Willkommen zur Vorlesung Statistik

Thema dieser Vorlesung:
Maßzahlen für zentrale Tendenz, Streuung und andere Eigenschaften von Verteilungen

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Inhaltsübersicht

- Maße der zentralen Tendenz
- Quantile
- Streuungsmaße
- Schiefe und Wölbung

Maße der zentralen Tendenz

Maße der zentralen Tendenz: Wozu?

Maße der zentralen Tendenz sollen durch einen numerischen Wert charakterisieren, wo das „Zentrum“ oder der „Schwerpunkt“ einer Reihe von Datenwerten liegt. Ihre Aussagekraft ist daher beschränkt, wenn die Verteilung keinen eindeutigen Schwerpunkt hat.

In Lehrbüchern stehen diese Maße häufig (auch) unter dem Begriff „Lagemaße“. Diese Redeweise ist irreführend, da sich Lagemaße auch auf andere Punkte einer Verteilung als deren Zentrum beziehen können.

Maße der zentralen Tendenz: Überblick

Die wichtigsten Maße der zentralen Tendenz sind

- der Modus (oder Modalwert): Der häufigste Wert
- der Median (oder Zentralwert): Der Wert „in der Mitte“
- das arithmetische Mittel: Der „Durchschnitt“

Es gibt weitere Maße (z. B. das geometrische Mittel für Daten, die sich auf relative Änderungen beziehen, u. a. mehr), die hier nicht dargestellt werden.

Der Modus

Der Modus ist der Wert (die Ausprägung), der (die) in einer Verteilung am häufigsten vorkommt. Er kann aus einer Häufigkeitstabelle oder einer graphischen Darstellung derselben abgelesen werden.

	h_j	%
Nichts angekreuzt	6	22,2
Differenz	6	22,2
Steigung	2	7,4
Flächeninhalt	9	33,3
Kreisfläche	1	3,7
Schnittpunkt	3	11,1
n	27	

Der Modus: Einsatz und Grenzen

Der Modus ist die „allgemeinste“ Maßzahl der zentralen Tendenz, da er jedenfalls im Grundsatz bei Variablen jeglichen Messniveaus bestimmt werden kann. Aber:

- Im allgemeinen ist der Modus nur sinnvoll bei Daten mit wenigen diskreten Ausprägungen.
- Er ist umso aussagekräftiger, je mehr sich unter diesen eine Ausprägung hervorhebt.
- Gelegentlich entsteht das Problem, dass mehrere Werte gleich häufig vorkommen; dann ist kein Modus bestimmbar.

Der Median (Zentralwert)

- Die Datenwerte/Messwerte („Urliste“) werden der Größe nach geordnet ($\rightarrow$ Daten müssen mindestens ordinalskaliert sein).
- Der Wert, der
  - (salopp gesprochen) genau in der Mitte liegt, bzw.
  - (exakt formuliert) die Bedingung erfüllt, dass mindestens die Hälfte der Messwerte kleiner oder gleich und mindestens die Hälfte größer oder gleich diesem Wert ist,

  heißt Median (formal bezeichnet oft mit $\tilde{x}$).

Achtung: Bei einer geraden Zahl von Messwerten existiert kein Datenwert „in der Mitte“. Es gilt allgemein:

- $\tilde{x} = x_{((n+1)/2)}$ bei ungeradem n
- $\tilde{x} = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$ bei geradem n

mit $x_{(i)}$ als dem i-ten Wert der der Größe nach geordneten Datenreihe.

Der Median im Beispiel

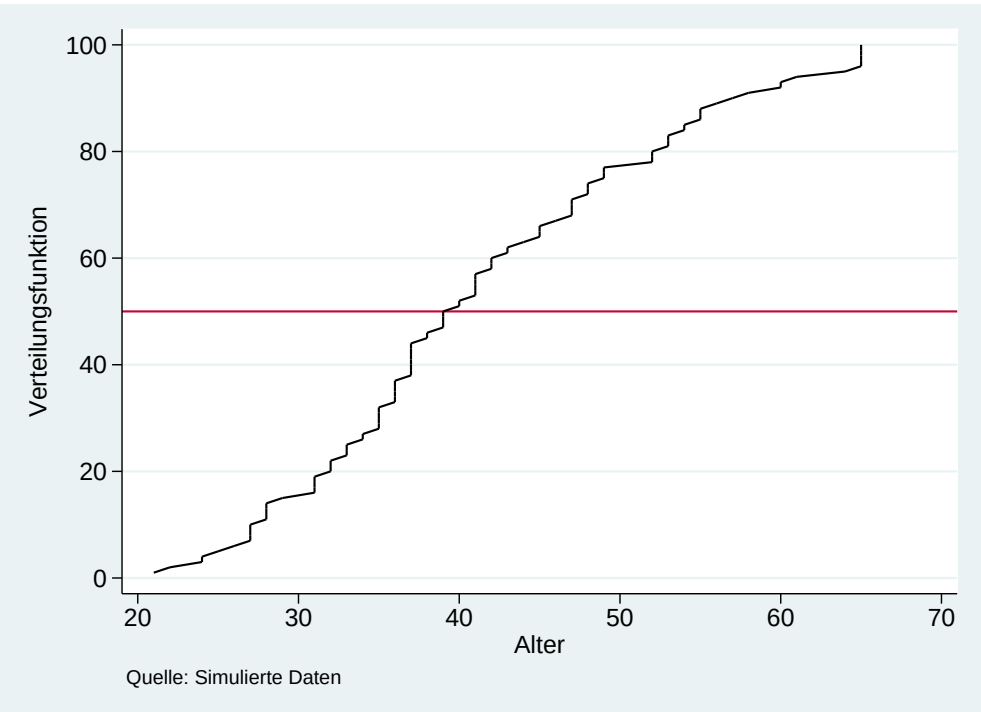
Messreihe 1				Messreihe 2			
i	x_i	(i)	$x_{(i)}$	i	x_i	(i)	$x_{(i)}$
1	2 000	1	1 500	1	2 000	1	1 500
2	5 000	2	2 000	2	5 000	2	2 000
3	4 000	3	2 500	3	4 000	3	2 500
4	1 500	4	4 000	4	1 500	4	3 500
5	2 500	5	5 000	5	2 500	5	4 000
				6	3 500	6	5 000

In Messreihe 1 steht der Wert 2500 genau in der Mitte der geordneten Reihe. Der Median ist daher 2500.

In Messreihe 2 gibt es keinen „Wert in der Mitte“. Der Median wird daher als arithmetisches Mittel aus den beiden Werten „um die Mitte herum“ gebildet, er beträgt also 3000.

Der Median graphisch

Der Median lässt sich graphisch auch (wenn auch oft nur annähernd) aus der Verteilungsfunktion ablesen (Schnittpunkt der roten Linie bei y = 0,5 mit der Verteilungsfunktion):



Der Median in Häufigkeitstabellen

Sind die Werte einer Häufigkeitstabelle der Größe nach geordnet, so entspricht der Median dem Wert, bei dem die kumulierten Anteilswerte 50 Prozent erreichen oder überschreiten.

Mathenote im Abitur

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1	3	11,1	11,1	11,1
	2	7	25,9	25,9	37,0
	3	9	33,3	33,3	70,4
	4	8	29,6	29,6	100,0
	Gesamt	27	100,0	100,0	

Im Beispiel beträgt der Median 3.

Das arithmetische Mittel

Das arithmetische Mittel, oft auch nur als Mittelwert bezeichnet, ist allgemein bekannt als „Durchschnitt“. Die Berechnung ist nur bei metrischen Daten sinnvoll. Errechnet wird es nach folgender Formel:

$$\bar{x} = \frac{1}{n}(x_1 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In Worten: Die Summe der Einzelwerte wird dividiert durch deren Anzahl.

„Schwerpunkteigenschaft“ des arithmetischen Mittels: Die Summe der quadrierten Abweichungen der Einzelwerte vom arithmetischen Mittel ist minimal (bei jedem anderen Bezugspunkt für die Abweichungen wird die Summe größer).

Das arithmetische Mittel im Beispiel

Gegeben seien wieder die fünf Messwerte aus dem Beispiel „Median“.

i	x_i
1	2 000
2	5 000
3	4 000
4	1 500
5	2 500

Das arithmetische Mittel wird berechnet als

$$\bar{x} = \frac{1}{n}(2\,000 + 5\,000 + 4\,000 + 1\,500 + 2\,500) = \frac{1}{n} \cdot 15\,000 = 3\,000$$

Median oder arithmetisches Mittel I

Wir vergleichen die zwei Messreihen X und X'

i	x_i	x'_i	geordnet
1	2 000	2 000	1 500
2	5 000	15 000	2 000
3	4 000	4 000	2 500
4	1 500	1 500	4 000
5	2 500	2 500	(1)5 000

$\bar{x}$ beträgt 3 000, $\bar{x}'$ beträgt 5 000.

