, extreme Werte).

Box-Plots können auch zum graphischen Vergleich von Verteilungen verwendet werden: