2 Eindimensionale Häufigkeitsverteilung

Die einfachste Möglichkeit, vorliegendes Datenmaterial zu beschreiben, besteht darin, die Häufigkeiten einzelner Ausprägungen auszuzählen. Die Gesamtheit aller ermittelten Häufigkeiten gibt uns dann an, wie sich die einzelnen Beobachtungswerte auf die unterschiedlichen Ausprägungsmöglichkeiten aufteilen, oder kurz wie diese verteilt sind. Wir sprechen in diesem Zusammenhang auch von der Häufigkeitsverteilung oder auch einfach nur von der Verteilung der Daten.

2.1 Aufbereitung von Stichprobenwerten

Urliste

Die Urliste enthält sämtliche Beobachtungswerte einer Studie in ihrer ursprünglichen Form ohne größere Aufbereitung und Manipulation. Man spricht in diesem Zusammenhang auch von den sog. Rohdaten.

Alter der Schülerinnen und Schüler einer Klasse

21	19	20	20	21	19	22	25	25
23	20	20	21	21	19	19	21	21

Liste von Beobachtungswerten
x_j Daten, Stichprobenwerte
x_1 bis x_n ; gleiche Werte sind möglich
$j = 1, \ldots, n$

Diese sind von den Merkmalsausprägungen zu unterscheiden:

a_i mit $i = 1, \ldots, k$

Strichliste

Alter	Häufigkeit	Absolute Häufigkeit

Absolute Häufigkeit

Kommt eine Merkmalsausprägung a_i in einer Urliste n_i-mal vor, so nennt man n_i die absolute Häufigkeit von a_i in der Urliste.
Eine Tabelle, die jeder Merkmalsausprägung ihre Häufigkeit zuordnet, heißt Häufigkeitstabelle.

Für die absoluten Häufigkeiten n_i gilt:

Klassenbildung

Werden die verschiedenen Merkmalsausprägungen zu neuen Ausprägungen zusammen gefasst, so spricht man von Klassenbildung oder Klassierung.

Relative Häufigkeiten

Tritt die Merkmalsausprägung a_i in einer Urliste mit Stichprobenwerten n_i-mal auf, so nennt man $\frac{n_i}{n}$ die relative Häufigkeit von a_i.

$$f_i = \frac{n_i}{n}$$

Summenhäufigkeiten

Summenhäufigkeiten geben Antworten auf Fragen wie: Wie viele Schüler sind jünger als 21 Jahre?
Summe der Häufigkeiten n_i bzw. f_i für $a_i \leq c$ ist die Summenhäufigkeit.

Exkurs – Umgang mit dem Summenzeichen

Beispiel: Angenommen die sechs Beobachtungswerte

$x_1 = 2;\ x_2 = 3;\ x_3 = 5;\ x_4 = 6; x_5 = 8$ und $x_6 = 9$ sollen addiert werden.

Wir können dann schreiben:

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 2 + 3 + 5 + 6 + 8 + 9 = 33$$

oder wir können als Kurzform das Summenzeichen setzen:

Allgemein:

$$x_1 + x_2 + x_3 + x_4 + \ldots + x_i + \ldots + x_n = \sum_{i=1}^{n} x_i$$

1. $\underbrace{c + c + \ldots + c}_{n-\text{mal}} =$
2. c ist eine beliebige Konstante. Dann gilt:

$$c \cdot x_1 + c \cdot x_2 + c \cdot x_3 + c \cdot x_4 + \ldots + c \cdot x_n =$$

3. Zwei verschiedenen Summationsvariablen a und b:

$$a_1 + b_1 + a_2 + b_2 + a_3 + b_3 + \ldots + a_n + b_n =$$

4. Sollen Zahlen eines rechteckigen Zahlenschemas

$$\begin{matrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{matrix}$$

aufsummiert werden so kann dies zeilenweise oder spaltenweise geschehen:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} =$$

Aufgaben:

1. Schreiben Sie ausführlich:

   a) $\sum_{i=1}^{n} (-1)^{i+1} i^2 =$

   b) $\sum_{k=2}^{n+1} (-1)^k (k-1)^2 =$

2. Schreiben Sie mit Summenzeichen:

a) $1 + 2^3 + 3^3 + \ldots + n^3 =$

b) $\frac{1}{2} + \frac{2}{2^2} + \frac{3}{2^3} + \ldots + \frac{n}{2^n} =$

c) $1 + 3 + 5 + \ldots + (2n+1) =$

d) $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + - \ldots - \frac{1}{100} =$

2.1.1 tabellarische Aufbereitung - Beispiel

Bei 30 Betrieben wird u. a. jeweils die Anzahl der Beschäftigten ermittelt. Man erhält:
12, 438, 623, 187, 216, 25, 98, 100, 617, 367, 560, 116, 270, 304, 36, 87, 54, 124, 517, 410,
160, 125, 44, 76, 62, 260, 342, 570, 520, 234
Für die Beschäftigungsanzahl der Betriebe sollen Klassen gebildet werden – und zwar 1 bis 100, 101 bis 200 usw.

Bestimmen Sie für diese Klassen zuerst nur die absoluten und relativen Häufigkeiten!

Kl. Nr.	Klasse	Strichliste	abs. Häufigkeit	relative Häufigkeit	abs. Summenhäuf.	rel. Summenhäuf.	abs. Resthäuf.	rel. Resthäuf.
1	1 bis 100							
2								

Wie viele Betriebe beschäftigen weniger als 301 Mitarbeiter?

Formel für die absolute Summenhäufigkeit N_i:

Formel für die relative Summenhäufigkeit F_i:

Wie viele Betriebe beschäftigen mehr als 500 Mitarbeiter?

Formel für die absolute Resthäufigkeit NR_i:

Formel für die relative Resthäufigkeit FR_i:

2.1.2 graphisch – Beispiel

Die Polizei führt an einem bestimmten Punkt in der Innenstadt Radar-Geschwindigkeitsmessungen durch. In einer Viertelstunde wurden folgende Häufigkeiten festgestellt:

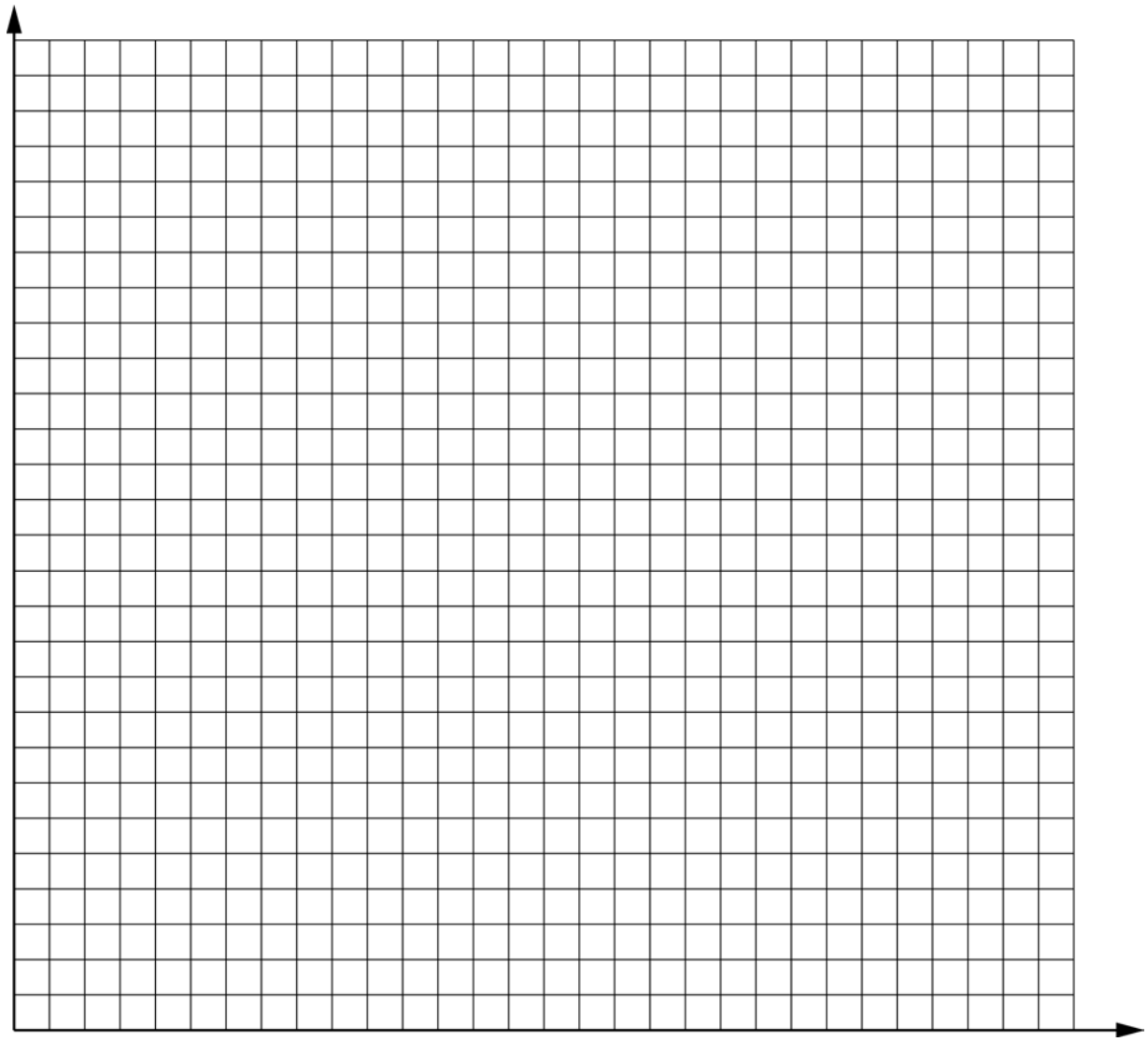
Nr.	Ausprägung a_i	abs. Häuf. n_i	relative Häufigkeit f_i	relative Summenhäufigkeit F_i
1	45	4		
2	46	3		
3	47	5		
4	48	7		
5	49	8		
6	50	12		
7	51	11		
8	52	10		
9	53	10		
10	54	8		
11	55	2		

m

2.1.3 Beispiel

Nach einem Ortsausgangsschild wurden exakt die Geschwindigkeiten von 59 Autos gemessen. Mehrere Geschwindigkeiten wurden zu Klassen zusammengefasst.

j	Klasse k_i	Geschwind.	n_i	f_i				
1	k_1	0 bis u. 25	1					
2	k_2	25 bis u. 50	14					
3	k_3	50 bis u. 71	16					
4	k_4	71 bis u. 86	15					
5	k_5	86 bis u. 96	10					
6	k_6	96 bis 100	3					



2.1.4 graphisch bei klassierten Daten

Angenommen bei einer Erhebung wurde die monatliche Absatzmenge einer bestimmten Brötchensorte in 30 Filialen eines Bäckereibetriebs erhoben. Die Beobachtungswerte lauten:

37176	29901	15144	20112	25432	18320	32770	38696	17160	8524
22138	13007	20556	24748	27936	28791	37322	19207	21086	21316
12941	44981	36180	18428	51525	12601	5588	39070	41004	47688

Häufigkeitsverteilung – Absatzmengen von Brötchen (in Tausend)

j	Klasse k_j von $(c_{j-1}, c_j]$	n_j	f_j
1	(0, 10]		
2	(10, 20]		
3	(20, 30]		
4	(30, 40]		
5	(40, 50]		
6	(50, 60]		

Die linken Klassengrenzen notieren wir mit c_{j-1} die rechte Grenze mit c_j. Demnach gilt $c_0 = 0$, c_1=10, c_2=20 usw.
Die jeweils runden Klammern der linken Grenzen bedeuten, dass die entsprechenden Werte jeweils nicht mehr zu diesen Klammern gezählt werden (ausschließend). Die jeweils eckigen Klammern der rechten Grenzen bedeuten, dass entsprechende Werte noch zu den Klassen gehören (einschließend). Der Wert 10 zählt also zur ersten und nicht etwa zur zweiten Klasse.

→ *Bestimmen Sie die absoluten Klassenhäufigkeiten n_j und die relativen Klassenhäufigkeiten f_j.*

→ *Zeichnen Sie ein passendes Säulendiagramm.*

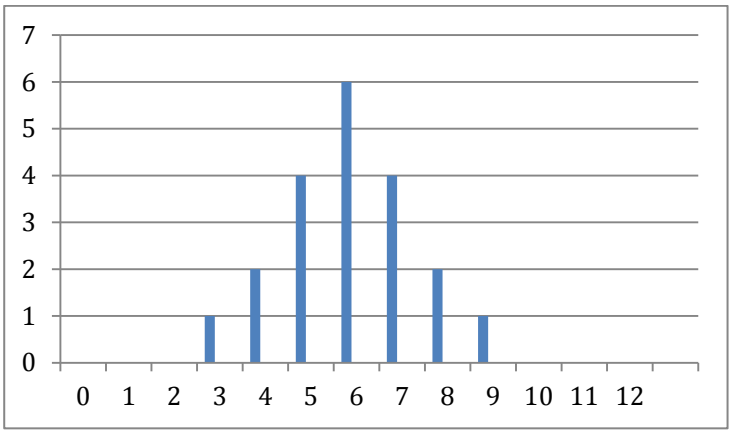
2.1.5 Empirische Verteilungen beschreiben

Zu den grundlegenden Aspekten, die bei der Charakterisierung von Verteilungen regelmäßig von Interesse sind, zählen Lage, Streuung und Schiefe. Mit der Lage ist das allgemeine Niveau der Daten gemeint, während die Streuung deren Variationsbreite (Verschiedenheit) umfasst. Schiefe beinhaltet die Art und Weise, wie eine Verteilung von der Symmetrie abweicht. Die Beschreibung einer Verteilung anhand dieser drei Aspekte ist häufig nur dann sinnvoll, falls die Verteilung unimodal ist.