Das arithmetische Mittel kann also durch einzelne extreme Werte beeinflusst werden; der Median bleibt dagegen unverändert.

Aber: Das arithmetische Mittel berücksichtigt alle Werte („suffiziente Statistik“).

Median oder arithmetisches Mittel II

Oft ist es sinnvoll, beides anzugeben – die Unterschiede verraten einiges über die Verteilung:

$\bar{x} \approx \tilde{x} \rightarrow$ symmetrische Verteilung

$\bar{x} > \tilde{x} \rightarrow$ rechtsschiefe Verteilung

$\bar{x} < \tilde{x} \rightarrow$ linksschiefe Verteilung

Achtung – bei diesen Regeln handelt es sich nur um Faustregeln; es gibt immer wieder Abweichungen.

Quantile (Perzentile)

Quantile: Einführung

In Analogie zum Median (50 Prozent der Datenwerte sind kleiner oder gleich, 50 Prozent sind größer oder gleich dem Median) lassen sich beliebige Quantile definieren, für die gilt:

p Prozent der (nach der Größe geordneten) Datenwerte sind kleiner oder gleich, $100 - p$ Prozent sind größer oder gleich dem betreffenden Quantil.

Anhand von mehreren Quantilen kann man eine Verteilung auf einfache Art und Weise charakterisieren.

Streng genommen bezieht sich der Begriff Quantile auf Anteilswerte; drückt man sich in Prozentwerten aus, müsste man von Perzentilen sprechen. Solange aus dem Kontext klar ist, was gemeint ist, verwende ich den Begriff Quantil (nicht zu verwechseln mit Quartil, siehe die nächsten Folien).

Die am häufigsten verwendeten Quantile: Quartile

Quartile trennen (geordnete) Datenwerte in vier gleich große Gruppen:

- Ein Viertel der Datenwerte ist kleiner oder gleich dem 1. Quartil, dem 25 %-Quartil ($Q_{0,25}$, manchmal auch als Q_1 bezeichnet)
- Die Hälfte der Datenwerte ist kleiner oder gleich dem 2. Quartil, dem 50 %-Quartil ($Q_{0,5}$, manchmal auch als Q_2 bezeichnet) (entspricht dem _____)
- Drei Viertel der Datenwerte sind kleiner oder gleich dem 3. Quartil, dem 75 %-Quartil ($Q_{0,75}$, manchmal auch als Q_3 bezeichnet)

Der Quartil(s)abstand (oder Interquartilabstand, IQR [von Interquartile Range]) ist die Differenz $Q_{0,75} - Q_{0,25}$.

Die Fünf-Punkte-Zusammenfassung

Ergänzt um Minimum und Maximum der Datenwerte ergeben die drei Quartile die sog. Fünf-Punkte-Zusammenfassung (five point summary) nach John Tukey.

Im Falle der Altersdaten (siehe vorherige Vorlesung) lautet die Fünf-Punkte-Zusammenfassung

21 – 33,5 – 39,5 – 49 – 65

Der Abstand von $Q_{0,25}$ zum Median ist etwas kleiner als der vom Median zu $Q_{0,75}$. Ebenso ist der Abstand vom Minimum zu $Q_{0,25}$ geringer als der von $Q_{0,75}$ zum Maximum. Dies entspricht dem visuellen Eindruck aus dem Histogramm: Die Verteilung ist leicht rechtsschief, es liegen etwas mehr Datenwerte in der linken als in der rechten Hälfte des Wertebereichs.

Berechnung von Quantilen

Wie beim Median wird es auch bei anderen Quantilen oft vorkommen, dass der gesuchte Wert „zwischen“ zwei Datenpunkten liegt. Eine gängige Regel für den Umgang mit diesem Problem lautet:

- Wir berechnen $n \cdot p$, d. h., Stichprobenumfang mal gesuchtes Quantil p.
  Beispiel für $Q_{0,25}$ bei $n = 14$: $14 \cdot 0{,}25 = 3{,}5$.
- Ist das Ergebnis keine ganze Zahl, wird der Wert trunkiert und 1 hinzu addiert.
  Im Beispiel: $3[,5] + 1 = 4 \rightarrow Q_{0,25} = x_{(4)}$, das 25-Prozent-Quantil ist also der vierte Wert in der (nach der Größe geordneten) Datenreihe.
- Ist das Ergebnis von Schritt 1 eine ganze Zahl, so liegt der Wert des Quantils zwischen $x_{(np)}$ und $x_{(np+1)}$. In diesem Fall muss interpoliert werden. (Meist wird das ar. Mittel aus beiden Werten berechnet.)

Berechnung von Quantilen im Beispiel

Ein Beispiel nach Ben Jann, Statistik, München/Wien: Oldenbourg 2002, S. 36:

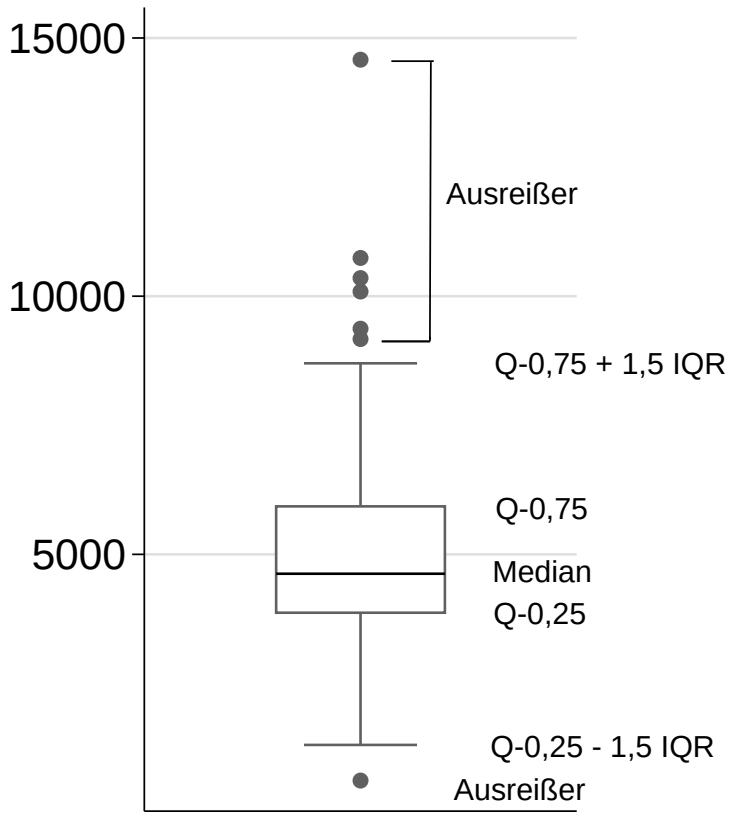
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x	0	0	3	6	6	8	9	10	12	14	18	18	22	23

$Q_{0,25}$ ist also der 4. Wert ($x_{(4)} = 6$). Das entspricht der Definition, dass mindestens 25 Prozent der Datenwerte kleiner oder gleich 6 und mindestens 75 Prozent größer oder gleich 6 sind.

Der Median liegt zwischen dem 7. und 8. Wert; nach der Regel aus dem Vorlesungsteil über „Lagemaße“ beträgt der Wert des Medians also 9,5.

Der Boxplot oder Box-and-Whisker-Plot

Ein Box-and-Whisker-Plot (meist nur: Boxplot) nach John Tukey ist eine weitere Möglichkeit, eine Verteilung zu visualisieren.



Die wichtigsten Konstruktionsregeln für Boxplots

- Es wird eine „Box“ konstruiert, deren untere Begrenzung durch $Q_{0,25}$ und deren obere Begrenzung durch $Q_{0,75}$ gebildet wird.
- An der Stelle des Medians wird eine horizontale Linie (evtl. betont durch einen Punkt in der Mitte) gelegt.
- Oberhalb und unterhalb der Box erstrecken sich Whisker. Die Länge der Whisker beträgt maximal das 1,5-fache des Quartilabstands. Ist das Minimum bzw. das Maximum der Daten weniger als $1{,}5 \cdot \text{IQR}$ von $Q_{0,25}$ bzw. $Q_{0,75}$ entfernt, reicht der betreffende Whisker nur bis zum Minimum bzw. Maximum.
- Liegen Werte außerhalb der Whisker, werden diese als „Ausreißer“ einzeln dargestellt, meist durch einen Punkt oder ein anderes Symbol.

Abschließendes zu Quantilen

Im Prinzip kann man Werte für jedes beliebige Quantil bestimmen. Relativ häufig werden beispielsweise noch Dezile eingesetzt. Dezile teilen die Daten in zehn gleich große Teile. Speziell gilt: Das unterste Dezil (Grenze zwischen den unteren 10 % und den oberen 90 % der Daten) heißt erstes Dezil, das oberste (Grenze zu den obersten 10 %) heißt neuntes Dezil. Diese beiden Werte werden manchmal anstelle von Minimum und Maximum angeführt.