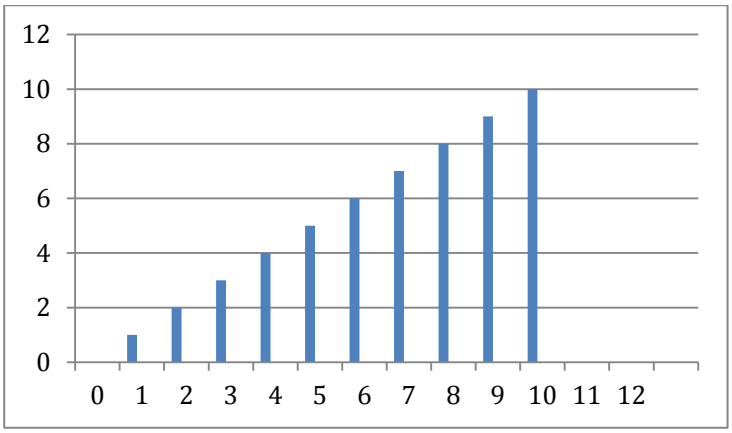
→ *Ordnen Sie zu: linksschiefe Verteilung, rechtsschiefe Verteilung, bimodale Verteilung, symmetrische Verteilung.*

→ *Berechnen Sie den Mittelwert (artihmetisches Mittel) und tragen Sie dieses ein.*

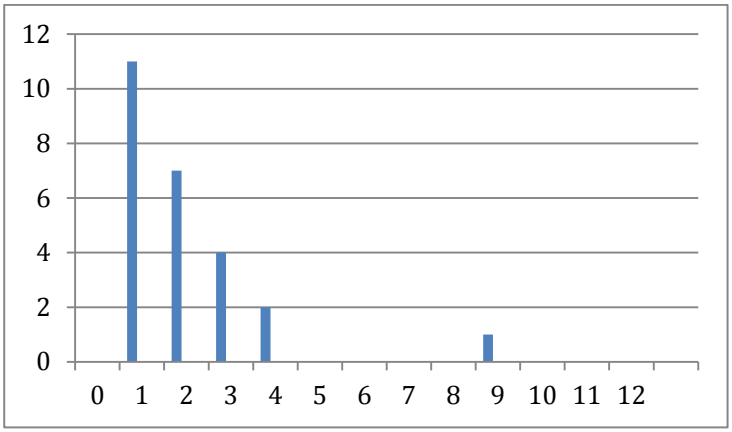
Beispiel 1



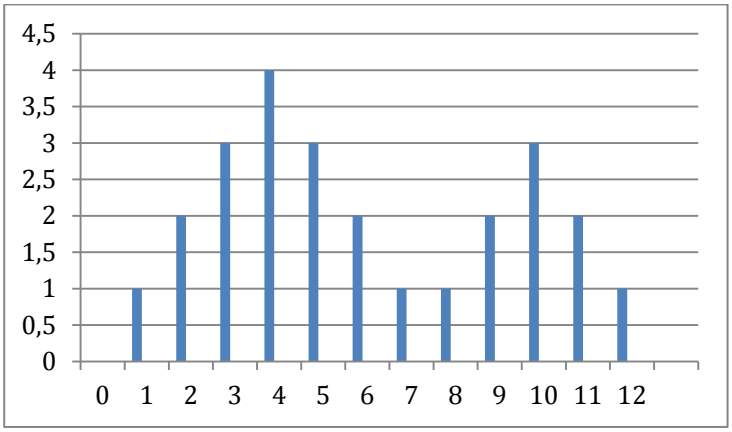
Beispiel 2



Beispiel 3



Beispiel 4



2.2 Statistische Maßzahlen

2.2.1 Lagemaße (arithmetisches Mittel, Median, geometr. Mittel, Modus)

Arithmetisches Mittel $\overline{x}$

a) $\bar{x}$ mit n Stichprobenwerten x_j

Beispiel: In einer Klasse bekommen Schüler ein wöchentliches Taschengeld von:
2 Euro, 5 Euro, 2 Euro, 1 Euro, 1,50 Euro, 2 Euro, 3 Euro, 3,5 Euro, 4 Euro und 3 Euro

b) $\bar{x}$ mit Merkmalsausprägungen a_i und absoluten Häufigkeiten n_i

Beispiel: Bei 25 statistischen Beobachtungen wurden die folgenden Merkmalsausprägungen betrachtet:

Alter in Jahren a_i	20	22	23	24	25	30
Anzahl n_i	3	7	6	3	2	4

c) $\bar{x}$ mit Merkmalsausprägungen a_i und relativen Häufigkeiten f_i

Beispiel: In einem Betrieb sind die monatlichen Bruttoeinkommen wie folgt verteilt:

Gehalt in Euro	2600	3500	4200	8000
Anteil der Arbeitnehmer in Prozent	5%	55%	27,5%	12,5%

Vergleichen Sie das arithmetische Mittel des ersten Beispiels mit dem arithmetischen Mittel folgender Stichprobe und vergleichen Sie:

2 Euro, 5 Euro, 2 Euro, 1 Euro, 1,50 Euro, 2 Euro, 3 Euro, **30 Euro**, 4 Euro und 3 Euro

Median

Ein Median der n Skalenwerte $x_{1,}x_{2,}$ ... x_n ist jede Zahl $\tilde{x}$ mit der Eigenschaft:

Höchstens 50% der Skalenwerte sind kleiner als $\tilde{x}$ und

Höchstens 50% der Skalenwerte sind größer als $\tilde{x}$.

Man kann alle Mediane zu den n Skalenwerten $x_{1,}x_{2,}$ ... x_n leicht angeben, wenn man die Skalenwerte der Größe nach anordnet und dann durchnummeriert.

Modus

Der einfachste Lageparameter ist der sogenannte Modalwert kurz Modus. Er ist der am häufigsten vorkommende Wert.

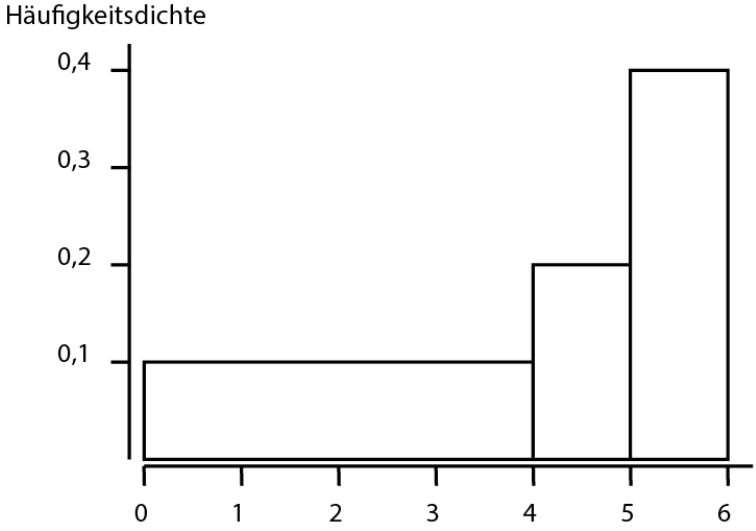
Taschengeld	1 Euro	2 Euro	3 Euro	4 Euro	5 Euro
Häufigkeit	2	3	2	1	2

Bei klassierten Daten ist der Modus die Klassenmitte der Klasse mit der höchsten Häufigkeitsdichte.

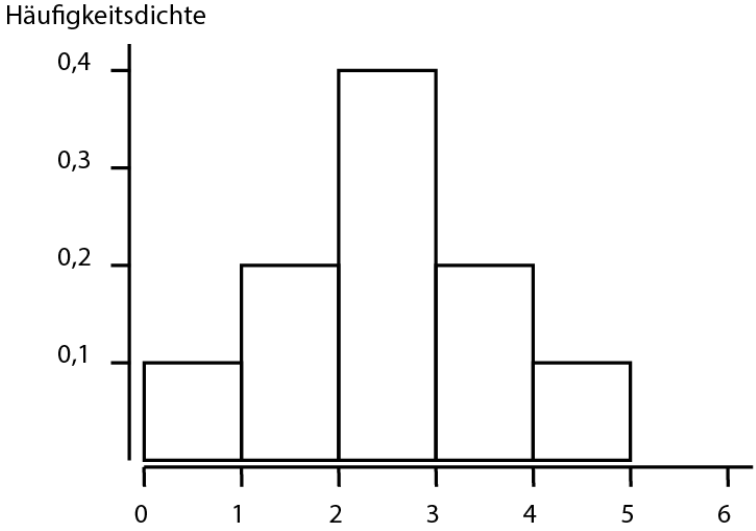
Fechner´sche Lageregel

→ *Wo liegen Modus, arithmetisches Mittel und Median? Zeichnen Sie sinnvoll ein!*