Für Quantile gibt es eine Reihe anderer Regeln zur Berechnung, vielfach sind das Interpolationsregeln. Die Details müssen nur Spezialisten kennen.

Achtung: Quantile werden uns bald wiederbegegnen, also am besten gleich gut merken (nicht unbedingt die Berechnungsregeln, aber die Grundidee, was Quantile sind)!

Maßzahlen für die Streuung von Daten: Varianz, Standardabweichung und Variationskoeffizient

Die Varianz

Die Varianz ist eine Maßzahl, welche die Streuung der Daten in einem einzigen numerischen Wert ausdrückt. Gleichzeitig werden alle Datenwerte berücksichtigt. Sie wird berechnet als durchschnittliche quadrierte Abweichung der Datenwerte vom arithmetischen Mittel:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2$$

Zur Beachtung: Was die Varianz eigentlich „ist“ (nämlich: quadrierte Abweichungen vom Mittelwert), wird nur aus der ersten Formel deutlich. Bitte verwenden Sie diese beim Nachrechnen!

Die Standardabweichung

Als Folge der Quadrierens der Abweichungen hat die Varianz eine andere Dimension als die Ausgangswerte. In der Standardabweichung wird das Quadrieren wieder rückgängig gemacht.

$$s_x = \sqrt{s_x^2}$$

Die Standardabweichung (englisch standard deviation, abgekürzt „s. d.“) wird daher meist verwendet, wenn in wissenschaftlichen Veröffentlichungen Daten charakterisiert werden.

Man beachte: Varianz und Standardabweichung dürfen (ebenso wie die noch folgenden Größen) nur bei metrischen Daten berechnet werden.

Varianz und Standardabweichung im Beispiel

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2 000	−1 000	1 000 000
5 000	2 000	4 000 000
4 000	1 000	1 000 000
1 500	−1 500	2 250 000
2 500	−500	250 000
Summe		8 500 000

Varianz: $s_x^2 = \frac{1}{5} \cdot 8\,500\,000 = 1\,700\,000$

Standardabweichung: $s_x = \sqrt{1\,700\,000} = 1\,304$

Man sieht: Werte, die weit weg vom Mittelwert liegen, haben den größten Einfluss auf die Varianz. Je mehr Werte weitab vom Mittelwert es gibt, desto größer sind Varianz und Standardabweichung

Varianz und Standardabweichung: weitere Beispiele

Wegen der Quadrierung haben Werte, die weit weg vom Mittelwert liegen, den größten Einfluss auf die Varianz. Je mehr Werte weitab vom Mittelwert es gibt, desto größer sind Varianz und Standardabweichung. Dazu zwei Beispiele:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1 500	−1 500	2 250 000	1 500	−1 500	2 250 000
2 500	−500	250 000	1 500	−1 500	2 250 000
3 000	0	0	2 000	−1 000	1 000 000
3 000	0	0	5 000	2 000	4 000 000
5 000	2 000	4 000 000	5 000	2 000	4 000 000
Summe		6 500 000			13 500 000
Varianz		1 300 000			2 700 000
S. D.		1 140			1 643

Varianz: Schätzung für Grundgesamtheit

Die auf den vorherigen Folien angegebenen Formeln für Varianz und Standardabweichung beschreiben die Streuung in den vorliegenden Daten.

Handelt es sich bei den Daten um eine Stichprobe, und sollen anhand der Stichprobe Varianz und Standardabweichung der Grundgesamtheit geschätzt werden, muss die Varianz nach der folgenden Formel berechnet werden:

$$\widehat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Entsprechend gilt für die Standardabweichung:

$$\widehat{\sigma}_x = \sqrt{\widehat{\sigma}_x^2}$$

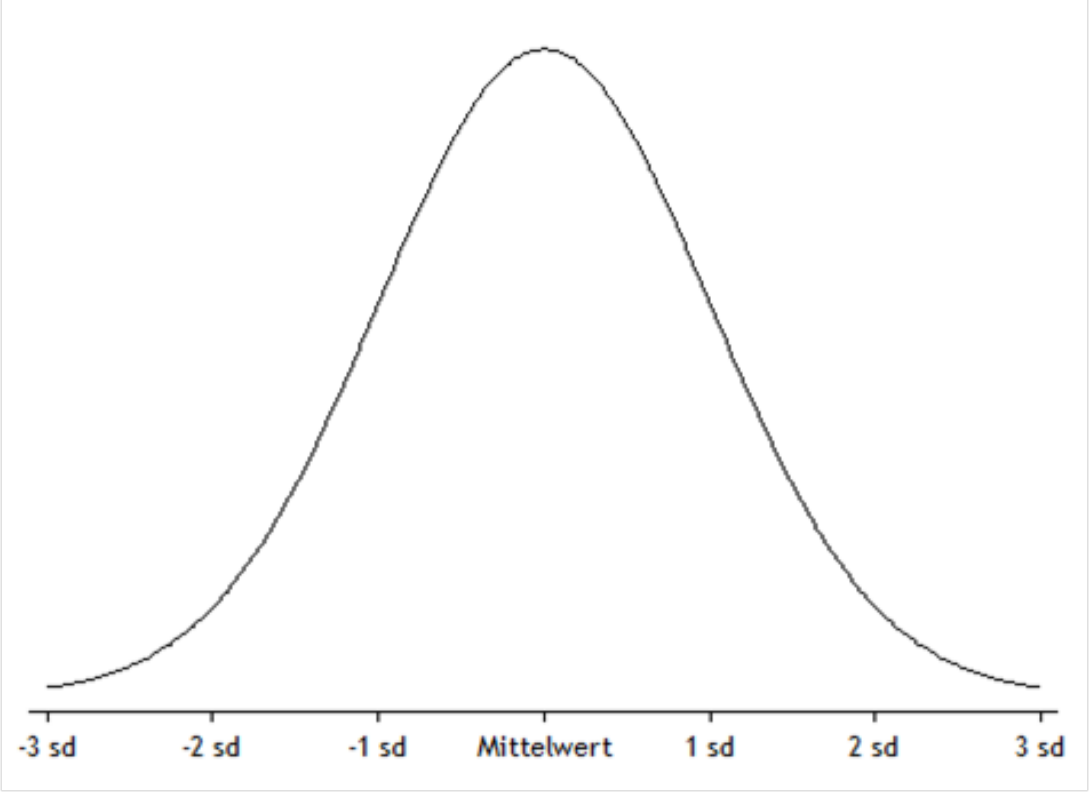
Achten Sie auf die Symbole!

Grundsätzlich gilt (in den meisten Darstellung der Statistik):

- Lateinische Buchstaben stehen für Kennwerte, die eine gegebene Stichprobe kennzeichnen. Beispiele: $\bar{x}$, s, s^2
- Griechische Buchstaben stehen für Werte der Grundgesamtheit. Dabei ist zu unterscheiden:
  - Trägt der griechische Buchstabe ein Dach, so heißt das, dass es sich um eine Schätzung des Wertes der Grundgesamtheit anhand einer Stichprobe handelt. Beispiele: $\widehat{\sigma}_x^2$, $\widehat{\sigma}_x$
  - Trägt der griechische Buchstabe kein Dach, so heißt das, dass es sich um einen bekannten (oder angenommenen) Wert der Grundgesamtheit handelt. Beispiele: σ_x^2, σ_x

Die Standardabweichung in der Beschreibung von Daten I

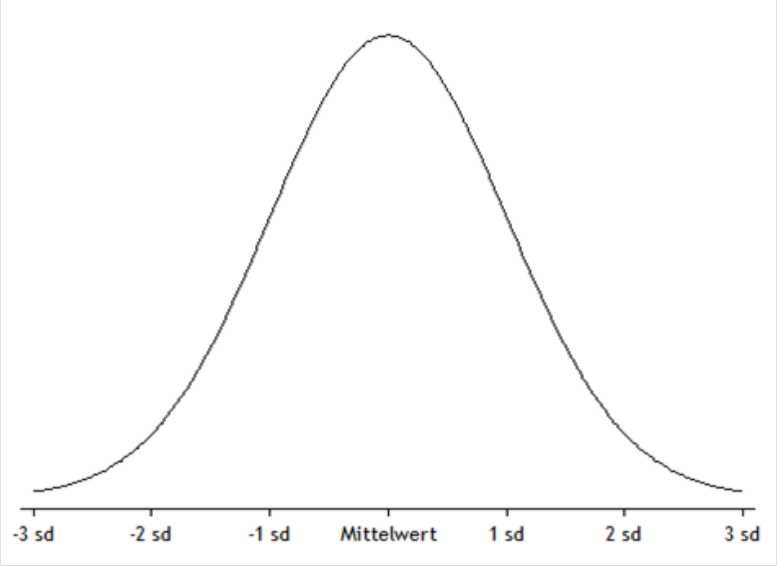
Empirische Variablen folgen manchmal (mehr oder weniger) einer Normalverteilung – einer symmetrischen Verteilung, die man oft als glockenförmig beschreibt.