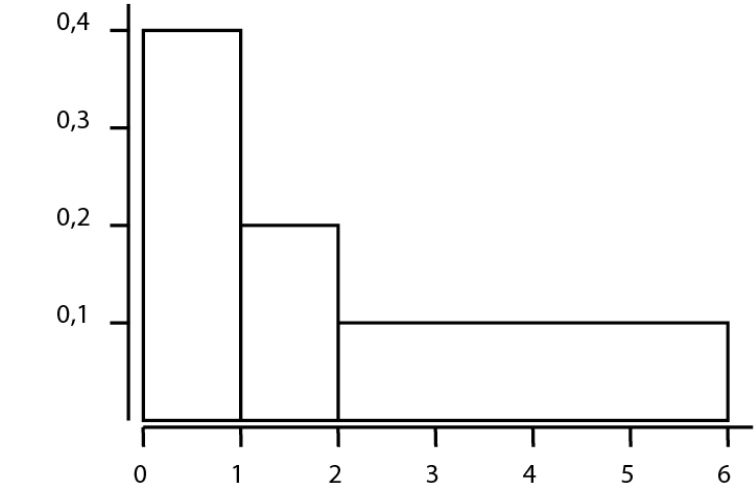
Beispiel 1



Beispiel 2



Beispiel 3



2.2.2 Durchschnittseinkommen – Median und Mittelwert

Einkommensverteilungen sind typischerweise rechtsschief. Deshalb liegt das Durchschnittseinkommen gewöhnlich deutlich über dem Medianeinkommen wie folgende Tabelle anhand von Deutschland zeigt.

Lebensbedingungen, Armutsgefährdung

Einkommensverteilung (Nettoäquivalenzeinkommen)[1] in Deutschland

Soziodemographische Untergliederung	Erhebungsjahr							
	2008	2009	2010	2011	2012	2013	2014	2015
Median des Äquivalenzeinkommens in EUR je Jahr								
Insgesamt	18 309	18 586	18 797	19 043	19 595	19 582	19 733	20 668
Männer	18 777	18 927	19 186	19 389	20 074	20 081	20 228	21 194
Frauen	17 909	18 219	18 448	18 700	19 137	19 067	19 319	20 238
Durchschnittliches Äquivalenzeinkommen in EUR je Jahr								
Insgesamt	21 086	21 223	21 470	21 549	22 022	22 471	22 537	23 499
Männer	21 595	21 648	21 937	22 077	22 663	23 127	23 131	24 042
Frauen	20 595	20 814	21 018	21 037	21 401	21 840	21 964	22 973
Ungleichheit der Einkommensverteilung (S80 / S20 – Rate)[2]								
Insgesamt	4,8	4,5	4,5	4,5	4,3	4,6	5,1	4,8
Gini - Koeffizient								
Insgesamt	30,2	29,1	29,3	29,0	28,3	29,7	30,7	30,1

[1] Referenzjahr für die Ermittlung des Nettoäquivalenzeinkommens ist bei Leben in Europa jeweils das dem Erhebungsjahr vorangegangene Jahr.
[2] Der Quotient stellt das Verhältnis zwischen dem Gesamteinkommen des oberen Fünftels und dem des unteren Fünftels der Einkommensverteilung dar.

Quelle: Leben in Europa (EU-SILC).

Äquivalenzeinkommen
Definition für die EVS und Leben in Europa (EU-SILC)
Das Äquivalenzeinkommen ist ein Wert, der sich aus dem Gesamteinkommen eines Haushalts und der Anzahl und dem Alter der von diesem Einkommen lebenden Personen ergibt. Das Äquivalenzeinkommen wird vor allem für die Berechnung von Einkommensverteilung, Einkommensungleichheit und Armut verwendet. Mithilfe einer Äquivalenzskala werden die Einkommen nach Haushaltsgröße und -zusammensetzung gewichtet. Dadurch werden die Einkommen von Personen, die in unterschiedlich großen Haushalten leben vergleichbar, da in größeren Haushalten Einspareffekte (Economies of Scale) auftreten (z. B. durch gemeinsame Nutzung von Wohnraum oder Haushaltsgeräten).

2.2.3 Arithmetisches Mittel bei gruppierten Daten

Beispiel – Haushaltsgröße in drei Stadtbezirken

Stadtbezirk	Durchschnittliche Haushaltsgröße	Anzahl von Haushalten
Bezirk 1	1,5	282
Bezirk 2	2,4	585
Bezirk 3	1,6	250

→ *Wie lässt sich die durchschnittliche Haushaltsgröße in allen drei Bezirken ermitteln?*

2.2.4 Arithmetisches Mittel und Median bei klassierten Daten

Klassierte Daten stellen einen Spezialfall von Gruppierung dar, bei dem Gruppen in Form von Größenklassen gebildet werden. sofern für alle Klassen (Gruppen) Klassenmittelwerte vorliegen, kann das Gesamtmittel für den klassierten Fall exakt berechnet werden.

Als Beispiel können wir das Beispiel „Absatzmenge von Brötchen“ aus Kapitel 2.1.4 nehmen.
Wenn die Urliste noch vorhanden ist, kann das exakte artithmetische Mittel berechnet werden. Liegen aber nur die Häufigkeiten pro Klasse vor, muss das arithmetische Mittel über die Klassenmitte angenähert werden.

Häufigkeitsverteilung – Absatzmengen von Brötchen (in Tausend)

j	Klasse k_j von $(c_{j-1}, c_j]$	n_j	f_j	m_j
1	(0, 10]	2	0,067	5
2	(10, 20]	8	0,267	15
3	(20, 30]	10	0,333	25
4	(30, 40]	6	0,200	35
5	(40, 50]	3	0,100	45
6	(50, 60]	1	0,033	55

Eine näherungsweise Berechnung des arithmetischen Mittels ergibt:

2.2.5 Geometrisches Mittel

Angenommen ein Unternehmen steigerte seinen Umsatz im Jahr 2011 um 10%. Im Jahr 2012 ging der Umsatz um 10% zurück und im Jahr 2013 konnte das Unternehmen wieder eine Umsatzsteigerung von 30% verzeichnen. Es stellt sich nun die Frage, wie sich in diesem Zusammenhang ein sinnvoller Durchschnittswert für das jährliche Wachstum in diesem Zeitraum ermitteln lässt.
Zur Illustration des Problems sei vereinfachend von folgenden Umsatzzahlen ausgegangen:

Jahr	2010	2011	2012	2013
Umsatz	1000			

Rechnungen:

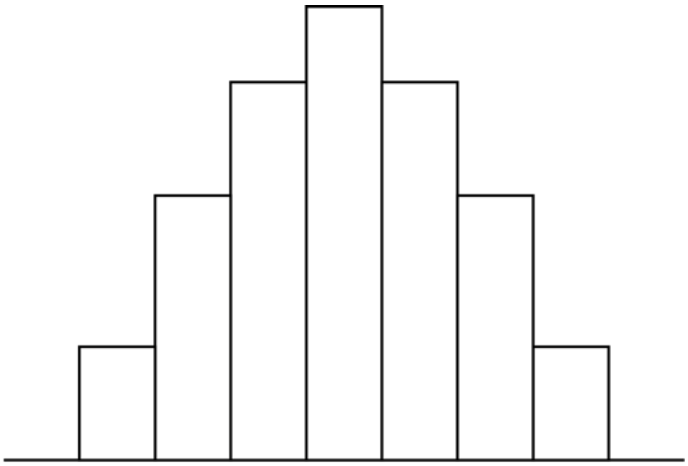
Das arithmetische Mittel der Wachstumsraten ergibt dann:

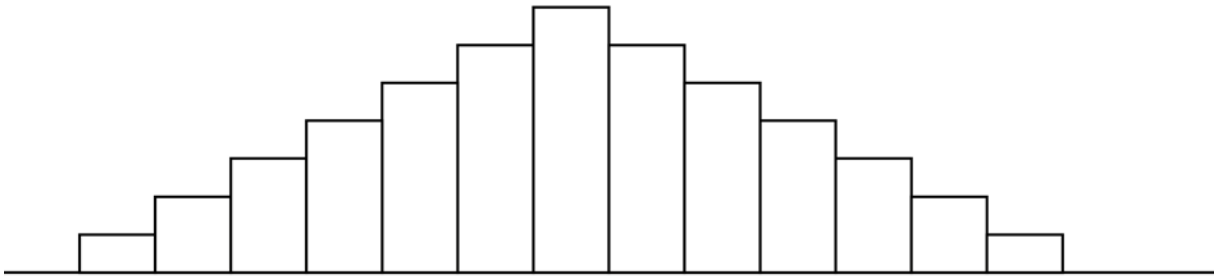
Mit dieser durchschnittlichen Steigerung ergibt sich folgender Umsatz nach drei Jahren:

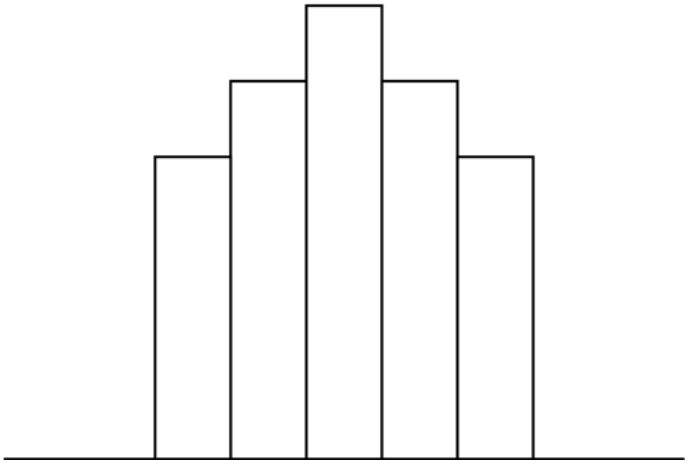
Wachstumsfaktoren: 1,1; 0,9; 1,3
Wachstumsraten +10%, -10%, +30%

Geometrisches Mittel

2.2.6 Streuungsparameter – Spannweite R







Gegeben sind folgende Urlisten mit gleichem arithmetischen Mittel, aber sehr unterschiedlicher Streuung.

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

2.2.7 Streuungsparameter – Mittlere lineare (absolute) Abweichung

Gegeben seien die n metrischen Beobachtungswerte $x_1, x_2, \dots, x_n$. Dann heißt die Kennzahl

$$d_a = \frac{1}{n}\sum_{j=1}^{n}|x_j - a|$$

Die mittlere lineare oder auch absolute Abweichung von einem Mittelwert a. Als Mittelwert kann der Median oder auch das arithmetische Mittel verwendet werden.

Berechnung der mittleren absoluten Abweichung am Beispiel von Seite 25.
Berechnen Sie mit artihmetischem Mittel und mit dem Median.

2.2.8 Streuungsparameter – Median absoluter Abweichungen: MAD

Mittlere absolute Abweichungen sind als Streuungsparameter nicht robust gegenüber Ausreißern.

Beispiel

Urliste 1: 1, 2, 3, 4, 5

Arithmetisches Mittel:

Median

$d_{\bar{x}}$

$d_{\tilde{x}}$

Urliste 2: 1, 2, 3, 4, 500

Arithmetisches Mittel:

Median

$d_{\bar{x}}$

$d_{\tilde{x}}$

Median der absoluten Abweichungen vom Median (MAD) median absolute deviation MAD

Gegeben seien die n metrischen Beobachtungswerte $x_1, x_2, \dots, x_n$. Die absoluten Abweichungen vom Median sind definiert als

$$|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, |x_3 - \tilde{x}|, \dots |x_n - \tilde{x}|$$

Dann heißt der Median dieser Abweichungen Median der absoluten Abweichungen vom Median oder auch MAD.

2.2.9 Streuungsparameter – Varianz und Standardabweichung

Empirische **Varianz** oder **mittlere quadratische Abweichung** vom arithmetischen Mittel

$$\bar{s}^2 = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2$$

Empirische **Standardabweichung**

$$\bar{s} = \sqrt{\bar{s}^2} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}$$

Berechnung anhand der Beispiele von Kapitel 2.2.6

2.2.10 Quantile

Quantile stellen eine Verallgemeinerung des Mediankonzeptes dar. Ein x% Quantil wird (grob gesagt) von x% der Werte unterschritten und von (100-x)% überschritten. Beispielsweise wird das 25% Quantil (auch 1. Quartil genannt) von 25% der Werte unterschritten und von 75% der Werte überschritten.

Als Quartile werden die beiden Quantile mit 25% (unteres Quartil) und 75% (oberes Quartil) bezeichnet. Zwischen oberem und unterem Quartil liegt die Hälfte der Stichprobe, unterhalb des unteren Quartils und oberhalb des oberen Quartils jeweils ein Viertel der Stichprobe. Auf Basis der Quartile wird der Interquartilsabstand definiert, ein Streuungsmaß.

Allerdings wäre das in unserem Beispiel von 10 Werten problematisch. 2,5 Werte sollen unterhalb des 1. Quartils liegen, 7,5 Werte oberhalb des 1. Quartils.
Hier gibt es mehrere Ansätze:

- zwischen 2. und 3 Wert mitteln
- *den zweiten Wert nehmen*
- den dritten Wert nehmen.

Wir definieren folgendermaßen allgemein

$$\tilde{x}_\alpha = \begin{cases} x_{([n\alpha]+1),} & falls\ n\alpha\ keine\ natuerliche\ Zahl\ ist \\ \frac{1}{2}\left(x_{(n\alpha)} + x_{(n\alpha+1)}\right), & falls\ n\alpha\ eine\ natuerliche\ Zahl\ ist. \end{cases}$$