Die Standardabweichung in der Beschreibung von Daten II

Je mehr die Verteilung einer Variablen einer Normalverteilung entspricht, desto eher kann man die Standardabweichung wie folgt interpretieren:

- Ca. 68 Prozent der Messwerte liegen im Bereich von ± 1 Standardabweichung (sd) symmetrisch um den Mittelwert.
- Ca. 95 Prozent der Messwerte liegen im Bereich von ± 2 Standardabweichung (sd) symmetrisch um den Mittelwert.



Der Variationskoeffizient

Verschiedene Merkmale können ganz unterschiedliche Größenordnungen aufweisen. Die Standardabweichungen können dann nicht sinnvoll verglichen werden.

Der Variationskoeffizient

$$V_x = \frac{s_x}{\bar{x}}$$

drückt die Standardabweichung als Anteil des Mittelwerts aus (im Bsp.: ca. 0,43).

Voraussetzung für die Berechnung des Variationskoeffizienten ist (offenkundig) ein Mittelwert > 0. Wirklich vergleichbar anhand des Variationskoeffizienten sind jedoch nur verhältnisskalierte Variablen.

Schiefe und Wölbung von Verteilungen

Schiefe

Die Schiefe einer Verteilung (einer metrischen Variablen) kann durch die Maßzahl

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3}$$

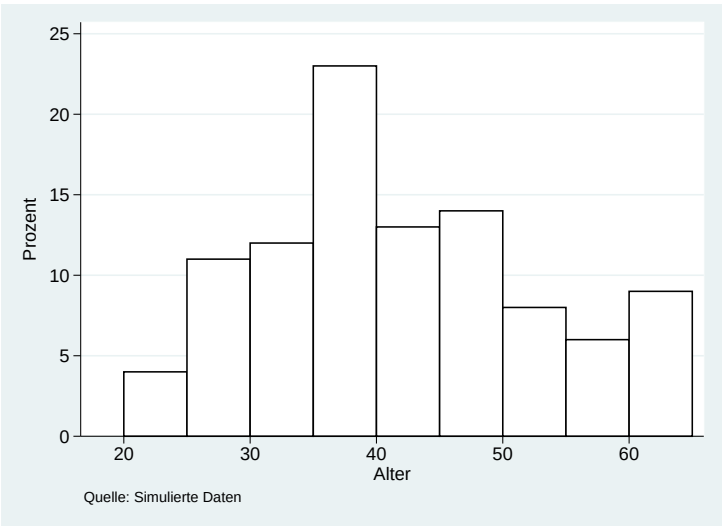
beschrieben werden („Schiefekoeffizient“).

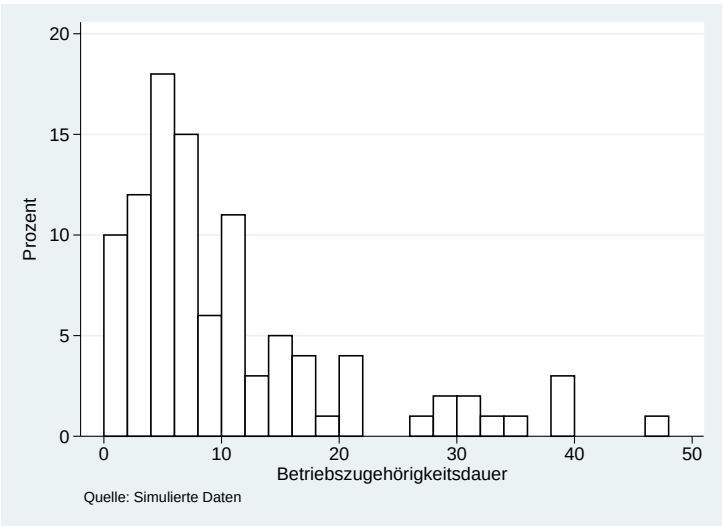
„Messlatte“ für die Schiefe ist eine normalverteilte Variable mit gleichem Mittelwert und gleicher Standardabweichung. Es gilt:

- Ist der Koeffizient größer als 0, ist die Verteilung rechtsschief,
- ist er (ungefähr) gleich 0, entspricht die Verteilung (in etwa) einer Normalverteilung,
- ist er kleiner als 0, ist die Verteilung linksschief.