Beispiel

1. Bei 200 Messwerten:

2. Bei 10 Messwerten:

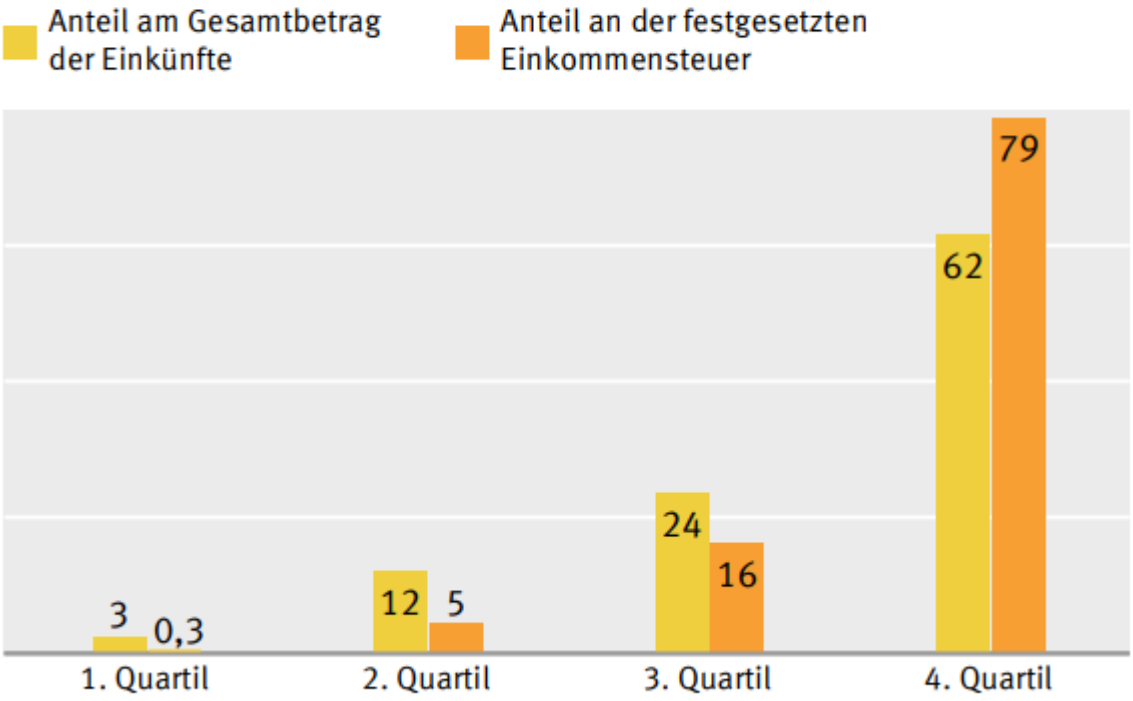
Bestimmen Sie das untere und das obere Quartil und den Quartilabstand jeweils von den Stichproben auf der Seite 25.

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

Beispiel aus dem statistischen Jahrbuch zu Quartilen

Einkünfte und festgesetzte Einkommensteuer[1] 2013
nach Einkunftshöhe der Steuerpflichtigen in Quartilen, in %

Die 25 % der Steuerpflichtigen mit den höchsten Einkünften (4. Quartil) erzielten 61,5 % der Einkünfte und zahlten 78,8 % der gesamten festgesetzten Einkommensteuer. Die 25 % der Steuerpflichtigen mit den niedrigsten Einkünften (1. Quartil) vereinigten 2,5 % der Einkünfte auf sich und trugen 0,3 % zur festgesetzten Einkommensteuer bei.



1 Für Fälle ohne Einkommensteuer-Veranlagung: Einbehaltene Lohnsteuer.

2.2.11 Übungen zu den Streuungsparametern

Übung 1

Eine Gruppe von Tieren mit einer bestimmten Krankheit wird mit einem Medikament behandelt und die Dauer des einsetzenden Heilungsprozesses (in Tagen) festgestellt:

13 17 12 9 10 13 15 11 14 15

Berechnen Sie das arithmetische Mittel und die Standardabweichung.

Übung 2

Bei der medizinischen Untersuchung einer Schulklasse wurden folgende Körpergewichte (in kg) festgestellt.

35 27 36 42 50 32 35 29 44 40 36 38 45 40 42

34 38 43 45 42 37 45 52 48 31 34 46 30 38 35

Berechnen Sie das arithmetische Mittel, die Varianz und die Standardabweichung.

Übung 3

In einem Sportverein sind 15% aller Aktiven 15 Jahre alt, 45% 16 Jahre alt, 20% 17 Jahre, 15% 18 Jahre und 5% 19 Jahre alt.

Bestimmen Sie Varianz und Standardabweichung.

Übung 4

Berechnen Sie geschickt Varianz und Standardabweichung der folgenden Daten:

3	3	1	2	2	4	5	2	2	4	6	4	5	3	1	5	1	4
6	1	4	3	6	5	2	2	6	6	3	2	2	5	3	5	2	3
5	6	2	3	6	3	4	1	6	2	6	2	4	2				

2.2.12 Variationskoeffizient

Angenommen, in einer Marktstudie wird die Streuung von Preisen für bestimmte Produkte bei drei verschiedenen Lebensmitteldiscountern verglichen. So wird für 100 Gramm des jeweils günstigsten Speisesalzes entweder 29, 39 oder 49 Cent verlangt. Eine 10-Kilogramm-Packung des gleichen Waschmittels kostet entweder 19,79 Euro, 19,89 Euro oder 19,99 Euro. Für beide Produkte gibt es die gleiche Standardabweichung von 8,16 Cent.

Rechnung:

Die Preisvariation gemessen an der Standardabweichung wäre bei beiden Produkten somit gleich. Dennoch ist klar, dass eine Standardabweichung von 8,16 Cent im Falle des wesentlich günstigeren Salzes anders zu bewerten ist als beim Waschmittel.

Definition des Variationskoeffizienten:

$$v = \frac{\bar{s}}{\bar{x}}$$

Berechnung am Beispiel:

$v_{Salz} =$

$v_{Waschmittel} =$

2.3 Konzentrationsmessung – Lorenzkurve

Im statistischen Jahrbuch 2017 sind die Einkommenssteuerdaten wie folgt wiedergegeben:
https://www.destatis.de/DE/Publikationen/StatistischesJahrbuch/FinanzenSteuern.pdf;jsessionid=A4A055DA1E570364765ABA9444D29DED.InternetLive1?__blob=publicationFile

9.6 Lohn- und Einkommensteuer
9.6.3 Lohn- und Einkommensteuerpflichtige 2013

Einkommensteuerpflichtige sind alle natürlichen Personen, soweit sie Einkünfte aus einer der im Einkommensteuergesetz bezeichneten sieben Einkunftsarten beziehen (Land- und Forstwirtschaft, Gewerbebetrieb, selbstständige Arbeit, nichtselbstständige Arbeit, Kapitalvermögen, Vermietung und Verpachtung, sonstige Einkünfte). Die unbeschränkte Einkommensteuerpflicht betrifft Personen mit Wohnsitz oder gewöhnlichem Aufenthalt im Inland. Die Gruppe der veranlagten Steuerpflichtigen umfasst die gesetzlich zur Veranlagung verpflichteten und freiwillig veranlagten Personen.

Gesamtbetrag der Einkünfte von ... bis unter ... EUR	Gesamtbetrag der Einkünfte				Zu versteuerndes Einkommen				Festgesetzte Einkommensteuer[1]			
	Steuerpflichtige	%	1 000 EUR	%	Steuerpflichtige	%	1 000 EUR	%	Steuerpflichtige	%	1 000 EUR	%
Insgesamt	**39 780 671**	**X**	**1 411 478 634**	**X**	**37 204 574**	**X**	**1 179 996 342**	**X**	**29 726 791**	**X**	**246 267 710**	**X**
Verlustfälle (Steuerpflichtige mit negativem Gesamtbetrag der Einkünfte)												
Zusammen	**242 090**	*100*	**– 4 410 026**	*100*	**242 087**	*100*	**– 5 156 700**	*100*	**3 239**	*100*	**192 830**	*100*
Gewinnfälle (Steuerpflichtige mit positivem Gesamtbetrag der Einkünfte von ... bis unter ... EUR)												
0 – 7 500	8 015 802	*20,3*	17 919 947	*1,3*	5 440 264	*14,7*	12 054 376	*1,0*	1 458 384	*4,9*	410 160	*0,2*
7 500 – 15 000	5 086 465	*12,9*	57 536 984	*4,1*	5 086 168	*13,8*	43 253 949	*3,6*	2 981 361	*10,0*	1 548 446	*0,6*
15 000 – 25 000	6 532 199	*16,5*	130 236 109	*9,2*	6 532 090	*17,7*	104 865 796	*8,8*	5 529 087	*18,6*	9 294 559	*3,8*
25 000 – 50 000	11 707 883	*29,6*	418 708 358	*29,6*	11 707 807	*31,7*	349 883 280	*29,5*	11 576 070	*38,9*	52 609 446	*21,4*
50 000 – 100 000	6 303 278	*15,9*	427 498 738	*30,2*	6 303 235	*17,1*	357 429 653	*30,2*	6 291 481	*21,2*	77 187 954	*31,4*
100 000 – 250 000	1 653 269	*4,2*	228 784 244	*16,2*	1 653 250	*4,5*	193 323 327	*16,3*	1 649 072	*5,5*	58 981 631	*24,0*
250 000 – 500 000	179 175	*0,5*	59 526 950	*4,2*	179 171	*0,5*	53 542 857	*4,5*	178 202	*0,6*	19 747 929	*8,0*
500 000 – 1 000 000	43 081	*0,1*	28 738 860	*2,0*	43 077	*0,1*	26 524 308	*2,2*	42 619	*0,1*	10 309 239	*4,2*
1 000 000 – 2 500 000	13 353	*0,0*	19 293 553	*1,4*	13 349	*0,0*	18 038 229	*1,5*	13 208	*0,0*	6 880 063	*2,8*
2 500 000 – 5 000 000	2 667	*0,0*	9 074 911	*0,6*	2 667	*0,0*	8 539 692	*0,7*	2 659	*0,0*	3 116 547	*1,3*
5 000 000 und mehr	1 409	*0,0*	18 570 007	*1,3*	1 409	*0,0*	17 697 577	*1,5*	1 409	*0,0*	5 988 908	*2,4*
Zusammen	**39 538 581**	*100*	**1 415 888 660**	*100*	**36 962 487**	*100*	**1 185 153 043**	*100*	**29 723 552**	*100*	**246 074 880**	*100*

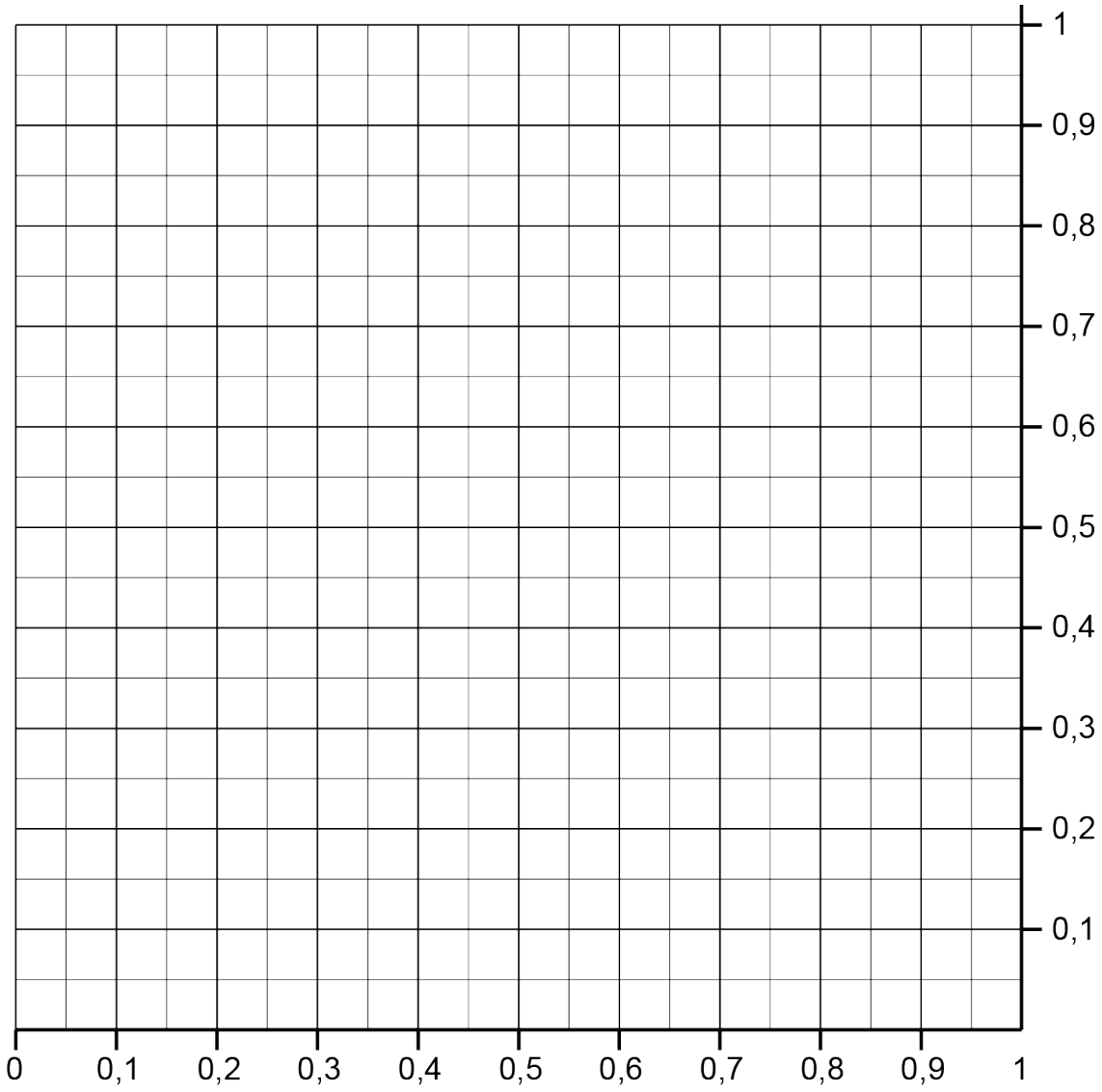
1 Für Fälle ohne Einkommensteuer-Veranlagung: Einbehaltene Lohnsteuer.

Betrachten wir die Gesamtbetrag der Einkünfte. Wir wollen der Frage nachgehen, wie ungleich (oder gleich) die Einkünfte in Deutschland verteilt sind.

Gesamtbetrag der Einkünfte von ... bis	Steuer-pflichtige	Anteil der Steuer-pflichtigen	Gesamt-einkünfte in Tausend Euro	Anteil an den Gesamt-einkünften		
0-7.500	8.015.802	20,27%	17.919.947	1,27%		
7.500-15.000	5.086.465	12,86%	57.536.984	4,06%		
15.000-25.000	6.532.199	16,52%	130.236.109	9,20%		
25.000-50.000	11.707.883	29,61%	418.708.358	29,57%		
50.000-100.000	6.303.278	15,94%	427.498.738	30,19%		
100.000-250.000	1.653.269	4,18%	228.784.244	16,16%		
250.000-500.000	179.175	0,45%	59.526.950	4,20%		
500.000-1.000.000	43.081	0,11%	28.738.860	2,03%		
1.000.000-2.500.000	13.353	0,03%	19.293.553	1,36%		
2.500.000-5.000.000	2.667	0,01%	9.074.911	0,64%		
5.000.000 und mehr	1.409	0,00%	18.570.007	1,31%		
	39538581		1.415.888.661			

Konstruktion der Lorenzkurve

Mit der von Max Otto Lorenz [1905] entwickelten Lorenzkurve wird grafisch beschrieben, wie sich die Merkmalssumme, also die Summe aller Beobachtungswerte, auf die einzelnen Beobachtungswerte aufteilt. Das Grundkonzept besteht darin, in einem Diagramm die kumulativen Anteile der Merkmalssumme gegen die kumulativen Anteile der Beobachtungen abzutragen.