Schiefe im Beispiel

Die leicht rechtsschiefe Variable Alter hat einen Schiefekoeffizienten von 0,41. Die stark rechtsschiefe Variable Betriebszugehörigkeitsdauer hat einen Koeffizienten von 1,67.





Steilheit oder Wölbung

Die Steilheit oder Wölbung (auch: Kurtosis oder Exzess) einer Verteilung kann durch die Maßzahl

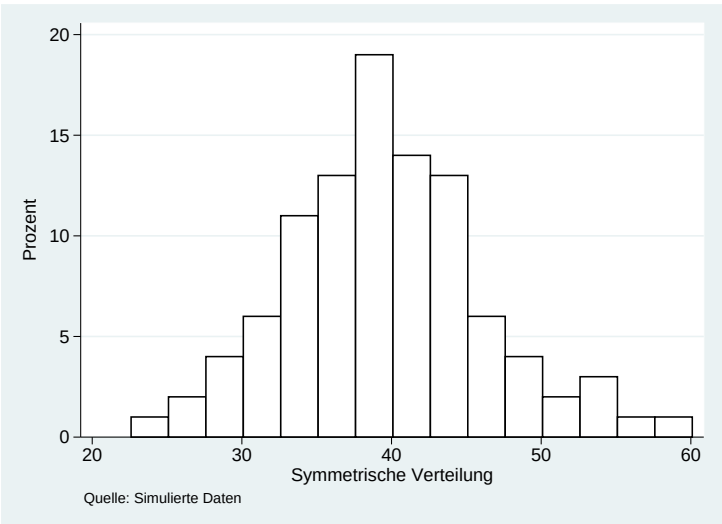
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4} - 3$$

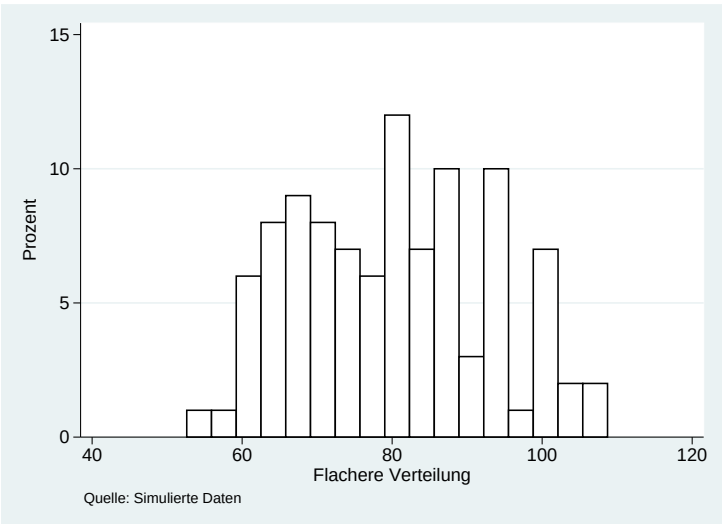
beschrieben werden. „Messlatte“ für die Wölbung ist eine normalverteilte Variable mit gleichem Mittelwert und gleicher Standardabweichung. Es gilt:

- Ist der Koeffizient größer als 0, ist die Verteilung eher steil (besonders viele Werte in der Mitte, wenige an den Enden der Verteilung),
- ist er (ungefähr) gleich 0, entspricht die Verteilung (in etwa) einer Normalverteilung,
- ist er kleiner als 0, ist die Verteilung flach oder hat sogar hat viele Werte an Anfang und Ende.

Wölbung im Beispiel

Die mehr oder weniger normalverteilte Variable links hat einen Wölbungskoeffizienten von 0,27. Die Variable rechts, deren Werte wesentlich weniger im Zentrum konzentriert sind, hat einen Koeffizienten von $-0{,}83$.





Abschließender Hinweis zu Schiefe und Wölbung

Statistik-Software (u. a. SPSS und Excel) verwendet gelegentlich etwas andere Maßzahlen (die nicht immer [leicht zugänglich] dokumentiert sind). Die Tendenz der Ergebnisse ist aber ähnlich wie bei den hier vorgestellten Formeln.

Beachten Sie grundsätzlich: Excel enthält viele Fehler, es kann teilweise einfachste Dinge nicht richtig berechnen (z. B. Quantile).

Die Software Stata verwendet die o. a. Formeln für Schiefe und Wölbung, allerdings unterlässt Stata bei der Wölbung die Subtraktion von 3. (Stata-User sind intelligent – sie können selber von einer gegebenen Zahl 3 subtrahieren!).