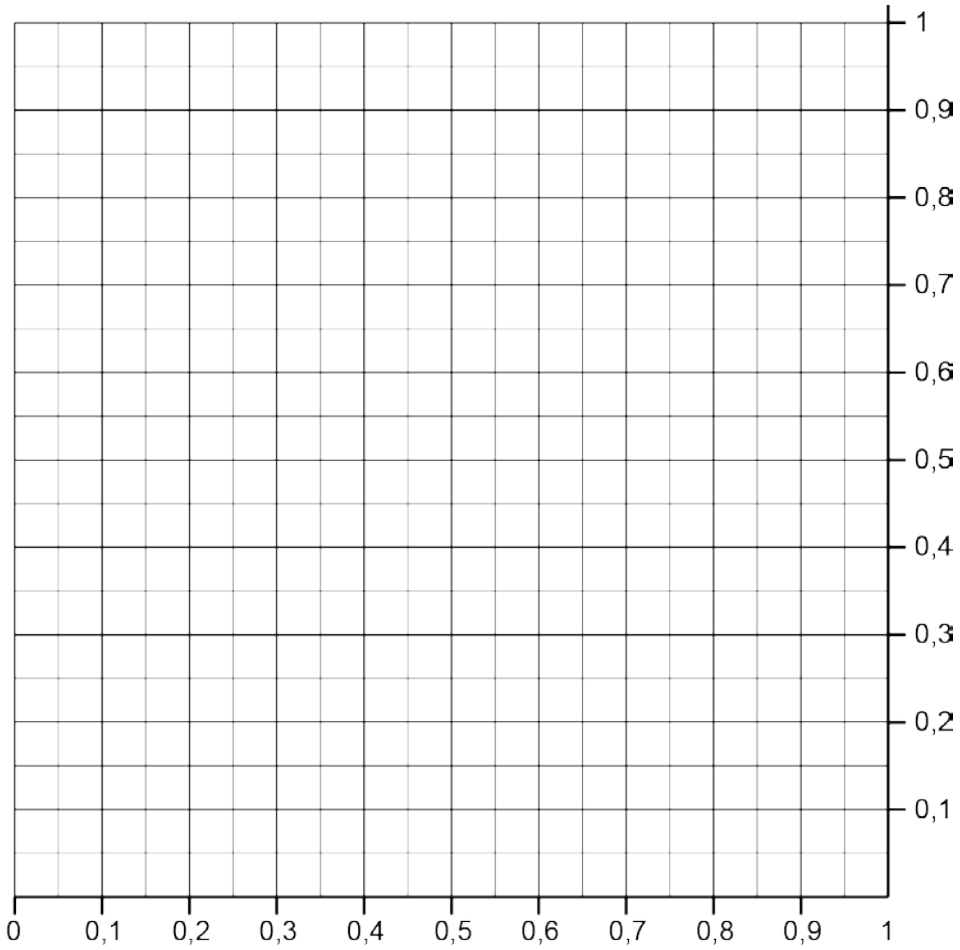
2.3.1 Beispiel Lorenzkurve

In dem Weltentwicklungsbericht 1999/2000 wird die Bevölkerung jedes Landes in eine unterste 10%-Gruppe, eine unterste 20%-Gruppe usw. eingeteilt und diesen Gruppen jeweils der prozentuale Anteil am Einkommen oder Verbrauch zugeordnet.

Für **Brasilien** lassen sich folgende Werte ablesen: (Vorsicht!)

Gruppen	Unterste 10%-Gruppe	Unterste 20%-Gruppe	Zweite 20%-Gruppe	Dritte 20%-Gruppe	Vierte 20%-Gruppe	Oberste 20%-Gruppe	Oberste 10%-Gruppe
Prozentualer Anteil am Einkommen oder Verbrauch	0,8	2,5	5,7	9,9	17,7	64,2	47,9

Zeichnen Sie eine Lorenzkurve, in der Sie alle angegebenen Werte berücksichtigen.



2.3.2 Weitere Länder im Vergleich

Gruppen	Unterste 10%-Gruppe	Unterste 20%-Gruppe	Zweite 20%-Gruppe	Dritte 20%-Gruppe	Vierte 20%-Gruppe	Oberste 20%-Gruppe	Oberste 10%-Gruppe
Prozentualer Anteil am Einkommen oder Verbrauch Dänemark	3,6	9,6	14,9	18,3	22,7	34,5	20,5
Deutschland	3,7	9,0	13,5	17,5	22,9	37,1	22,6
Frankreich	2,5	7,2	12,7	17,1	22,8	40,1	24,9
USA	1,5	4,8	10,5	16,0	23,5	45,2	28,5

2.4 Robustheit

Der Begriff der Robustheit wird in der Statistik mal mehr und mal weniger genau definiert. Im Zusammenhang der induktiven Statistik kann man unter Robustheit allgemein eine „Unempfindlichkeit" gegenüber Abweichungen von in einem Modell geforderten Annahmen verstehen.

Robuste und nicht robuste Kennwerte

Wie bereits festgestellt wurde, ist das arithmetische Mittel keine robuste Statistik da sich der physikalische Schwerpunkt der Daten bei Ausreißern stark verlagert. Im Gegensatz dazu ist der Median robust. Zieht ein Vorstandsvorsitzender einer großen Aktiengesellschaft in eine kleines Dorf, so verändert sich mit Sicherheit das Durchschnittseinkommen in diesem Dorf erheblich, nicht aber das entsprechende Medianeinkommen.

Lagekennwerte		**Streuungskennwerte**	
nicht robust	**robust**	**nicht robust**	**robust**

Anmerkung zur Verwendung

Ein in einem Dorf wohnhafter Millionär ist bedingt durch sein Vermögen oder Einkommen ein Ausreißer. Er passt statistisch nicht richtig dazu. Das Ergebnis statistischer Auswertungen wird durch seine Präsenz „gestört". Ausreißer mit einer solch negativen Konnotation im Sinne einer „Störung" können sich beispielsweise auch durch fehlerhafte Datenerfassungen ergeben.
Untersucht man dagegen die Einkommensverteilung einer Großstadt oder eines ganzen Landes, so liegt es möglicherweise in der Natur der Sache, mit einem gewissen Anteil von Millionären zu rechnen. In einem solchen Fall, sind diese Werte keine Ausreißer, sondern sie gehören zu einem vollständigen statistischen Bild dazu.

2.4.1 Beispiel Robustheit

Eine Befragung der Klassebezüglich der Entfernung zwischen der ADV und dem Geburtsort hat zu folgendem Ergebnis geführt.

Nr	Entfernung ADV Geburtsort in km						
1	40						
2	35						
3	0						
4	325						
5	37						
6	18						
7	32						
8	5623						
9	35						
10	14						
11	16						
12	3						
13	3						
14	47						
15	5						
16	18						
17	1						
18	18						
19	1						
20	0						

Berechnen Sie die Lagekennwerte

- Artithmetisches Mittel
- Median
- Modus

Sowie die Streuungskennwerte

- Spannweite
- mittlere absolute Abweichung vom Median
- mittlere absolute Abweichung vom arithmetischen Mittel
- Quartilsabstand
- Varianz
- Standardabweichung
- Variationskoeffizient
- MAD